# Computational Models of Anaphora Resolution: A Survey

Massimo Poesio (University of Essex),
Simone Paolo Ponzetto (University of Heidelberg),
Yannick Versley (University of Tübingen)

August 25, 2010

**Abstract**

Interpreting anaphoric expressions is one of the most fundamental aspects of language interpretation. The study of **anaphora** and **anaphora resolution** (also known in Computational Linguistics as **coreference resolution**) has brought about many fundamental developments in theoretical linguistics (e.g., the development of dynamic models of language interpretation) and computational linguistics (e.g., the developments of theories of local and global salience) and has important practical applications, e.g., in work on information extraction, summarization, and entity disambiguation. We present in this paper a comprehensive survey of approaches to anaphora and anaphora resolution covering the results of almost forty years of research in the field. The survey is divided into three main parts. The first part covered the the linguistic and psycholinguistic-related issues of the area, as well as early models developed before the widespread availability of corpora annotated with anaphoric information. The second part is dedicated to data-driven methods for coreference resolution: this includes a survey of the corpora annotated with anaphoric information, an overview of the machine learning methods developed for this task and methodologies to evaluate these approaches. Finally, we present in the last part work dedicated on extracting lexical and encyclopedic knowledge features required for anaphora resolution.

## 1 Introduction

**Anaphora resolution** or, as it is also been known since the Message Understanding Initiative (MUC), **coreference resolution** (Aone and Bennett 1995; McCarthy and Lehnert 1995; Kehler 1997; Vieira and Poesio 2000; Soon et al. 2001; Ng and Cardie 2002b; Yang et al. 2003; Luo et al. 2004; Hoste 2005; Daumé III and Marcu 2005, *inter alia*)[1], can be defined as the the task of identifying which parts of a text refer to the same **discourse entity**. The rationale for this task is that the same entity can be referred to in texts through different linguistic expressions. For instance, in (1) *Joshua*, *he* and *his* are **mentions** of the same entity, and the same is true of *cornflakes* and *them*.

[1] See Section 2.6 for a discussion of the difference between these terms.

(1) Joshua really likes cornflakes, but he gets them all over his face.

Coreference is a pervasive phenomenon in natural language and it is one of the fundamental ingredients of semantic interpretation: it is therefore not surprising that it has been intensively studied in linguistics, psycholinguistics, and computational linguistics (CL).

The aim of this survey is to cover CL research on anaphora resolution, starting from the early days and framing this work within the context of work on anaphora and anaphora resolution in other fields as well, in particular linguistics and psycholinguistics. The survey consists of three main parts. In this first part we begin by presenting the linguistic background on context dependence and anaphora, and the terminology (Section 2). We then discuss the factors affecting the interpretation of anaphoric expressions, also considering evidence from psycholinguistics and neuroscience (Section 3). Finally, we briefly survey early work on anaphora resolution (Section 4). In the second part, we concentrate on the so-called 'data-driven revolution', namely the widespread employment of empirically-based techniques for this task, similarly to the trend observed in all subfields of CL. This 'empirical turn' has led in the case of anaphora resolution to the development of annotated corpora and appropriate experimental settings, e.g. evaluation metrics (Section 5), as well as the development of sophisticated machine learning models (Section 6). In the last part (Section 7), we finally discuss methods for extracting from corpora and other resources the syntactic, semantic, and encyclopedic knowledge required to resolve anaphors.

We should point out that this survey's coverage of anaphora resolution will be rather narrow. To begin with, we will only cover dependence on the linguistic context, limiting our discussion of **exophora** (dependence on the visual context) to the absolute minimum notwithstanding the importance of this phenomenon in, for instance, computational work on multimodal and embodied interfaces (Landragin et al. 2002; Kelleher et al. 2005). Secondly, we will focus on interpretation and not discuss generation, notwithstanding the importance of work on generation in the development of theories of discourse models (see, e.g., (Grosz et al. 1995; Dale 1992)). A second limitation is that we will concentrate on nominal anaphora and not discuss ellipsis, even though this type of anaphora has had great importance for in both theoretical and computational linguistics (Webber 1979; Sag and Hankamer 1984; May 1985; Dalrymple et al. 1991; Hardt 1997).

# 2 The Linguistics of Anaphora

## 2.1 Context Dependence

The interpretation of many natural language expressions depends on the context of interpretation; in particular, the interpretation of many noun phrases depends on the entities mentioned in the **linguistic context**–the previous utterances and their content. Such dependency on the entities in the linguistic context is particularly obvious in the case of pronouns, whose interpretation entirely depends on them, as illustrated by the following dialogue fragment from the TRAINS dialogues (Heeman and Allen 1995). In this example, the very same expression, personal pronoun *it*, is interpreted in totally

different ways in utterances 3.1 (where it refers to engine E2) and 5.4 (where it refers to engine E1). Demonstrative pronouns as well may depend on entities introduced in the linguistic context, as illustrated by demonstrative *that* in 4.3. We will use the term **anaphoric** to indicate expressions that depend on the linguistic context, i.e., on objects explicitly mentioned or objects whose existence can be inferred from what has been said.

(2)

| | | | |
|---|---|---|---|
| 1.1 | M | : | all right system |
| 1.2 | | : | we've got a more complicated problem |
| 1.4 | | : | first thing _I'd_ like you to do |
| 1.5 | | : | is send engine E2 off with a boxcar to Corning to pick up oranges |
| 1.6 | | : | uh as soon as possible |
| 2.1 | S | : | okay |
| 3.1 | M | : | and while it's there it should pick up the tanker |
| 4.1 | S | : | okay |
| 4.2 | | : | and that can get |
| 4.3 | | : | we can get that done by three |
| 5.1 | M | : | good |
| 5.3 | | : | can we please send engine E1 over to Dansville to pick up a boxcar |
| 5.4 | | : | and then send it right back to Avon |
| 6.1 | S | : | okay |
| 6.2 | | : | it'll get back to Avon at 6 |

Pronouns are not the only noun phrases whose interpretation depends on the entities in the context, and context-dependent noun phrases may depend on a broader context. (2) also contains the definite NP *the tanker* in 3.1, whose interpretation depends on the **visual** context, which in the TRAINS dialogues is a map of the 'TRAINS world' shared between the participants. *The tanker* has not been mentioned before, but it's on this map, and therefore it is shared and has high salience and can be referred to (Clark and Marshall 1981); this type of context dependence is usually called **(visual) deixis**.[2] Such examples illustrate the fact that the noun phrases in these examples are better viewed as depending on what is usually called the **discourse situation** or **utterance situation** (Barwise and Perry 1983), that includes both the linguistic context and the surroundings in which the participants operate.[3] Following the terminology of Discourse Representation Theory (DRT) (Kamp and Reyle 1993), we will call the set of entities introduced in the discourse situation U, for 'Universe of Discourse'. There is plenty of work in CL on interpreting references to the visual context (Poesio 1993, 1994a; Beun and Cremers 1998; Landragin et al. 2002; Kelleher et al. 2005), but this work falls outside the scope of this survey, in which we will focus on anaphora.

More in general, the interpretation of noun phrases depends on the **domain of interpretation**: the subset of the world under discussion. This subset is completely specified in artificial settings like the TRAINS world or the visual world of some psy-

[2]Visual deixis is a type of **exophora**, but this term is not much used in CL.

[3]There are some constraints on what can be referred to in this way (Sag and Hankamer 1984); see below for a discussion of visual focus.

chological experiments, but in the general case depends on shared knowledge. For instance, the proper name *David Mitchell* refers to at least twenty different individuals (that's the number of interpretations in Wikipedia), among which those in (3a) and in (3b). Which particular *David Mitchell* is being referred to in a conversation or in a text depends on the domain of interpretation.

(3) a. David Mitchell (born 12 January 1969) is an English novelist. He has written four novels, two of which were shortlisted for the Booker Prize.

b. David Mitchell (born 14 July 1974) is a British actor, comedian and writer. He is one half of the comedy duo Mitchell and Webb, alongside Robert Webb, whom he met at Cambridge University.

Indeed, under the most widely accepted theory about their meaning (Kripke 1972), the interpretation of proper nouns only depends on the domain of interpretation, because proper nouns are **directly referring**: they are the natural language encoding of constants, and therefore the object they are referring to is directly encoded in their semantics (as opposed to being recovered from the discourse situation). A reference to *David Mitchell* would be infelicitous if the domain of interpretation would not allow to uniquely identify which is being referred to.

Under this view, the process of interpreting proper nouns is completely different from that of interpreting pronouns and nominals. The interpretation of the second mention of *Avon* in (2), for instance, would not be obtained by finding an antecedent in the discourse situation, but it would come straight from the lexical semantics of proper noun *Avon*–or, more plausibly, through a pragmatic process of identifying the appropriate domain of interpretation, and the object referred to within that domain. The conclusion that the two instances of proper noun *Avon* are mentions of the same object would be obtained indirectly, through the fact that they both refer to the same object: it would be a genuine 'coreference' task.

There is plenty of work in CL on disambiguating direct references to the domain of interpretation, particularly now that Wikipedia provides unique identifiers for many objects–e.g., the two interpretations of *David Mitchell* above correspond to different Wikipedia pages (Bunescu and Paşca 2006; Csomai and Mihalcea 2008a)– but this work, as well, falls outside the scope of this paper, and even CL systems concentrating on identifying links between named entities do not identify coreference indirectly through the reference of proper nouns, whereas systems that attempt to interpret all noun phrases still need to model the context-modifying effect of proper nouns as they make antecedents available for pronouns and nominals.[4]

The choice of the domain of interpretation also affects the interpretation of nominals by fixing their **domain of quantification**–the set of objects of the type specified by the nominal complex which are included in the domain of interpretation (Partee 1995; Cooper 1996). For instance, what makes the use of definite NP *the tanker* in (2) felicitous is the fact that the domain of quantification of nominal **tanker** consists of a single object (in the TRAINS dialogues the domain of interpretation coincides with the visual context).[5] The domain of quantification can also be specified by the linguistic context.

[4] And the direct reference theory of proper names is being challenged in semantic theory (Geurts 1997).

[5] Readers may have noticed that the interpretation of expressions like *tanker* is 'context dependent' also in the sense that it depends on the sense of the word *tanker* intended in the circumstances of utterance. We

In the following example, the expression *most employees* is evaluated with respect to the firm mentioned in the first sentence, whereas *the management* is interpreted as the management of the firm.

(4) Kim worked for three years in a large firm. Most employees were friendly, but the management was very distant.

However, we are not aware of much work on identifying the domain of quantification of a nominal apart from (Poesio 1993, 1994a).

## 2.2 Types of Context-dependent Expressions

Nominals are not the only expressions whose interpretation is dependent on the discourse situation in the sense above. Other examples include expressions that could be viewed as the analogous for the verbal interpretation domain of pronouns, such as **pro-verbs** like *did* in (5a) and ellipsis such as **gapping** in (5b). But just as pronouns are only the most extreme example of context-dependence among nominals, full verbal expressions have a context-dependent component as well. In (5c), for instance, the time of listening to the messages is pragmatically determined by the discourse (Partee 1973; Dowty 1986; Kamp and Reyle 1993).

(5) a. Kim is making the same mistakes that I did.
b. Kim brought the wine, and Robin _ the cheese.
c. Kim arrived home. She listened to the messages on her answering machine.

A great deal of interest was paid to ellipsis in the early years of computational linguistics (Woods et al. 1972; Webber 1979; Dalrymple et al. 1991) but modern corpus-based work on the interpretation of anaphoric expressions in computational linguistics (and psycholinguistics) tends to focus on the identification of the antecedents of nominal expressions, primarily because of the lack of annotated resources for studying other types of anaphora.[6] For this reason, we will concentrate on nominal anaphoric expressions in this survey.

Nominal expressions play four types of semantic function in theoretical linguistics:

**Referring** Following the terminology used in functional linguistics and natural language generation, we will use the term **referring noun phrases** to indicate noun phrases that introduce new entities in a discourse, or require a link to previously introduced entities. Examples include sentences like *A train arrived soon after*, where *A train* introduces a new discourse entity; or *It left immediately*, where *it* refers to a previously introduced entity. We discuss referring noun phrases and their semantics in greater detail shortly.[7]

will not be concerned here with this sense of context dependence, but only with expressions that are context dependent in that their interpretation depends on the entities contained in universe of discourse U.

[6] A notable exception is the work by Hardt, e.g., (Hardt 1997). Also, the presence of a VP ellipsis detection and resolution task at SEMEVAL-2010 indicates a renewed interest.

[7] This sense of 'referring noun phrase' is clearly distinct from the sense in which the term 'referring' is used in the philosophical and semantics literature. See Section 2.6.

**Quantificational** Quantificational noun phrases denote relations between the set of objects denoted by the nominal complex and the set of objects denoted by the verbal phrase: e.g., in *Few trains arrived in time*, the quantificational noun phrase *few trains* expresses a relation between the set of trains and the set of objects arriving late–namely, that few of the members of the first set are members of the second set as well:

$$\textbf{few}(\lambda\, x.\textbf{train}(x),\lambda\, x.\textbf{arrive-late}(x))$$

**Predicative** Predicative noun phrases express properties of objects. For instance, in *Kim is a preacher*, the noun phrase *a preacher* expresses a property of Kim (as opposed to referring to a second object).

**Expletive** In languages like English, where verbal arguments always have be filled on syntactic grounds, forms like *it* and *there* can also be used to express semantically vacuous **expletives** as well as pronouns, as in example (6).

(6) It is half past two.

However, an important advance deriving from corpus linguistics work in the last twenty years is the realization that such distinctions are not always easy to make, even for humans (Poesio and Vieira 1998; Poesio and Artstein 2005). For instance, pronoun *it* in utterance 37.7 in fragment (7), also from the TRAINS dialogues, could be interpreted either as an expletive or as a reference to the proposed action of 'going through Dansville'.

(7)

| | | | |
|---|---|---|---|
| 37.1 | M | : | um |
| | | | [5sec] |
| 37.2 | | : | oh kay |
| 37.3 | | : | um |
| 37.4 | | : | ... then I guess we might as well go through Dansville |
| 37.5 | | : | so |
| 37.6 | | : | th / dz / cn / dyou / |
| 37.7 | | : | does it seem like a reasonable alternative to |
| 37.8 | | : | dealing with the engine that's hanging out in Elmira |

Whether a noun phrase is considered referring or quantificational is often a matter of one's theoretical assumptions: in some theories all nominals are considered quantifiers, wheres in DRT and other theories, definites and indefinites are considered of a different type from other nominals. Determining whether an NP is predicative or referring, as well, depends to some extent on the theoretical framework. In annotation schemes such as MUC's all NPs occurring in copular clauses and appositions are treated as referential, which leads to the kind of problems discussed by van Deemter and Kibble (2000) (see Section 2.6). By contrast, most linguistic theories assume that NPs in these positions are predicative, which is not always the case, as shown by example (8), where both *That* and *The Cedars* are referring.

(8) See that light among the trees? That is The Cedars, and beside that lamp sits a woman whose anxious ears have already, I have little doubt, caught the clink of our horse's feet. (Conan Doyle, *The man with the twisted lip*)

A more complex example of difficulty in classifying a noun phrase as predicative or referring is the underlined NP in (9) from the ARRAU corpus, which would seem to be coreferring with *Mr. Hoffman* yet appears to be playing more of a predicative role.

(9) Mr. Lieber, the actor who plays Mr. Hoffman, says he was concerned at first that the script would "misrepresent an astute political mind, one that I admired," but that his concerns were allayed.
The producers, he says, did a good job of depicting someone "who had done so much, but who was also a manic-

Predicative noun phrases usually depend less on the universe of discourse $U$ than other types of nominals (although they can depend on context in other respects of course). Thus, our only concern with predicative NPs in this paper will be the fact that many types of noun phrases can be used referentially in some contexts and predicatively in others. Quantificational NPs are often context dependent, but in the sense that their domain of quantification is contextually specified, as discussed above. We will therefore concentrate here on referring expressions, and on the problem of selecting the discourse entity they are associated with, that we will call **anchor** in the most general case, **antecedent** in the case the relation between the referring expression and the anchor is one of identity (see below).

There are many varieties of referring noun phrases, which differ primarily according to the rules that govern their anaphoric behavior (Reinhart 1976; Chomsky 1981; Gundel et al. 1993; Garrod 1993; Garnham 2001). Such varieties include:

**Reflexives,** as in *John bought himself a parrot*;[8]

**Pronouns,** which in turn can be divided into

- **Definite pronouns** such as *he* or *she*, as in *Ross bought {a radiometer / three kilograms of after-dinner mints} and gave {it / them} to Nadia for her birthday*. (Hirst 1981)
- **Indefinite pronouns** such as *one* in *Kim bought a t-shirt so Robin decided to buy one as well* (Webber 1979).
- **Demonstrative pronouns** such as *that* in example (2), utterance 4.3.

**Nominals,** i.e., noun phrases that have a noun as head, such as *a man*, *a woman*, and *the man* in (10).

(10) A man and a woman came into my shop yesterday. The man wore a baseball hat.

**Proper names,** such as *Kim* and *Robin* in *Kim and Robin are good friends even though Kim likes sports whereas Robin prefers reading*.

[8]Reflexives are also known as 'anaphors' in Binding theory (see below).

As said above, proper names differ from other referring noun phrases from a semantic point of view, in that they are directly referring rather than referring to an entity introduced in the linguistic context; demonstratives, as well, can be directly referring (both pronouns and nominals) (Kaplan 1977). In this survey however we will concentrate on methods for establishing coreference rather than identifying the referent of noun phrases, thus these claims about proper names and demonstratives will be primarily of interest in that they suggest that such nominals will often used to introduce new entities in the linguistic context. But this is true for nominals as well, as shown by (11) (from the 1993 TRAINS corpus; reported by J. Gundel) where *the maximum number of boxcars of oranges that I can get to Bath by 7 a.m. tomorrow morning* is not anaphoric–indeed, most studies find that a majority of definite NPs serve this purpose (Fraurud 1990; Poesio and Vieira 1998). (We discuss some statistics about the distribution of referring NPs in corpora below.)

(11)

| | |
|---|---|
| S | hello can I help you |
| U | yeah I want t- I want to determine |
| | the maximum number of boxcars of oranges that I can get |
| | to Bath by 7 a.m. tomorrow morning |
| | so hm so I guess all the boxcars will have to go through oran- |
| | through Corning because that's where the orange juice factory is |

Another difference between types of referring expressions intensively discussed in Linguistics is that between reflexives and (personal) pronouns, illustrated by (12), in which *herself* must corefer with *Susan*, but *her* cannot, has been investigated in depth in generative syntax, even leading to the development of a whole new Chomskyan paradigm in the '80s (Government and Binding) (Reinhart 1976; Chomsky 1981).

(12) Susan considered herself fortunate to meet her.

Several researchers have concerned themselves with the factors influencing the choice among multiple admissible linguistic forms (Ariel 1990; Garrod 1993; Gundel et al. 1993; Passonneau 1993; Almor 1999; Poesio 2000). Gundel et al. (1993) investigated in depth the difference between personal and demonstrative pronouns (i.e., the difference between *it* and *that*) using corpus data (see also (Linde 1979; Passonneau 1993)), whereas several papers by Garrod and colleagues (e.g., (Garrod 1993)) discuss behavioral evidence concerning the difference between definites and pronouns and between definites and proper names. We will discuss these differences in the Section 3.

It is important for the purposes of the following discussion to point out that no form of referring expression is invariably referring or invariably context dependent. Even pronouns can sometimes be non-referring, as shown by the example of expletives.

## 2.3 The Relation of Referring Expressions to their Context

Apart from their surface form, a second important difference between referring expressions concerns the relation between the entities they refer to and the entities already in the context. Linguists make a first, broad distinction between referring expressions that introduce **discourse-new** entities in a linguistic context–entities not mentioned before–like the first mention of *Joshua* in (1), and expressions that refer to **discourse-old**, i.e.,

an entity already introduced, like the pronouns *he* and *his* in the same example, that refer to Joshua as well (Prince 1992). Forcing Prince's terminology a bit, we will say that the first mention of Joshua is a discourse-new *expression*, whereas the second and third are *discourse old expressions*. Discourse-new expressions can be further differentiated between expressions referring to entities that are completely new to the hearer (**hearer-new**) and expressions referring to entities that can be expected to be known to the hearer / reader (e.g., references to *New York City*), which are called **hearer-old** (Prince 1992).

Some discourse-new entities can be related to the linguistic context as well, albeit more indirectly, and thus are also considered anaphoric. Indefinite pronouns *one* and *another* generally stand in an **identity of sense** relation to their anchor: they refer to a different object of the same type, as in (13). Definite pronouns may also be used in the same way, as in so-called **paycheck pronouns** from famous example (14).

When the anchor is a quantified expression, as in (15), a pronoun with that anchor behaves like a variable in a procedure that gets repeatedly called over the elements specified by the restriction of the quantifier; that the relation between the pronoun and its anchor is not of identity in these cases is seen most clearly when the quantifier is downward entailing, like *no* in this case. We talk in these cases of **bound** anaphora.

Finally, in **associative** anaphora, the context-dependent nominal is related to its anchor by a relation such as part-of, as in (16). In these cases, to identify the antecedent a **bridging inference** is generally required (Clark 1977; Sidner 1979; Vieira 1998).

(13) Sally admired Sue's jacket, so she got one for Christmas. (Garnham 2001)

(14) The man who gave his paycheck to his wife is wiser than the man who gave it to his mistress. (Karttunen 1976)

(15) No Italian ever believes that the referee treated his team fairly.

(16) We saw a flat yesterday. The kitchen is very spacious but the garden is very small.

None of these distinctions is always easy to make (Poesio and Vieira 1998). As Poesio and Vieira and others showed, readers can usually discriminate between discourse-new and discourse-old expressions, but even agreement on these distinctions is never complete; such studies typically find $\kappa$ values of around .7.[9] More complex distinctions, such as trying to distinguish between hearer-new and hearer-old expressions, tend to lead to disagreements.

Surprisingly, even when an expression is clearly anaphoric, it is not always easy to tell what its anchor is, or what is the the exact relation between an anaphor and its anchor (Poesio and Vieira 1998; Poesio and Artstein 2005; Versley 2008). This difficulty is illustrated by examples like the following, from the WSJ portion of the ARRAU corpus, where the possessive description *its machines* in sentence (17e) could refer either to the 'three small personal computers' introduced in sentence 1, or to the entire range of computers sold by Texas Instruments (thus paralleling the reference to

[9] $\kappa$ is a **coefficient of agreement** measuring the extent to which different coders agree on a judgment. It has value 1 for perfect agreement; a value of .7 is normally taken to indicate moderate agreement (Carletta 1996; Artstein and Poesio 2008).

the machines sold by Compaq in the previous clause), but it's not clear which. We refer to these cases as cases of **underspecified identity**.

(17) a. Texas Instruments Inc., once a pioneer in portable computer technology, today will make a bid to reassert itself in that business by unveiling three small personal computers.

b. The announcements are scheduled to be made in Temple, Texas, and include a so-called "notebook" PC that weighs less than seven pounds, has a built-in hard disk drive and is powered by Intel Corp. 's 286 microprocessor.

c. That introduction comes only two weeks after Compaq Computer Corp., believing it had a lead of three to six months on competitors, introduced the first U.S. notebook computer with such features.

d. Despite the inevitable comparison with Compaq, however, Texas Instruments' new notebook won't be a direct competitor.

e. While Compaq sells its machines to businesses through computer retailers, Texas Instruments will be selling most of its machines to the industrial market and to value-added resellers and original-equipment manufacturers.

## 2.4 Discourse Models

The ideas about context (and in particular, of linguistic context) and anaphora introduced in the previous sections were made more precise through the development of the so-called **discourse model** hypothesis (Karttunen 1976; Webber 1979; Kamp 1979, 1981; Sanford and Garrod 1981; Heim 1982; Garnham 1982, 2001) and **dynamic** models of discourse interpretation. The discourse model hypothesis states that context dependent expressions are interpreted with respect to a discourse model which is built up dynamically while processing a discourse, and which includes the objects that have been mentioned (the universe of discourse $U$ introduced above). This hypothesis may at first sight seem to be vacuous or even circular, stating that context dependent expressions are interpreted with respect to the context in which they are encountered. But in fact three important claims were made in this literature. First, that the context used to interpret utterances is itself continuously updated, and that this **update potential** needs to be modelled as well. Second, that the objects included in the universe of discourse $U$ / discourse model are not limited to those explicitly mentioned. The following examples illustrate the fact that a number of objects that can be 'constructed' or 'inferred' out of the explicitly mentioned objects can also serve as antecedents for context dependent nominals, including sets of objects like the set of John and Mary in (18), or propositions and other abstract objects like the fact that the court does not believe a certain female individual in (19). In fact, the implicitly mentioned object may have been introduced in a very indirect way only, as in the case of (20), where *the government* clearly refers to the government of Korea, but the country itself has not yet been mentioned either in the text or the title. These implicitly mentioned objects constitute what Grosz (1977) called the '**implicit focus**' of a discourse.

(18) John and Mary came to dinner last night. They are a nice couple.

(19) We believe her, the court does not, and that resolves the matter. (NY Times, 5/24/ 00, reported by J. Gundel)

(20) For the Parks and millions of other young Koreans, the long-cherished dream of home ownership has become a cruel illusion.
For the government, it has become a highly volatile political issue. (Poesio and Vieira 1998)

The idea of discourse model, originally formulated by Karttunen (1976), was then developed by Sanford and Garrod (1981) and Garnham (2001) in psycholinguistics, and made more formal, among others, by Heim (1982) and Kamp (1981) in theoretical linguistics and Webber (1979) in computational linguistics.

The theories developed by Heim and Kamp collectively took the name of Discourse Representation Theory (DRT); DRT has become the best known linguistic theory of the semantics of anaphora, and has served as the basis for the most extensive treatment of anaphora proposed in linguistics, (Kamp and Reyle 1993) In DRT, a discourse model is a pair of a set of discourse referents and a set of conditions (statements) about these discourse referents:

$$\langle x_1..x_n, c_1..c_n \rangle$$

represented in the linear notation of Muskens (1996) as

$$[x_1 \ldots x_n | c_1 \ldots c_n].$$

For instance, suppose A addresses utterance (21a) to B in an empty discourse model[10]. Then according to DRT update algorithms such as those proposed in (Kamp and Reyle 1993; Muskens 1996), when we process this utterance, we update the existing discourse model with information contributed by this utterance: that a discourse-new entity, engine e3, has been mentioned (hence a discourse referent $x_1$ 'representing' that entity gets introduced in the discourse model); and that 'we' (speaker A and addressee B) are supposed to take $x_1$. This fact, as well as the fact that $x_1$ is an engine, are new conditions added to the discourse model. The resulting discourse model is as in (21b). Note in particular that interpreting a discourse-new nominal expression like *engine E3* results in a new discourse referent being added to the universe of discourse U. (Here and elsewhere we'll ignore illocutionary force and simply treat all such utterances as statements.)

(21) a. We're gonna take engine E3

b. $[x_1 | x_1 = e_3, \textbf{engine}(x_1), \textbf{take}(A + B, x_1)]$

This discourse model is the context in which the interpretation of the following utterance takes place. Say that (21a) is followed by (22a), which contains a pronoun and a discourse-new expression, *Corning*. The pronoun has only one interpretation in the discourse model in (21b)–as a discourse-old mention of discourse entity $x_1$. Interpreting utterance (22a) –i.e., establishing that an instruction to send engine E3 to Corning– leads to a second update of the discourse model; the resulting model is as in (22b) and contains, in addition to the discourse entities and the conditions already present

[10] An extreme abstraction!

in (21b), the new discourse entities $x_2$–the interpretation of pronoun *it* and which pronoun resolution identifies with $x_1$– and $x_3$, the discourse-new interpretation of proper noun *Corning* and interpreted as directly referring to (real world) object $corning$. The discourse model also includes new conditions on these entities.

(22) a. and shove it to Corning

b. $[x_1, x_2, x_3 | x_1 = e_3, x_2 = x_1, x_3 = corning, \textbf{engine}(x_1), \textbf{take}(A{+}B, x_1), \textbf{send}(A{+}B, x_2, x_3)]$

Associative references would instead be interpreted as introducing new discourse entities related by relations other than identity to existing discourse entities. Two key contributions of dynamic theories of anaphora developed in formal linguistics have been to show that the construction of such discourse models can be characterized in a formal way, and that the resulting interpretations can be assigned a semantics just as in the case of interpretations proposed for other semantic phenomena. The original approach to discourse model construction proposed by Heim (1982) and Kamp (1981) – and later spelled out in painstaking detail by Kamp and Reyle (1993) – was highly idiosyncratic, but later work demonstrated that the methods of syntax-driven meaning composition used in mainstream formal semantics can be used to develop a theory of discourse model construction as well (Heim 1983; Rooth 1987; Groenendijk and Stokhof 1991; Muskens 1996).

These formal approaches to discourse model construction center around the idea of **file card**. According to Heim (1983), a discourse model can be seen as a collection of file cards, each representing the information about a single discourse entity introduced in the discourse. More precisely, in most recent versions of DRT, mentions of referring expressions are interpreted as follows:

**indefinite (a P, some P):** a new file card $x_i$ is added to the discourse model and asserted to be of type **p**. This update is formally written $[x_i, |\textbf{p}(x_i)]$.

**proper nouns:** as a result of a reference to object *b* via a proper name, a new file card $x_i$ is added to the discourse model and asserted to be identical with *b*. This update is formally written $[x_i, |x_i = b]$. (See for instance proper name *Corning* in (22).)

**pronouns:** a new file card $x_i$ is added to the discourse model and noted as needing resolution via the condition $x_i =?$. This update is formally written $[x_i, |x_i =?]$. Resolution leads to this condition being replaced with an equality with the file card of the anchor. (See for instance pronoun *it* in (22).)

**definite nominals (the P, that P):** this is the type of referring expression on which there is the least agreement. Most researchers propose that definite descriptions have a **uniqueness presupposition**: the existence of an object of type P is presupposed instead of asserted, and furthermore this object is meant to be unique (Barker 1991; Roberts 2003). This semantics can be translated as follows: a new file card $x_i$ is added to the discourse model and asserted to be identical with the unique object of type **p** (in the context). This update is formally written $[x_i, |x_i = \iota y.\textbf{p}(y)]$.

Crucially for what follows, the file card for *x* contains all information that is known in the context about *x*. Thus for instance after reading the first sentence of example (23) our Universe of Discourse will contain an entity $x_i$ whose file card will contain the information that her name is Miss Watson, that she is the sister of the widow, that she is an old maid, etc etc.

(23) The widow's sister, Miss Watson, a tolerable slim old maid, with goggles on, had just come to live with her, and took a set at me now with a spelling-book. She worked me middling hard for about an hour, ...(from M. Twain, *Huckleberry Finn*).

The notion of file cards, or discourse entities, played a crucial role in subsequent work on anaphora resolution of the '80s and early '90s (Webber 1979; LuperFoy 1992; Vieira and Poesio 2000; Poesio and Kabadjov 2004) but then took a back seat to more primitive notions such as single anaphor-antecedent links, although it is now being revived, as we will see below.

A crucial feature of theories such as DRT is that DRSs are logic representations with their own truth conditions, different although equivalent to traditional first-order logic, and from which inferences can be made. For instance, (21b) is equivalent to the pseudo-existential statement that there is an object, this object is identical to $e_3$, and that A+B take this object. The existence of a deductive system over these representations is essential because many cases of anaphora resolution require complex inference, as we will see in a moment.

DRT has been used to develop accounts of a range of anaphoric phenomena beyond the simple case of nominal reference to antecedents introduced by nominals, covering reference to events as in (24a), to plurals as in (24b), or to more abstract objects such as propositions as in (24c).

(24) a. John met Mary. That happened at 3 o'clock.
b. John saw Mary. They had gone to school together.
c. John met Mary. This fact stroke him as strange ....

Kamp and Reyle (1993) and others provide detailed treatments of anaphora to events and plurals. Their treatment of reference to events is based on the assumption that events are individuals that introduce discourse referents in the common ground, as in (25).

(25) a. John met Mary. That happened at 3 o'clock.
b. $[x_1, x_2, e_1, x_3 | x_1 = john, x_2 = mary, e_1 : \textbf{meet}(x_1, x_2), x_3 @ 3pm, x_3 = e_1]$

By contrast, Kamp and Reyle's analysis of plurals, like that of most researchers in the area, is based on the assumption that resolving such references (i.e., finding an anchor for discourse entity $x_3$ in in (26b)) requires bridging inferences on the discourse model as a result of which the model is augmented with new objects. In the case of plurals, these new objects are sets or groups, such as new object $x_4$, defined as $x_1 + x_2$ in

(26c).[11] In the case of propositional references, these new objects are propositions. Resolving the discourse referent $x_3$ in (26e) requires introducing a new propositional variable $K_1$, as in (26f). As already discussed, one of the key claims of the discourse model hypothesis is that resolving anaphoric references in general requires inferences on the discourse model.

(26) a. John met Mary. They had gone to school together.

b. $[x_1, x_2, e_1, e_2, x_3 | x_1 = john, x_2 = mary, e_1 : \textbf{meet}(x_1, x_2),$
$e_2 : \textbf{gone-to-school-together}(x_3)]$

c. $[x_1, x_2, e_1, e_2, x_3, x_4 | x_1 = john, x_2 = mary, e_1 : \textbf{meet}(x_1, x_2),$
$e_2 : \textbf{gone-to-school-together}(x_3), x_4 = x_1 + x_2, x_3 = x_4]$

d. We believe her, the court does not, and that resolves the matter.

e. $[x_1, s_1, x_2, s_2 | s_1 : \textbf{believe}(we, x_1),$
$\textbf{court}(x_2), \neg s_2 : \textbf{believe}(x_2, x_1)]$

f. $[x_1, s_1, x_2, s_2, x_3, e_1, K_1 | s_1 : \textbf{believe}(we, x_1),$
$\textbf{court}(x_2), K_1 : [\neg s_2 : \textbf{believe}(x_2, x_1)], \textbf{matter}(x_4),$
$e_1 : \textbf{resolves}(x_3, x_4), x_3 = K_1]$

Even richer, if less formalized, models (usually called **mental models** instead of **discourse models** were proposed in psycholinguistics on the basis of work by Bransford *et al.*, Garnham, and Sanford and Garrod, among others (Bransford et al. 1972; Sanford and Garrod 1981; Garnham 2001). Such models are assumed to encode the results of rich inference and to be more distant from language than the models usually assumed in computational and theoretical linguistics.

## 2.5 Statistics about Anaphora from Corpora

Statistics from anaphorically annotated corpora can give a rough quantitative indication of the relative importance of different types of nominal anaphoric phenomena.

Kabadjov (2007) reports several statistics about the relative frequency of different types of nominals in the GNOME corpus and the Vieira-Poesio corpus. The GNOME corpus (Poesio 2004) was designed to study local and global salience (Poesio et al. 2004c, 2006) and in particular, their effect on generation, including text structuring (Karamanis 2003), aggregation (Cheng 2001) and determining the form of referring expressions (Poesio 2000). It consists of texts from three different genres widely studied in NLG: museum labels, pharmaceutical leaflets, and tutorial dialogues.

The subset of the GNOME corpus analyzed by Kabadjov includes 3354 NPs, classified into 28 mutually exclusive types. The five most frequent types are bare-np, the-np and the-pn, pers-pro, pn and a-np, representing 22%, 18%, 10%, 10%, and 8% of the total, respectively.

Concerning the types of relations, the part of the GNOME corpus studied by Kabadjov includes 2075 anaphoric relations; of these, 1161 (56%) are identity relations, whereas the rest are bridging. Among the anaphors, 44% of all anaphors related to their antecedent by an identity relation are pronouns (of which 27% personal pronouns and

[11] *Bare* plurals like *dogs* (and bare singulars like *water*) are generally interpreted in semantics, following Carlson (1977), as involving references to **kinds** accompanied in some case by an existential quantification.

| Type of anaphoric expression | Percentage (of anaphors) | Percentage anaphoric | Source |
|---|---|---|---|
| **Pronouns** | 44% | | Kabadjov 2007 |
| Pers. Pronouns | 27% | 95% | Kabadjov 2007 |
| *it* | | 68–72% | Evans, Boyd *et al* |
| Poss. Pronouns | 17% | 95% | Kabadjov 2007 |
| **Definites** | 16% | 30%[Gnome] | Kabadjov |
| | | –40%[WSJ] | Poesio and Vieira |
| (first mention) | | 50%[WSJ] | Poesio and Vieira |
| (bridging) | | 10%[WSJ] | Poesio and Vieira |
| **Proper names** | 10% | 38% | Kabadjov 2007 |

Table 1: Anaphors and degree of anaphoricity in written text: Summary

17% possessive pronouns), 16% are definite descriptions, and 10% are proper nouns. Conversely, 97% of possessive pronouns are anaphoric, as are 95% of pers-pro, 38% of proper names, and 30% of definite descriptions.

The anaphoricity (or lack thereof) of pronouns has been studied in a number of papers concerned with detecting expletives. (Evans 2001) collected statistics from 77 texts from the SUSANNE and BNC corpus chosen to sample a variety of genres, and which contained 3171 examples of *it*. Of these, he classified 67.9% as being nominal anaphoric, 26.8% expletives, 2.2% used in idiomatic / stereotypical constructions, 2% discourse topic mentions, 0.8% clause anaphoric, 0.1% cataphoric. Very similar figures are reported by (Boyd et al. 2005), who studied expletives in text as well (the BNC sampler corpus). Of the 2337 instances of *it* in their corpus, 646 (28%) are expletives. Arguably the most careful analysis of the distribution of pronouns has been carried out by Müller (2008), who studied the distribution of third-person pronouns *it*, *this* and *that* in multi-party dialogue. Mueller asked his coders to classify these pronouns as either 'normal' (i.e. referring to either a nominal or clausal antecedent), 'extrapos-it' and 'prop-it' (two types of expletives), 'vague' (i.e., referring but without a clearly identifiable antecedent), 'discarded' (i.e., included in utterances that were not continued) and 'other'. For *it*, he found that of the around 1,000 cases in his corpus, about 62.5% were classified as referential (of which 57.8% were 'normal' and 4.7% 'vague') and 37.5% as either expletive or discarded (22% as 'discarded', 15.5% as expletive). He also observed however significant disagreements on the classification ($\kappa = .61$). These figures are summarized in Table 1.

The distribution of the referents of pronouns, and in particular whether they were introduced by NPs or more indirectly, was studied by (Passonneau 1993), Eckert and Strube (2001), Byron (2002), and Gundel et al. (2002). Eckert and Strube found that around 22% of the pronouns in their corpus (Switchboard) had a non-NP antecedent, whereas 33% had no antecedent at all. Byron reported that 16% of pronouns in her corpus had non-NP antecedents. Gundel *et al.* analyzed 2000 personal pronouns in the Santa Barbara Corpus of Spoken American English and found that 16% lacked an NP antecedent: around 5% had a non-NP antecedent, 4.5% were expletives, and 4.2% had what Gundel *et al.* call 'inferrable' antecedent, like *she* in the following example, that

refers to the mother of the kids just mentioned.

(27) [Talking about how the kids across the street threw paint in their yard.] Those kids are just - And she's pregnant with another one. (2.294)

An extensive study of the uses of definite descriptions was carried out by Poesio and Vieira (1998), who were in particular concerned with the percentage of definite descriptions that were first mention, as opposed to anaphoric. Poesio and Vieira carried out two experiments in which definite descriptions were classified according to two slightly different schemes. In both cases, they found that around 50% of definite descriptions were first mention, around 40% were anaphoric, and 10% bridging. However, Poesio and Vieira also raised the issue of agreement on classification, only finding reasonable agreement among their coders on the distinction between first mention and anaphoric ($\kappa = 0.76$) with finer distinctions leading to more disagreements, and the distinction between bridging and first mention in particular being difficult.

## 2.6 More Terminology

We conclude this first section with a bit more discussion on terminology. As we said above, we use the term **anaphoric** to indicate expressions whose interpretation depends on objects introduced in universe of discourse $U$ either by virtue of being explicitly mentioned (like *engine E3* in (21)) or by being inferred (as in the cases of plurals and propositional anaphora). As we said, in this survey we will primarily be concerned with these expressions and this characterization of the interpretation problem. However, quite a lot of other terms are used in the literature and there is a great degree of confusion about their use, so a few remarks on these issues are in order.

First of all, note that that this use of the term 'anaphoric'– although, we would argue, the most common in linguistics– is not the only use of the term. Many researchers use the term to indicate links at the *textual* level of representation (i.e., between expressions rather than with respect to discourse entities)–indeed, this seems to be the use of the term in the well-known (van Deemter and Kibble 2000). Other researchers use the term anaphora to indicate the study of pronominal interpretation, reserving the term coreference for the study of anaphoric reference via proper nouns.

Second, with the first MUC initiative the term **coreference** was introduced for a task which is closely related (although not identical with) the task of anaphoric resolution. As a result, the term coreference has become virtually synonymous with anaphora. Unfortunately, the term coreference has a technical meaning in linguistics, which has caused all sorts of discussions (van Deemter and Kibble 2000). To add to the confusion, the term coreference is used in different ways in formal linguistics and in functional linguistics.

As we saw earlier in this Section, in formal semantics the term 'reference' is used to indicate the relation between an expression of the language and an object in the world, if any: proper names are the typical example of expression which is referring in this sense. Two expressions are thus **co-referring** if they refer to the same object. However, not all expressions in the language, and not even all the nominal expressions that we called 'referring' earlier on, are referring in this sense, yet this does not prevent them serving as antecedents of anaphoric expressions. A typical example are expressions

occurring in hypothetical or negated contexts, as shown in the examples in (28) (Partee 1972): neither the hammer mentioned in (28a) nor the car mentioned in (28a) exist, yet they can happily serve as antecedents of anaphoric expressions.

(28) a. If I had a hammer I would use it to break your head.
b. I can't buy a car - I wouldn't know where to put it.

Vice versa, there are expressions which are coreferent but are not anaphoric in the sense discussed above –e.g., references to *Barack Obama* in distinct conversations, or in distinct documents, are co-referring (the term used in the case of documents is **cross-document coreference**) but not anaphoric (because distinct universes of discourse are built during each conversation).

This distinction between coreference and anaphora is the reason why computational linguists have generally preferred to avoid the term coreference and introduce other ones (Sidner 1979; van Deemter and Kibble 2000). We should however note that in other types of linguistics–particularly in systemic functional grammar and related functional frameworks–the term coreference is used in an entirely different manner (Halliday and Hasan 1976; Gundel et al. 1993). In these frameworks, there is no notion of 'reference to the world': all we can do is to refer to objects in our cognitive state–i.e., discourse referents–and therefore the term 'coreferring' is synonymous with 'anaphoric' in the sense here. (The use of the term 'referring expression' in the sense used in this Section comes from this tradition, via NLG.)

As the CL use of the term coreference is here to stay, anticipating some of the issues discussed in the second Part of this survey we will note here that the 'coreference task' as defined by the MUC guidelines (Hirschman 1998) is not the same as coreference either in the sense of formal semantics or in the sense of functional linguistics.

Given the focus on applications, most instantiations of the 'coreference task' concentrate on entities of a restricted number of semantic classes frequently occurring in newspaper text (persons, organisations, locations, events, vehicles, weapons and facilities in the case of the Automatic Content Extraction (ACE) effort, or include the marking of textual relations that would not necessarily be viewed as 'coreference' in linguistics. The most discussed example (van Deemter and Kibble 2000) is that of the relation between *John* and *a fool* in (29).

(29) John is a fool.

In linguistics, the relation is typically seen as one of predication–being a fool is viewed as a property of John, as discussed earlier in this section. In the MUC / ACE guidelines, the relation is marked as coreference. The problem is that coreference is generally taken to be transitive so these guidelines result in John, 'Mayor of Buffalo', and 'Senator for New York' being coreferent in (30).

(30) John was mayor of Buffalo last year and is now Senator for New York.

## 2.7 Summary of Section

The main points to keep made in this Section are as follows. First of all, we introduced some terminology, and in particular the notions of anaphora, reference, and corefer-

ence. Second, we pointed out that anaphoricity is a very general phenomenon, as the interpretation of most natural language expressions depends on the linguistic context to some extent. Computational Linguists however have primarily focused on the anaphoricity of nominal expressions, and so will we in this Survey. Third, we discussed how the notion of linguistic context has been formalized in terms of discourse models containing discourse entities, which are introduced by discourse-new expressions and referred to by discourse-old expressions. We also saw however that nominal expressions are not only used to refer to discourse entities; they can also be semantically empty, predicative, or quantificational. Last but not least, we pointed out that many of the judgments involved in determining the interpretation of anaphoric expressions, such as determining whether an expression refers to an entity not previously mentioned or not, are complex and substantial disagreements can be observed among coders.

# 3 The Interpretation of Anaphoric Expressions: Evidence from Corpora and Psycholinguistics

As illustrated by example (31), there is often more than one matching antecedent for anaphoric expressions in a discourse model, so that the 'one sense per context' heuristic, so successful in the case of wordsense disambiguation, does not work well for this case of ambiguity. Starting with the second sentence, there are two potential antecedents masculine in gender, that become three the next sentence if the system does not recognize that *the skipper of a minesweeper* is an apposition on *his father*). After the fifth sentence, a third potential antecedent appears, the sailor.

(31) Maupin recalls his mother trying to shield him from his father's excesses.
"Your father doesn't mean it,", she would console him.
When Maupin was born, his father was in the thick of battle, the skipper of a minesweeper.
He didn't see his son for two years.
He learned of his birth from a sailor on another ship, by semaphore.
"I got very sentimental about six months ago, and asked him to tell me exactly where he was when he found out." (From *The Guardian Weekend*, August 15th, 1998, p. 22.)

Interpreting anaphoric expressions–i.e., resolving this ambiguity–requires a combination of many different types of information, as illustrated by the example above. One of the strongest factors is gender: *she* in the second sentence is totally unambiguous. Commonsense knowledge can be an equally strong factor: clearly *Maupin* and *his father* cannot corefer if *his* is taken to have *Maupin* as its antecedent. Syntactic constraints also play a role: even if *his son* was replaced with *him* in sentence four (obtaining *he didn't see him*), coreference between subject and object would still be ruled out. Other types of disambiguation depend on factors that appear to behave more like preferences than hard constraints. For instance, the preferred interpretation for pronoun *He* at the beginning of the fourth sentence would seem to be Maupin's father rather than Maupin himself, but that preference appears to be more the result of the

preference for pronouns in subject position to refer to antecedents in subject position than a hard constraint or complex reasoning. The same motivation seems to justify the preference for pronoun *He* in the subject position of the following sentence. This difference between **constraints** and **preferences** plays an important role in many computational models of anaphora resolution and is also followed in standard expositions such as (Mitkov 2002) so we'll follow it here even though there is not conclusive evidence about the existence of two distinct mechanisms. In this section we will discuss these constraints and preferences and the psychological evidence in their favor; in the following sections we will discuss evidence coming from computational work.

## 3.1 Constraints

Much of the early linguistic work on anaphora focused on the identification of morphological and syntactic constraints on the interpretation of anaphoric expressions. Among these constrains the better known are **agreement** constraints (syntactic and semantic) and **binding** constraints. We'll discuss them in turn.

**Morphological constraints** Agreement constraints include gender, number and person constraints. We have an example of gender constraint in (31): *him* in the second sentence can only refer to Maupin or his father, not to his mother. The role of gender matching has been intensively studied in psychology (Ehrlich and Rayner 1983; Garnham et al. 1995; Arnold et al. 2000). Such studies demonstrated that gender affects disambiguation very early, and considered also the differences in gender use between languages with semantic gender such as English and languages with syntactic gender such as Italian or Spanish. As we will see below, most modern systems attempt to incorporate agreement constraints. The problems such systems encounter are that gender is not always used consistently–witness cases like (32), an error reported in (Tetreault 2001) but due to erroneous use of pronoun *it*:

(32) to get a customer's 1100 parcel-a-week load to its doorstep

Even when gender is not used erroneously, systems run into difficulties when pronouns are used to refer to entities referred to using uncommon proper names, as in the examples in (33).

(33) a. Maja arrived to the airport. [Maja a man] He ...
b. John brought Maja to the airport. [Maja a small dog] It ...

This second problem can be in part addressed by attempting to infer the gender of unknown names (Ge et al. 1998; Bergsma 2005) but more in general it is clear that people can often infer gender from context (see (Cornish 1986) and other references mentioned by (Garnham 2001), p. 67).

There has been much less work on examining number, but several studies have compared the relative difficulty of interpreting plural and singular anaphoric references (e.g., (Gordon et al. 1999)), and Clifton and Ferreira (1987) showed that plural pronoun *they* was equally easy to read following a conjoined noun phrase (*Bill and Sue met*) than when the antecedents were syntactically divided (*Bill met Sue*) suggesting that the

antecedent for the plural pronoun was found in the discourse model instead of in the syntactic representation. In Computational Linguistics, the main problem with number are nouns which are syntactically singular but semantically plural such as *the Union* in (34).

(34) The Union said that they would withdraw from negotiations until further notice.

**Syntactic constraints** The study of constraints on anaphoric reference played an important role in the development of modern generative linguistics, to the point of giving the name to one of its best-known paradigms, Government and Binding theory (Chomsky 1981). The aim of this work was understanding why *him* cannot corefer with *John* in (35a) (the asterisk indicates that the sentence is ungrammatical under the interpretation specified by the indexing) whereas *himself* must obligatorily be interpreted as referring to *John* in (35b).

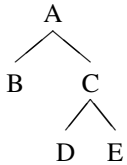
(35) a. *John$_i$ likes him$_i$.
b. John$_i$ likes himself$_i$

Langacker (1969) proposed an account based on a relation that he called **command** holding between nodes in a syntactic tree. The definition of the relation was subsequently refined by Lasnik (1976) and then by Reinhart (1976), who introduced the **c-command** relation, defined as follows:

**Definition 1** *Node A c-commands node B iff*

1. *A* $\neq$ *B*
2. *A does not dominate B and B does not dominate A, and*
3. *every X that dominates A also dominates B.*

For instance, in the following tree, A does not c-command anything, B c-commands C, C c-commands B, D c-commands E, and E c-commands D.




The c-command relation is at the heart of the classic definition of what is now called the **binding theory** due to (Chomsky 1981), which is articulated around three Principles. Principle A specifies constraints on reflexives and reciprocals (rather misleadingly called 'anaphors'), and says that they must have a c-commanding antecedent in their **governing category** (the smallest clause or noun phrase in which they are included). Principle B states that pronouns cannot have an antecedent in this governing category. Together, Principles A and B claim that reflexives and pronouns are in complementary distribution. Finally, Principle C states that R-expressions –proper names and nominals–cannot have c-commanding antecedents.

Binding theory subsequently underwent numerous revisions to address empirical limitations of the 1981 version. In (Chomsky 1986) the alternative notion of **m-command** was introduced. In HPSG, an alternative definition of **o-command** was introduced based on argument structure instead of phrase structure (Pollard and Sag 1994), to account for exceptions to binding theory in so-called picture NPs, as in (36).

(36) John was going to get even with Mary. That picture of himself in the paper would really annoy her, as would the other stunts he had planned.

But perhaps the main development was the proposal by (Reinhart and Reuland 1993) that some reflexives are **logophors**, i.e., have discourse antecedents–examples being cases like *himself* in (37a), which is grammatical, in contrast with the ungrammaticality in (37b).

(37) a. Bill$_i$ told us that Elisabeth had invited Charles and himself$_i$
b. * Bill$_i$ told us that Elisabeth had invited himself$_i$

Substantial experimental testing of binding constraints has been carried out over the years. Nicol and Swinney (1989) using a priming technique found that only associates of *the doctor* would be primed by *himself* in (38a), whereas only associates of *the skier* would be primed by *him* in (38b).

(38) a. The boxer told the skier that the doctor for the team would blame himself for the recent injury.
b. The boxer told the skier that the doctor for the team would blame him for the recent injury.

Gordon and Hendrick (1997) found broad support for Principles A and B of binding theory but poor support for Principle C. Runner et al. (2003) found confirmation that many reflexives in picture NPs behave like logophors.

**Semantic constraints** The main semantic constraint on anaphoric reference is the so-called **scope constraint**, that prevents anaphoric reference to antecedents introduced in the scope of downward-entailing operators (Karttunen 1976). Thus, in (39a), the reference in the second sentence to the car introduced in the scope of a negation is claimed to be infelicitous. In (39b), the car can be referred to within the conditional, but now outside it. (39c), illustrates that anaphoric reference to indefinites in the scope of modals is problematic (Karttunen 1976; Roberts 1989).

(39) a. John doesn't have a car. * It is in the garage.
b. If John has a car, he doesn't use it much. * Let's drive it around the park.
c. A wolf might have come in. *It ate John first. (Roberts 1989)

Semantic constraints have recently become the object of interest among psycholinguists because ERP experiments[12] is showing that examples of anaphoric reference like

[12]The experimental paradigm of *event-related potentials* look for correlations between text that subjects read and brain activity as measured by EEG.

those in (39) result in so-called 'semantic' violation effects (i.e., N400 effects) - see, e.g., (Dwivedi et al. 2006) for such effects in cases like (39c).

## 3.2 Preferences

Linguistic constraints by themselves do not completely eliminate anaphoric ambiguity. None of the constraints discussed above would prevent interpreting *him* in the second sentence of (31) as referring to Maupin's father. Neither do these constraints rule out interpreting *He* in the fourth sentence as referring to Maupin instead of his father. Yet these interpretations are clearly **dispreferred**. Much research has been carried out on the factors determining such preferences.

**Commonsense knowledge** One such factor is plausibility based on commonsense knowledge. One of the best known illustrations of the effect of plausibility is the minimal pair in (40), due to Winograd and also reported in (Sidner 1979). The only difference between (40a) and (40b) is the verb in the second clause, but that change is sufficient to change the preference from the council (in (40a)) to the women (in (40b)).

(40) a. The city council refused the women a permit because they feared violence.
b. The city council refused the women a permit because they advocated violence.

One type of plausibility effect intensively studied in the literature is the so-called **implicit causality** effect (Garvey and Caramazza 1974; Stevenson et al. 1994). Garvey and Caramazza (1974) observed that subjects, when asked to write a continuation to a sentence like (41), would tend to continue in a way consistent with *he* being Bill (i.e., by assuming that the *because* clause explains why Bill is to blame.

(41) John blamed Bill because he ...

Stevenson et al. (1994) found that these preferences are affected by the thematic structure of the verb (so that agent-patient verbs behave differently from experience-stimulus ones) and by the connective.

In a forced choice experiment, Kehler et al. (2008) presented subjects with a short discourse and a question uncovering the subjects' interpretation of a pronoun in the second sentence, as in (42).

(42) Samuel threatened Justin with a knife, and he blindfolded Erin with a scarf.
*Who blindfolded Erin?*

Kehler *et al.* found that in discourses with one semantically coherent interpretation, this interpretation was chosen regardless of other salience factors, whereas in sentences where both interpretations were equally plausible, subjects' choice of interpretation more or less reflected general salience.

Another simple form of preference carried by verbs are so-called **selectional restrictions**: restrictions on the type of argument a verb may have. Their effect is shown by minimal pair (43), from Mitkov (2002). In (43a), the preferred antecedent for *it* is the computer, presumably because *disconnect* prefers an electric appliance. In

(43b), however, the preferred antecedent for *it* is the disk, because *copied* prefers an information-carrying device.

(43) a. George removed the disk from the computer and then disconnected it.
b. George removed the disk from the computer and then copied it.

Because of evidence such as that above the early models of anaphora resolution in CL concentrated on developing theories of commonsense reasoning (Charniak 1972; Wilks 1975; Hobbs et al. 1993), but there is clear evidence that other factors are at play as well. In (44a), one could argue that it's more plausible for Bill to know the combination of his own safe–yet the interpretation that has John as antecedent of *he* is clearly preferred. And if commonsense reasoning was the only factor determining anaphoric resolution, then (44b) should not be funny –the reason it is is that the preferred interpretation for *it* is as referring to the head rather than the bomb.

(44) a. John can open Bill's safe - he knows the combination (Hobbs 1979)
b. If an incendiary bomb drops near you, don't lose your head. Put it in a bucket and cover it with sand (Hirst 1981)

**Syntactic Preferences** The next factor obviously playing a role in anaphora resolution is syntactic structure and syntactic preferences. Corpus statistics suggest that in most English corpora, about 60-70% of pronouns occur in subject position, and of these, around 70% have an antecedent also realized in subject position. This preference for pronouns in subject position to refer to antecedents in subject position has been called **subject assignment** and has been extensively studied in psycholinguistics (Broadbent 1973; Crawley et al. 1990).

Researchers also observed a preference for object pronouns to refer to antecedents in object position, suggesting a preference for **parallel** interpretations (Sheldon 1974; Kameyama 1985). Parallelism effects were studied, among others, by Smyth (1994), who showed that the closer the syntactic function, the stronger the effect; and by Stevenson et al. (1995), who observed a similar phenomenon, but a much stronger preference for subject pronouns than for object pronouns (80% to 60%).

Researchers including Smyth and Stevenson and colleagues also hypothesized that parallelism might be semantic rather than syntactic in nature; this approach was developed by Hobbs and Kehler (1997), among others.

**Salience** Another factor that clearly plays a role in anaphora resolution is **salience**, at least in its simplest form of **recency**: generally speaking, more recently introduced entities are more likely antecedents. Hobbs (1978) reported that in his corpus, 90% of all pronoun antecedents were in the current sentence, and 98% in the current or the previous sentence, although there was no fixed distance beyond which no antecedent could be found (one pronominal antecedent was found 9 sentences back). This importance of the antecedents in the current and previous sentence for pronouns has been confirmed by every study of referential distance , if with slightly different figures: e.g., Hitzeman and Poesio (1998) found that around 8% of pronoun antecedents in their corpora were not in the current or previous sentence. Distance is less important for other types of

anaphoric expressions: e.g., Givon (1992) found that 25% of definite antecedents were in the current clause, 60% in the current or previous 20 clauses, but 40% were further apart. Vieira (1998) found that a window of 5 was optimal for definites. This is true cross-linguistically (Givon 1983)

This is not to say, however, that choosing the most recently mentioned antecedent is an effective strategy. (Several studies suggest that this strategy would have mediocre results: e.g., Tetreault (2001) reports that choosing the most recent antecedent for pronouns that satisfies gender number and binding constraints would result in a 60% accuracy.) On the contrary, there is a lot of evidence for a **first mention advantage**–a preference to refer to first mentioned entities in a sentence (Gernsbacher and Hargreaves 1988; Gordon et al. 1993). Combined, these results provide support for a search strategy like that proposed by Hobbs (1978): going back one sentence at a time, then left-to-right.

A stronger version of the claim that there are differences of salience between entities is the hypothesis that attentional mechanisms of the type found in visual interpretation also affect the interpretation of anaphoric expressions. Authors such as Grosz (1977), Linde (1979), Sanford and Garrod (1981), and others have claimed that linguistic **focusing** mechanisms exist and play an important role in the choice of an antecedent for anaphoric expressions. Gundel et al. (1993) and others suggested that such mechanisms also affect production, and in particular, the choice of form of referring expression.

The best-known theory of this type is the framework proposed by Grosz and Sidner (1986) and articulated in two levels: the **global focus** specifying the articulation of a discourse into segments, and the **local focus** of salience specifying how utterance by utterance the relative salience of entities changes. That discourses are segmented according to 'topics' or the episodic organization of the story is widely accepted and backed up by evidence such as that presented by Anderson et al. (1983). Anderson and colleagues presented their subjects with a passage like in Figure 1, introducing a main character (in this case, female) and a secondary character (in this case, male) tied to the scenario. This first passage was followed either by a sentence expressing immediate continuation of the episode (*Ten minutes later . . .* ) or by one indicating that the story had moved on (*Ten hours later . . .* ). Finally, the subjects were presented with either a sentence referring to the main entity, or to one referring to the scenario entity. Anderson *et al.* found an entity x delay effect: after the sentence expressing immediate continuation there was no difference in processing a pronoun referring to the main entity or a pronoun referring to the scenario entity, but when the text indicated a longer delay (and hence, a closure of the previous episode) the pronominal reference to the scenario entity was harder to process.

Grosz and Sidner (1986) add the further hypothesis that this segmentation is hierarchical and that it is parasitical upon the intentional structure of the discourse–the intentions that the participants are trying to achieve. Grosz and Sidner proposed that the global focus is like a stack; by contrast, Walker (1998) proposes a cache model. The two models were evaluated by Poesio et al. (2006) in terms of the way they limit accessibility. Knott et al. (2001) argued that the intentional structure proposed by Grosz and Sidner, while perhaps appropriate for task-oriented dialogue, is not appropriate for many types of text.

AT THE CINEMA

Jenny found the film rather boring.
The projectionist had to keep changing reels.
It was supposed to be a silent classic.
a. Ten minutes later the film was forgotten
Ten hours later the film was forgotten
b. She was fast asleep
c. He was fast asleep

Figure 1: The materials from (Anderson et al. 1983)

The second level of attention is the so-called **local focus**. According to Grosz and Sidner and other researchers including Linde, Garrod and Sanford, and others, at every moment during a conversation or while reading text some entities are more salient than the others and are preferred antecedents for pronominalization and other types of anaphoric reference. Sidner (1979) proposed the first detailed theory of the local focus, articulated around two distinct foci: the **discourse focus**, meant to account for the phenomena normally explained in terms of the notion of 'discourse topic' (Gundel 1974; Reinhart 1981; Vallduvi 1993) is usually introduced. In (45), the meeting with Ira is the discourse focus and serves as privileged antecedent for certain types of anaphoric reference.

(45) a. I want to schedule a meeting with Ira.
b. It should be at 3p.m.
c. We can get together in his office

Sidner also introduces an **actor focus**, supposed to capture some of the effects accounted in previous theories through subject assignment, such (46).

(46) John gave a lot of work to Bill. He often helps friends this way.

According to Sidner, the local focus changes after every sentence as a result of mention and coreference. Extremely complex algorithms are provided for both foci and for their use for anaphoric reference.

Centering theory (Grosz et al. 1995) was originally proposed as just a simplified version of Sidner's theory of the local focus (Grosz et al. 1983) but eventually it evolved in a theory of its own –in fact, the dominant paradigm for theorizing about salience in computational linguistics and, to some extent, in psycholinguistics and corpus linguistics as well (see, e.g., the papers in Walker et al. 1998). According to Centering, every **utterance** updates the local focus by introducing new **forward looking centers** (mentions of discourse entities) and updating the focal structure. Forward looking centers are **ranked**: this means that each utterance has a most highly ranked entity, called **Preferred Center** (CP), which corresponds broadly to Sidner's actor focus. In addition, Centering hypothesizes the existence of an object playing the role of the discourse topic or discourse focus: the **backward looking center**, defined as follows:

**Constraint 3** CB($U_i$), the **Backward-Looking Center** of utterance $U_i$, is the highest

ranked element of CF($U_{i-1}$) that is realized in $U_i$.

Several psychological experiments have been dedicated to testing the claims of Centering, and in particular those concerning pronominalization, known as Rule 1:

**Rule 1** If any CF in an utterance is pronominalized, the CB is.

Hudson-D'Zmura and Tanenhaus (1998) found a clear preference for subjects, which could however also be accounted for in terms of subject assignment. Gordon and colleagues carried out a series of experiments that, they argued, demonstrated certain features of the theory. Gordon et al. (1993), for instance, revealed a **repeated name penalty**–a preference for avoiding repeating full names when an entity is mentioned in subject or first mention position, and using pronouns instead. Thus for instance Gordon *et al.* found an increase in reading time when processing sentences b–c of (47), with respect to reading sentences b–c of ex:RNP:2 in which the proper name in subject position *Bruno* has been replaced by pronoun *He*.

(47) a. Bruno was the bully of the neighborhood.
b. Bruno chased Tommy all the way home from school one day.
c. Bruno watched Tommy hide behind a big tree and start to cy.
d. Bruno yelled at Tommy so loudly that the neighbors came outside.

(48) a. Bruno was the bully of the neighborhood.
b. He chased Tommy all the way home from school one day.
c. He watched Tommy hide behind a big tree and start to cy.
d. He yelled at Tommy so loudly that the neighbors came outside.

Poesio et al. (2004c) carried out a systematic corpus-based investigation of the claims of Centering, that revealed among other things that entity coherence between utterances is much less strong than expected, so that the majority of utterances do not have a CB. Gundel et al. (1993) proposed an account of the factors affecting the choice of NP based on a theory of salience with some similarities to Centering but also some important differences. Gundel *et al.* argued that the choice of NP form is the result of a process that, among other factors, takes into account the **cognitive status** of the entities being referred. Gundel *et al.*'s theory distinguishes several levels of 'givenness', including **in focus**, **activated**, **familiar** and several levels of lexical acquaintance. 'Activation' corresponds to Grosz and Sidner's implicit focus, and 'in focus' is related to the notion of CB and CP, except that more than one entity may be in focus and there may also be no entity in focus (for the relation between Gundel *et al.*'s theory and Centering see (Gundel 1998; Poesio and Modjeska 2005)).

In addition to these **discrete** models of salience, **activation-based** models have also been proposed in which there is no fixed number of foci, but in which all entities have a level of activation (Klapholz and Lockman 1975; Alshawi 1987; Lappin and Leass 1994; Strube 1998; Tetreault 2001).

Models that integrate salience and commonsense knowledge have also been proposed, such as Carter's (Carter 1987). Carter combined Sidner's theory of focus with Wilks' causal reasoning. Among psychologists, the interaction of Centering with commonsense preferences has been studied by Gordon and Scearce (1995), who found

evidence that pronouns are interpreted according to Centering first and only later is commonsense knowledge used.

### 3.3 Perceived Ambiguity

In most cases, the combination of constraints and preferences is sufficient to ensure that anaphoric expressions have a single most preferred interpretation in context. This is now always the case, however. Consider again example (2), repeated here for convenience. Experiments with multiple coders (Poesio and Artstein 2005) suggest that coders systematically disagree on the interpretation of pronoun like *it* in 5.4, which ambiguously refer to one of two objects which has been joined together.

(2)

| | | | |
|---|---|---|---|
| 1.1 | M | : | all right system |
| 1.2 | | : | we've got a more complicated problem |
| 1.4 | | : | first thing _I'd_ like you to do |
| 1.5 | | : | is send engine E2 off with a boxcar to Corning to pick up oranges |
| 1.6 | | : | uh as soon as possible |
| 2.1 | S | : | okay |
| 3.1 | M | : | and while it's there it should pick up the tanker |
| 4.1 | S | : | okay |
| 4.2 | | : | and that can get |
| 4.3 | | : | we can get that done by three |
| 5.1 | M | : | good |
| 5.3 | | : | can we please send engine E1 over to Dansville to pick up a boxcar |
| 5.4 | | : | and then send it right back to Avon |
| 6.1 | S | : | okay |
| 6.2 | | : | it'll get back to Avon at 6 |

### 3.4 Summary of Section

In this Section we discussed evidence about linguistics and psycholinguistics concerning the factors that affect anaphora resolution, including a variety of linguistic constraints, and preferences deriving from syntactic information, commonsense knowledge, and salience. We'll see in the next Section how this evidence led to the development, in the early years of Computational Linguistics, of algorithms (especially for pronoun resolution) incorporating these preferences, typically by using hand-coded rules. In Section 6 we'll see how more modern approaches to anaphora resolution use such information, and how such information is extracted and approximated.

## 4 Early Computational Models

Between the '60s and the mid '90s a great number of computational models of anaphora resolution were developed implementing the theories of the effect on anaphora of syn-

tactic, commonsense, and discourse knowledge discussed in the previous section–in fact, many of these ideas were introduced as part of the development of these models.

There are substantial differences between these models in terms of their theoretical assumptions (some models assume that anaphora resolution is entirely a matter of commonsense knowledge, others that it is almost entirely a matter of syntactic information) and their level of formality (some models are very linguistically and formally oriented, others are very much pragmatically oriented). However, they all have two aspects in common that set them apart from later work: (i) no large scale evaluation was attempted: the models were either purely theoretical, or the implementation was a proof of concept (the larger evaluation attempts, such as Hobbs', consider a few hundred cases); (ii) development was guided near-exclusively by the researcher's own intuitions, rather than by annotated texts from the targeted domain.

For this reason it makes sense to cover them all in a single section before moving on to more recent work starting with the Message Understanding Initiative, and the development of the first medium-scale annotated resources, which allowed researchers in the field to overcome these early limitations. Our discussion will be short and focusing on the main ideas introduced in this work, many of which still valuable (and not yet incorporated in recent work).[13]

## 4.1 Syntax-based Models and The Hobbs Algorithm

We saw in Section 3 that (morpho) syntactic information plays an important role both in filtering certain types of interpretation (gender, binding constraints) and in determining preferred interpretations (subject assignment, parallelism). Several algorithms have been developed that incorporate these types of syntactic knowledge for anaphora resolution, in particular for the resolution of pronouns.

**Hobbs** The earliest and best-known of these syntax-based algorithms is the pronoun resolution algorithm proposed by Hobbs (1978). This algorithm, still often used as a baseline, traverses the **surface parse tree** breadth-first, left-to-right, and then going backwards one sentence at a time, looking for an antecedent matching the pronoun in gender and number. (See Figure 1.)

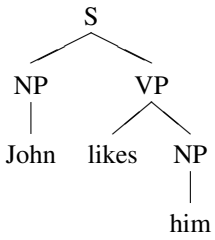
The algorithm incorporates both syntactic constraints, in particular from binding theory, and preferences, in particular subject and preference for first mentioned entities. Steps 2 and 3 ensure that no NP within the same binding domain as a pronoun will be chosen as antecedent for that pronoun, in that step 3 requires another NP o S node to occur in between the top node (node X) and any candidate: thus for example [NP John] will not be chosen as a candidate antecedent of pronoun *him* in example (49).

[13]For a more thorough introduction to early models before 1981, see (Hirst 1981). For the work between 1981 and the beginning of the empirical revolution, see (Mitkov 2002).

**Algorithm 1** Hobbs' Algorithm

1: Begin at the NP node immediately dominating the pronoun.
2: Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.
3: Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X.
4: **if** node X is the highest node in the sentence **then**
5: traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP is encountered, it is proposed as antecedent
6: **else**
7: (X is not the highest node in the sentence) continue to step 9.
8: **end if**
9: From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.
10: **if** X is an NP node and if the path p to X did not pass through the N node that X immediately dominates **then**
11: propose X as the antecedent
12: **end if**
13: Traverse all branches below node X to the left of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
14: **if** X is an S node **then**
15: traverse all branches of node X to the right of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered.
16: Propose any NP node encountered as the antecedent.
17: **end if**
18: Go to step 4

(49)




Because the search is breadth-first, left-to-right, NPs to the left and higher in the node will be preferred over NPs to the right and more deeply embedded, which is consistent both with the results of Gordon et al. (1993) concerning the effects of first mention and with the proposals and results of, e.g., Suri and McCoy (1994) and Miltsakaki (2002) concerning the preference for antecedents in matrix clauses.

Hobbs was possibly the first anaphora resolution researcher to attempt a formal evaluation of his algorithm. He tested it (by hand, apparently) with 100 pronoun examples from three different genres (a historical text, a novel, and a news article) achieving 88.3% accuracy, on the assumption of perfect parsing. Hobbs also claimed that with the addition of selection restrictions, his algorithm could achieve 91.7% accuracy. Several subsequent larger-scale evaluations showed that when perfect syntactic knowledge is available (i.e., using syntactically hand-annotated corpora) the algorithm is very competitive, if not quite as accurate as these initial figures would suggest. Lappin and Leass (1994) observed an accuracy of 82% over 360 pronouns from their corpus of computer manuals for their reimplementation of the algorithm. Tetreault (2001) found an accuracy of 76.8% over the 1694 pronouns in the Ge et al corpus of news text from the Penn Treebank, and of 80.1% over 511 pronouns from fictional texts. Hobbs' algorithm was also tested in a study by Matthews and Chodorow (1988), who found reading time evidence for a left-to-right top-down breadth-first search for antecedents.

## 4.2 Commonsense Knowledge: Charniak, Wilks, Hobbs' Abductive Model

The first computational models of the effect of commonsense knowledge on anaphora resolution go back to the very early days of CL (Charniak 1972; Winograd 1972; Wilks 1975); such research flourished in the 'knowledge-based' years of AI between the mid-70's and the mid-90's (Cohen 1984; Alshawi 1987; Carter 1987; Hobbs et al. 1993; Alshawi 1992; Poesio 1993; Asher and Lascarides 1998; Gardent and Konrad 2000). Some of this research set out to show that there was no need for syntax to account for anaphoric interpretation; a particularly explicit example of this was the work of (Charniak 1972), who developed a theory of language understanding based on the frame theory of commonsense knowledge (Minsky 1975), and as part of that explored the role of such knowledge in understanding references such as *the milk counter* in the following example.

(50) Jack was shopping at the supermarket. After getting a basket he went to the milk counter and picked up a carton of milk.

The use of frame and semantic network information in anaphora resolution was subsequently extensively investigated by Alshawi (1987) and, after the development of WordNet, by Poesio et al. (1997) and Harabagiu and Moldovan (1998).

Another researcher that ended up developing a theory of anaphora resolution while developing a general theory of semantic interpretation was Wilks (1975). Wilks developed **preference semantics**, a theory of language understanding according to which, again, syntax plays a very limited role. In Wilks' theory, all meanings are specifies in terms of a fixed number (around 70) of primitives (**semantic units**): entities, actions, roles, etc. In case of ambiguity, a sentence is assigned the interpretation which satisfies the greater number of preferences. Finally, commonsense reasoning –and, in particular, causal reasoning–is used to fill any remaining gaps.

The use of commonsense inference in anaphora resolution was greatly studied in the years between 1975 and 1995. Arguably, the most important contribution of this later work was the use of more formal (and more easily replicable) frameworks for inference. We will concentrate here on the best known among these later proposals, by Hobbs et al. (1993), but important theoretical contributions were made by Asher and Lascarides (1998); Gardent and Konrad (2000), and the group at SRI Cambridge who developed the Core Language Engine (Alshawi 1992)–the best known CL system used in actual applications yet incorporating full reasoning capabilities (in limited domains)–also developed anaphora resolution methods discussed in (Alshawi 1990).

Hobbs and colleagues proposed a theory of semantic interpretation based on a general defeasible interpretation mechanism based on **abduction**. Abduction is reasoning from effects to (the most plausible) causes: e.g., to conclude a friend must have woken up late in order to explain the observable fact that he hasn't showed up in time to go jogging in the morning. Hobbs and colleagues used abduction to develop accounts for a number of problems in language interpretation, ranging from interpreting noun noun compounds (e.g., to understand that *chess board* is a board to play chess whereas *wood board* is a board made of wood) to word sense disambiguation to anaphora resolution. The idea is that interpreting such expressions is akin to producing arguments for why these expressions were used.

For instance, let us suppose that we have the discourse

(51) John bought a new car. The engine is already acting up.

According to Hobbs and colleagues, identifying the link between the engine and the car amounts to constructing an 'explanation' for the second sentence on the basis of what is known from the first sentence:

$\exists x \mathbf{car}(x) \wedge \mathbf{buy}(j, x)$

and the following axiom, which can be used to infer the existence of an engine from the existence of a car.

$$\forall y \mathbf{car}(y) \supset \exists x \mathbf{engine\text{-}of}(x, y) \wedge \mathbf{engine}(x)$$

Given the right axioms, such methods can also provide formal accounts for the more complex reasoning involved, e.g., in example (52a).

(52) a. John hid Bill's car keys. He was drunk.

b. ?? John hid Bill's car keys. He likes spinach.

According to abduction theories, understanding discourse requires building an explanation tying the second utterance to the first–the lack of an easy explanation for the juxtaposition is what makes (52b) less felicitous. Such explanation would include abducing (i.e., assuming) that the second utterance is a reason for the first (i.e., they are tied by the **reason** rhetorical relation); that driving while drunk is dangerous, and therefore people may want to prevent other people from driving when they are in such condition; and that one way of preventing people from driving is to hide their car keys. Producing this explanation will also require assuming that *he* in the second utterance refers to Bill rather than John. All such assumptions have a cost; the lowest cost explanation will be chosen (Hobbs 1979; Hobbs et al. 1993); a modern presentation of the ideas can also be found in chapter 21 of Jurafsky and Martin (2009).

## 4.3 Salience: Discrete and Activation-based Models

The findings and hypotheses about salience discussed in Section 3.2, and in particular the original work by Grosz (1977), Linde (1979) and Sanford and Garrod (1981), motivated a great deal of work on computational models incorporating theories of salience (Sidner 1979; Reichman 1985; Kameyama 1985; Brennan et al. 1987; Alshawi 1987; Walker 1989; Lappin and Leass 1994; Suri and McCoy 1994; Walker et al. 1994, 1998; Strube 1998; Strube and Hahn 1999; Tetreault 2001).

**Sidner's Algorithm** The algorithms proposed in (Sidner 1979) remain to this day arguably the most detailed model not only of the effects of salience on anaphora resolution, but of anaphora resolution in general, although its accuracy is unclear given the lack of substantial evaluation. The central component of Sidner's theory is a discourse model with two key structural aspects:

- the organization of the entities in a semantic network inspired by the work of Charniak, although very few details about its organization are given (see discussion of Carter's work below)
- above all, data structures keeping track of which entities are currently most in focus. This aspect of the theory is the one which has had the greatest influence on subsequent research, in particular on the development of Centering (see next paragraph).

Sidner's theory of the local focus is articulated around three main data structures: the **discourse focus**–her implementation of the notion of 'discourse topic', see discussion in Section 3.2–the **actor focus**, accounting for the effects due to thematic role preferences or subject assignment; and a ranked list of the entities mentioned in the last sentence. In addition, stacks of previous discourse foci, actor foci, and sentence foci lists are maintained. Detailed algorithms are proposed to explain how each of these structures are updated as a discourse progresses.

Sidner proposed an extreme version of the 'bottom up' view of anaphora interpretation argued by psycholinguist: separate algorithms not just for each type of anaphoric

expression, but also depending on their (semantic) position–e.g., not only different algorithms for demonstrative and personal pronouns, but different algorithms for personal pronouns in agent position, non-agent position, and possessive position. There is clearly no space here for discussing these algorithms, but the key point is that each algorithm differs in the order in which the local focus structures are accessed. No evaluation of the theory was provided in the thesis apart from discussing how it would work with several examples, but an of evaluation was given by Carter (see below).

**Centering** Centering theory (Grosz et al. 1995) has been the theoretical foundation for a great deal of work both in anaphora resolution and in natural language generation (Dale 1992; Kibble and Power 2000; Karamanis et al. 2009). In anaphora resolution, the two best known algorithms based on Centering theory were developed by Brennan et al. (1987) and Strube and Hahn (1999).

The algorithm proposed by Brennan *et al.* (henceforth: BFP) takes as input utterance $u_n$ and updates the local focus by choosing the pair

$$\langle CB_n, [CF_n^1, \ldots, CF_n^m] \rangle$$

which is most consistent with the claims of Centering. This is done in a generate-filter-rank fashion:

1. Produce all possible $\langle CB_n, [CF_n^1, \ldots, CF_n^m] \rangle$ pairs. This is done by computing the CFs–which in turn involves generating all interpretations for the anaphoric expressions in utterance $u_n$–and ranking them.

2. Filter all pairs which are ruled out either by hard constraints (e.g., of the binding theory) or by the constraints of Centering (see Section 3.2): that if any CF is pronominalized, the CB is; and that the CB should be the most highly ranked element of the CF list of $u_{n-1}$ that is realized in $u_n$. The CFs are ranked according to grammatical function, with subjects ranking more highly than objects, and these than adjuncts.

3. Finally, the remaining pairs are ranked according to the preferences among transitions: namely, that maintaining the same CB as the most highly ranked (**continuing**) is preferred over maintaining the CB, but in less prominent position (**retaining**) which in turn is preferred over changing the CB (**shifting**).

The BFP algorithm has been extremely influential. Some of its features are grounded in solid empirical evidence–e.g., Poesio et al. (2004c) found very few exceptions for the preference for pronominalizing the CB if any other entity is pronominalized–but other characteristics found less empirical verification: e.g., there is little behavioral evidence for the preferences among transitions (Gordon et al. 1993) and real texts do not appear to be consistent with such preference either (Poesio et al. 2004c). BFP did not themselves provide an evaluation of the algorithm, but Walker (1989) evaluated it by hand comparing its performance for pronouns with that of Hobbs' algorithm, over the same texts used by Hobbs. The BFP algorithm performed slightly better than Hobbs' over the narrative texts (90% accuracy vs. 88%), whereas Hobbs' algorithm performed

slightly better over the task-oriented dialogues (51% vs. 49%) and clearly better with the news data (89% vs. 79%), the difference coming for Hobbs' algorithm preference for intrasentential antecedents, whereas the BFP algorithm tended to prefer intersentential ones. However, Tetreault's more extensive (and automatic) evaluation in (Tetreault 2001) suggests that the performance of Hobbs' algorithm is actually rather better than that of the BFP algorithm: Hobbs achieved 80.1% accuracy with fictional texts vs. 46.4% for BFP, whereas with news articles, Hobbs achieved 76.8% accuracy vs. 59.4% for BFP.

In the algorithm proposed by Strube and Hahn (1999), ranking by grammatical function is replaced by 'functional' ranking, i.e., ranking according to the taxonomy of given-new information proposed by Prince (1981): (hearer) old entities (i.e., anaphoric entities and entities referred to using proper names) are ranked more highly than 'mediated' (i.e., bridging) references, and these more highly than hearer-new entities. Strube and Hahn evaluated the performance of their algorithm by hand for both English and German, using both narrative and newspaper texts for a total of around 600 pronouns for each language, and comparing the accuracy with that of the BFP algorithm The performance using functional ranking was higher than using grammatical function ranking for both languages. For English, they obtained 80.9% accuracy as opposed to 76% for BFP, whereas for German, they achieved 83.7% with functional ranking vs. 74.8% with grammatical function ranking. The good performance of functional ranking was confirmed by the corpus study of Poesio et al. (2004c), which found that the parameter configuration with functional ranking was the one for which most of Centering's hypotheses were supported by the evidence.

**Graded Salience Models, Lappin and Leass** An alternative account of salience effects is centered around the notion of **activation**. Whereas Sidner's focusing theory and Centering account for salience effects by stipulating a discrete number of items in focus (the discourse focus, the CB, etc), activation-based models assume that every discourse entity has a certain level of activation on a graded scale (often values in the range $0 \ldots 1$), updated after every utterance, and that it is this level of activation that determines the likelihood of that entity being referred to. Activation-based models are less discussed, but in fact most commonly used in anaphora resolution systems than discrete models of salience.

The first known system of this type was proposed by Lockman and Klappholz (1980), but the best known models are the MEMORY system proposed by Alshawi (1987) (which also includes a detailed theory of semantic network use in anaphora resolution), and the RAP pronoun resolution algorithm proposed by Lappin and Leass (1994), that builds on Alshawi's work but includes several innovations, above all the first extensive treatment of expletives, and has become one of the best known pronoun resolution algorithms in CL. RAP also incorporates a sophisticated treatment of binding constraints.

Lappin and Leass's algorithm is another good example of the *generate-filter-rank* model of anaphora resolution. RAP takes as input the output of a full parser, and uses the syntactic information to filter antecedents according to binding constraints, specifically (i) antecedents of non-reflexives when the pronoun occurs in the argument,

adjunct or NP domain of the potential antecedent (e.g. *John$_i$ wants to see him$_{*i}$*, *She$_i$ sat near her$_{*i}$*, *John$_i$'s portrait of him$_{*i}$*), and (ii) non-pronominal antecedents that are contained in the governing phrase of the pronoun (*He$_i$ believes that the man$_{*i}$ is amusing*, *His$_i$ portrait of John$_{*i}$*). Reflexive pronouns are instead resolved to an antecedent that fulfills the binding criteria. Of all the candidates that pass the syntactic filter and are number and gender compatible with the pronoun, the one with the highest *salience weight* is selected, breaking ties by selecting the closest antecedent. Each mention receives an initial salience weight, consisting of:

- A *sentence recency* weight, which is always $100$.
- Additional weights for mentions not occurring in dispreferred position such as embedded in a PP (*head noun emphasis*, $80$), or in a topicalized adverbial PP (*Non-adverbial emphasis*, $50$).
- A weight depending on the grammatical function ($80$ for subjects, $50$ for direct objects, $40$ for indirect objects or oblique complements). Predicates in existential constructions also receive a weight ($70$).

The weight for each antecedent mention is halved for each sentence boundary that is between anaphor and then summed across all the members of the coreference chain of a candidate. To this salience value for the discourse entity, two local factors are added: one for parallelism of grammatical roles ($35$) and a penalty for cataphora ($-175$), which is applied to antecedent candidates that appear *after* the anaphoric pronoun.

Lappin and Leass evaluated RAP using 360 previously unseen examples from computer manuals. RAP finds the correct antecedent for 310 pronouns, 86% of the total (74% of intersentential cases and 89% of intrasentential cases). Without the combination of salience degradation and grammatical function/parallelism preferences, the performance gets significantly worse (59% and 64%, respectively), whereas other factors seem to have a much smaller impact (4% loss in accuracy for a deactivation of the coreference chains features, 2% loss for a deactivation of the cataphora penalty). By contrast, their reimplementation of Hobbs's algorithm achieves 82% accuracy on the same data.

Lappin and Leass use deep linguistic information in three places: firstly, to determine binding-based incompatibility and restrictions on the resolution of reflexives; secondly, to assign salience weights based on grammatical functions; thirdly, they use the parser's lexicon to assign the gender of full noun phrases. An approach based on shallow processing would have to approximate the syntax-based constraints based on the information in partial parses, and use a heuristic approach to reach full coverage for gender determination. Kennedy and Boguraev (1996) use a Constraint Grammar parser that determines morphological tags and grammatical functions and allows the identification of NP chunks, but does not yield enough information for constructing a complete tree, and report 75% resolution accuracy for news text, citing incomplete gender information and quoted passages as the most important source of errors.

**Strube / Tetreault** The algorithms proposed by Strube (1998) and Tetreault (2001) were inspired by Centering, but are in fact a version of the activation models in which activation scores (a partial order) are replaced by a list (a total order).

Tetreault's algorithm, Left-to-Right Centering (LRC), shown in 2, is the simplest and yet arguably the most effective algorithm inspired by Centering. It combines the idea of ranking of CFs from Centering with several ideas from Hobbs algorithm.

**Algorithm 2** Tetreault's LRC Algorithm

```
1:  for all U_n do
2:    parse U_n
3:    for all CF_i in the parse tree of U_n traversed breadth-first, left-to-right do
4:      if CF_i is a pronoun then
5:        search intrasententially in CF-partial(U_n), the list of CFs found so far in
          U_n, an antecedent that meets feature and binding constraints.
6:        if found matching antecedent then
7:          move to the next pronoun in U_n
8:        else
9:          search intersententially in CF(U_{n-1}) an antecedent that meets feature
            and binding constraints.
10:       end if
11:     else
12:       add CF_i to CF-partial(U_n)
13:     end if
14:   end for
15: end for
```

Tetreault evaluated his algorithm using a corpus of texts from two genres: news articles (a subset of the Penn Treebank containing 1694 pronouns annotated by Ge et al. (1998)), and fictional texts (also from the Penn Treebank, for a total of 511 pronouns). Tetreault also compared his algorithm with a variety of baselines, and with reimplementations of the BFP and Hobbs algorithms. On news articles, LRC achieved an accuracy of 80.4%, as opposed to 59.4% for BFP and 76.8% for Hobbs. On function, LRC achieved 81.1% accuracy, compared with 80.1% of Hobbs and 46.4% of BFP.

## 4.4 SPAR: Putting Syntactic, Commonsense and Focusing Preferences Together

The SPAR system proposed by Carter (1987) is arguably the most developed proposal in pronoun resolution prior to the data-driven revolution discussed in later sections. The main contribution of SPAR is not so much the development of novel algorithms, but the integration and implementation of several existing proposals, which led to a number of interesting observations concerning the interaction of, above all, focusing and commonsense knowledge, and resulted in several modifications. Carter used the pronoun rules from of Sidner's focusing algorithm for intersentential anaphora, and Hobbs' algorithm for producing a ranking to be used for intrasentential anaphora. He

used Wilks' preference semantics both to encode the semantic type of mentions and for causal reasoning.

The input to the algorithm is an analysis produced by Boguraev's English analyzer (Boguraev 1979). The algorithm then involves the following steps:

1. Compute selectional restrictions for pronouns depending on their syntactic position, using Wilks' semantics. (E.g., in *John ate it*, pronoun *it* would be restricted to an edible object.)

2. Use Sidner's rules to resolve each pronoun. Sidner's focusing rules are used to determine the order in which antecedents are tried (e.g., the actor focus is used first for pronouns in agent position). In Sidner's algorithm, however, the first antecedent that matches the pronoun syntactically, morphologically and semantically is chosen; in SPAR, all matching candidates are computed and maintained.

3. Binding constraints are used to eliminate inconsistent interpretations. For instance in the following text:

   John went to the doctor for a problem in his eye.
   He told him that an operation was needed.

   the two pronouns in the second sentence cannot both be resolved to *John*.

4. Causal inference is used if alternative interpretations are still possible for some pronouns–e.g., in cases like (44a).

5. Finally, a number of heuristics are used, some derived from Centering (prefer to pronominalize a discourse focus).

Carter evaluated SPAR using a corpus of 60 2-3 sentence stories. 40 of these stories (for a total of 65 pronouns) were written by himself to test specific features of SPAR; 20 (containing a total of 242 pronouns) by people who didn't know anything about the system. Carter claimed an accuracy of 100% with the stories he wrote himself, and of 93% with the other stories. We are not aware of any attempt at evaluating the system with a real corpus, or comparing it with other systems; however, many ideas from SPAR later found their way in the Core Language Engine (Alshawi 1992), for many years the most advanced NLP pipeline available.

## 4.5 Preliminary Discussion and Conclusions

This Section concludes the first part of our survey, in which we set the scene for the following discussion by first presenting the available linguistic and psychological evidence about anaphora and the resolution of anaphora, and then summarizing early work on anaphora resolution in which the relative importance of constraints and preferences, and the required information, were hand-coded.

In the next sections we will discuss how the creation of modern, substantial corpora of anaphoric information allowed an extensive empirical study of the phenomenon and

enabled the development of data-driven methods for anaphora resolution influenced by the theories that we discussed so far. Data-driven methods learn from annotated corpora how to combine, e.g., syntactic and lexical preferences in order to maximize performance on the data. Running such methods on large amounts of text, requires, however techniques for automatically and reliably extracting morphosyntactic knowledge, and large repositories of lexical and commonsense knowledge. Such techniques and repositories were not available at the beginning of the data-driven revolution, and as a result, simple approximations of the constraints and preferences discussed so far were used; but as we will see, more sophisticated techniques are becoming increasingly available, resulting in much more plausible models of the anaphora resolution process.

# 5 Towards an Empirical Approach to Anaphora Resolution: Developing an Experimental Setting

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 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, hosted two evaluations of coreference resolution systems, MUC-6 (Grishman and Sundheim 1995) and MUC-7 (Chinchor 1998), where annotated corpora were provided to the participants. In parallel with the development of the corpus, guidelines for the annotation of coreference were created and a common evaluation procedure for the comparative evaluation was thought out. The availability of these corpora, and of common evaluation metrics, made it possible to train and test anaphora 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 still seen in current evaluation campaigns such as the Automatic Content Extraction (ACE) initiative[14]. As a result, it is not an exaggeration to talk of a pre-MUC and post-MUC period in anaphora resolution research.

In this section we discuss some of the proposals concerning the annotation of corpora with anaphoric information and their use for evaluation of the data-driven approaches to coreference resolution.

## 5.1 Annotation Schemes for Anaphora

In a data-driven perspective, the design of the annotation scheme acquires a crucial importance. This is because linguistic data annotated with coreference information are used (1) to evaluate the performance of data-driven coreference resolvers (cf. Section

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

5.2), as well as (2) to train supervised systems, the most popular machine learning approach to this problem (cf. Section 6). So the annotation scheme defines what the problem of coreference is, and what is the linguistic phenomenon to be learned from the data. In this section we will briefly discuss some of the decisions taken in MUC and related initiatives, the most controversial issues, as well as the main subsequent developments.

**The MUC Annotation Scheme** One of the most influential, if less acknowledged, aspects of the annotation scheme developed for MUC (Hirschman 1998) was that it virtually defined the focus for research on anaphora for the last fifteen years. The scheme is focused on nominal coreference, where coreference is defined as *identity of reference*, i.e. whether two noun phases refer to the same object, set, activity, etc. Annotators were asked to mark all coreference relations involving two NPs, or a noun phrase and a nominal modifier, but no other type of relation (no identity of sense or bridging relation, for instance), and no relation where the anaphor or the antecedent are not both explicitly introduced as part of a noun phrase (i.e., no ellipsis, and no reference to implicitly mentioned objects as in discourse deixis).

One of the controversial issues in defining a coding scheme for anaphora is the definition of **markable** – also referred to as **mention**, i.e. *which* text constituents to chose as mentions of the entities. This definition depends on both syntactic and semantic factors. Syntactically, the coding scheme may require coders to mark the entire noun phrase with all postmodifiers, as in Example (53a), or just up to the head, as in Example (53b).

(53) 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".

Because of pre-processing errors such as e.g. parsing inaccuracies, the phrases annotated in the gold-standard and those automatically identified by a system can be partially misaligned, e.g. they differ on which phrases are postmodifiers of a noun. In order not to penalize coreference resolution systems on the incorrect identification of the markable boundaries, a solution was adopted in MUC where coders were instructed to mark the maximal span of noun phrases, as well as annotate its head in a separate MIN attribute. This way the systems can be also evaluated in a relaxed evaluation setting where they receive credit 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. However, since this assumption assumes an extra level of parsing annotations to be available, in subsequent proposals annotators were generally required to annotate the NP with all its modifiers (Poesio 2004; Poesio and Artstein 2008; Pradhan et al. 2007).

From a semantic perspective, a coding scheme can require coders to annotate mentions of entities of all types, or only of a subset of them. In the context of information

extraction applications, i.e. the general topic of the MUC and ACE campaigns, coreference resolution is most important for members of a small number of **semantic classes** that are relevant for the domain at hand. Indeed, many early machine learning approaches such as those of McCarthy and Lehnert (1995) and Aone and Bennett (1995), only concerned themselves with organizations and persons.

One potential 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. As a result, the coreference task in the ACE evaluation limits the consideration to persons, organizations, geopolitical entities (i.e., countries and regions with local government), locations, vehicles and weapons. But while being a reasonable simplification of the coreference task in an application-oriented setting, this approach leaves open the question of which entities to consider in different domains – e.g., the MUC types do not include any artifact, but these were key in one of the GNOME (Poesio 2004) domains, museum objects.

A particularly controversial aspect of the definition of the coreference task in MUC was the proposal to annotate apposition and copula constructions, which would normally not be considered cases of coreference but are important for information extraction. This drew criticism from researchers such as van Deemter and Kibble (2000), since the inclusion of intensional descriptions (as are the predicates in a copula construction) leads to counter-intuitive effects in cases such as the following one:

(54) [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. Later proposals for annotation schemes such as the MATE (Poesio 2004) and OntoNotes (Pradhan et al. 2007) guidelines distinguished between (transitive) coreference links and (directed, non-transitive) predication links. Some of these annotation schemes also allowed for other types of anaphoric relation (Passonneau 1997; Poesio 2004; Poesio and Artstein 2008; Pradhan et al. 2007).

A particularly difficult problem related to the issue of defining which markables to annotate is the treatment of **metonymy**, as in the following example:

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

The meaning of the example could be interpreted roughly as

(56) French government official made a statement to the effect of a French official position of disapproval regarding the "logic of ultimatums".

Given that NPs can be used metonymically, coreference annotation guidelines should specify whether the markable *Paris* in 55 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 annotation guidelines offer different (partial) solutions for this problem: The ACE guidelines resolve the ambiguity between 2 and 3 by assuming a semantic class for so-called *geopolitical entities* (GPEs), i.e., a conflation of a country, its government and its inhabitants. In the OntoNotes corpus, the diametrically opposite solution has been chosen: instead of merging different entities within a GPE semantic class, the annotation guidelines explicitly state that metonymies are to be distinguished from other uses of an NP, e.g. coreferential ones. Thus, in a document that contains the sentences:

(57) [$_1$ 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).

**Agreement** Given these problematic issues related to annotating coreference relations, it is essential to quantify the agreement between annotators. But while seminal efforts such as MUC reported agreement scores in terms of the MUC scoring metric (see subsection 5.2)[15], we note that more recent ones do not include a systematic study of the agreement between annotators (Artstein and Poesio 2008) – e.g., notably the ACE campaign. In contrast, in-depth studies of agreement on anaphoric annotations have been carried out by Poesio and Vieira (1998) and as part of the development of the GNOME and ARRAU (Poesio and Artstein 2008) corpora. Such results suggest that reasonable agreement can be obtained on the distinction between discourse old and discourse new, and that annotating bridging reference requires identifying very clearly the subset of bridging relations of interest.

**More Recent Coding Schemes** More recent coding schemes, such as the scheme developed for the GNOME corpus (Poesio 2004) and then for the English ARRAU corpus (Poesio and Artstein 2008) and ONTONOTES corpus (Pradhan et al. 2007), and those for the Dutch German TüBa-D/Z corpus Hinrichs et al. (2005b) and the Catalan / Spanish ANCORA corpus (Recasens and Martí 2009), tend to differ from the MUC / ACE schemes on the aspects we mentioned. The main difference is that all noun phrases are annotated, instead of only those referring to a limited number of types. All of these annotation schemes also distinguish between identity and predication, and some of these schemes also require annotation of associative relations (GNOME and ARRAU) or of some types of discourse deixis–e.g., reference to events in ONTONOTES. Finally, all modifiers are usually included in the markables. Agreement studies are generally carried out.

[15]The MUC-6 annotators reached an agreement level of $F_1$=0.83 (Hirschman et al. 1997), comparable with later efforts such as the German TüBa-D/Z corpus ($F_1$=0.83, Versley 2008), or the Dutch COREA corpus ($F_1$=0.76, Hendrickx et al. 2008) which relied on more refined annotation guidelines.

**Available Corpora** Table 2 summarizes the currently available anaphorically annotated corpora, with references to the main publications and sites with information. Ongoing efforts as part of the Anaphoric Bank initiative[16] aim at making some of these anaphorically annotated corpora available in compatible markup formats. Some data are also available from the SEMEVAL-2010 site.[17]

| Language | Name | Reference | Size (words) |
|---|---|---|---|
| Arabic | ACE-2005[18] | Walker et al. (2006) | 100k |
| | OntoNotes 3.0[19] | Weischedel et al. (2008) | 200k |
| Catalan | AnCora-CO-Ca[20] | Recasens and Martí (2009) | 300k |
| Chinese | ACE-2005 | Walker et al. (2006) | ≈200k |
| | OntoNotes 3.0 | Weischedel et al. (2008) | 1224k |
| Dutch | COREA[21] | Hendrickx et al. (2008) | 325k |
| English | MUC-6[22] | Grishman and Sundheim (1995) | 30k |
| | MUC-7[23] | Chinchor (1998) | 30k |
| | GNOME[24] | Poesio (2004) | 50k |
| | ACE-2005 | Walker et al. (2006) | 400k |
| | NP4Events[25] | Hasler et al. (2006) | 50k |
| | OntoNotes 3.0 | Weischedel et al. (2008) | 1150k |
| | ARRAU 1.0[26] | Poesio and Artstein (2008) | 300k |
| French | DEDE (definite descriptions)[27] | Gardent and Manuélian (2005) | 50k |
| German | Potsdam Commentary Corpus[28] | Stede (2004) | 33k |
| | TüBa-D/Z [29] | Hinrichs et al. (2005b) | 600k |
| Italian | Venex | Poesio et al. (2004a) | 40k |
| | i-Cab[30] | Magnini et al. (2006) | 250k |
| | LiveMemories 1.0 | Rodriguez et al. (2010) | 250k |
| Japanese | NAIST Text Corpus[31] | Iida et al. (2007b) | 38k sentences |
| Spanish | AnCora-CO-Es | Recasens and Martí (2009) | 300k |
| Tibetan | Tusnelda (B11) | Wagner and Zeisler (2004) | <15k |

Table 2: Referentially annotated corpora in different languages

[16] http://www.anaphoricbank.org
[17] http://stel.ub.edu/semeval2010-coref/
[18] http://projects.ldc.upenn.edu/ace/data/
[19] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T24
[20] http://clic.ub.edu/ancora/
[21] http://www.clips.ua.ac.be/~iris/corea.html
[22] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T13
[23] http://ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001T02
[24] http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/
[25] http://clg.wlv.ac.uk/projects/NP4E/#corpus
[26] http://cswww.essex.ac.uk/Research/nle/arrau/
[27] http://www.cnrtl.fr/corpus/dede/
[28] http://www-old.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.en.html

## 5.2 Evaluating Coreference Resolution Systems

A common question with both algorithms and systems for computational anaphora resolution is how ell these perform in comparison with each other. Many of the earlier approaches for pronoun resolution rely on accuracy as the main evaluation measure, i.e. the ratio of correctly resolved pronouns to the total number of pronouns to resolve. However, as Mitkov (2000) and Byron (2001) point out, a solid evaluation setting for such systems should be able to answer the following questions:

- Does the evaluation quantifies the performance of the resolution algorithm only (i.e., assuming perfect pre-processing, including agreement features like number or gender) or rather of the whole system, where pre-processing steps such as parsing and determination of gender features are done automatically?
- Does the evaluation include or exclude difficult cases such as first-person pronouns (which may not be resolvable to an antecedent), cataphora, cases of expletive pronouns, or pronouns and demonstratives that refer to clauses instead of noun phrases?
- What type of texts is the evaluation carried out on, as technical manuals seem to be easier to treat with pronoun resolution than newspaper text?

These two latter points seem to be less of an issue when considering results from systems using standard corpora from MUC or ACE. Nevertheless, we note that the interpretation of quantitative results can still be subtle, even when using standard datasets, since different evaluation metrics and evaluation conditions could have been used. For example, as (Stoyanov et al. 2009) point out, the practice of resolving only those NPs that are marked in the annotated corpus (and thus are part of a gold-standard coreference chain) severely distorts the evaluation results and it can lead to overestimate the performance by a very large margin in comparison with the same system resolving automatically extracted markables.

In the following we present an overview of the main evaluation measures developed to quantify the performance of coreference resolution systems. These can be broadly classified into three main classes: link-based measures (Section 5.2.1), set-based measures (Section 5.2.2) and alignment-based measures (Section 5.2.3).

### 5.2.1 Link-based Measures

The simplest way of evaluating algorithms for anaphora resolution is let the system choose an antecedent for each markable and determine the **accuracy** of these decisions, i.e., how many of them are correct. And because up until very recently most systems were based on the mention-pair model of anaphora resolution (see Section 6), in which the system has to decide whether two noun phrases are mentions of the same discourse entity, the simplest way of evaluating the correctness of the system's decision

[29] http://www.sfs.uni-tuebingen.de/tuebadz.shtml
[30] http://www.celct.it/projects/icab.php
[31] http://cl.naist.jp/nldata/corpus/

is **link-based** –check whether the mention chosen by the system as the last mention of the same entity is in fact the last mention in the gold standard. But by making these two simplifications we obtain an evaluation measure which is unsatisfactory in many respects.

We begin by noticing that, as in the case of information retrieval, accuracy may produce inflated assessments of the performance of a system, as typically only 30-40% of the markables are anaphoric. And it is not very informative for comparing the performance of systems with expressions that are sometimes anaphoric and sometimes non-anaphoric, such as definite noun phrases. Definite NPs like "*the city*" refer to an already introduced entity about 50% of the time, and accommodate a new referent about 50% of the time (Poesio and Vieira 1998). Therefore, one system might choose to resolve it and link the expression to a plausible antecedent, and another one might choose to treat it as non-anaphoric and not resolve it at all. Since there is a merit both to more conservative strategies that leave potentially ambiguous cases unresolved and to more greedy strategies that would resolve as many potentially anaphoric noun phrases as possible, it is sensible to replace accuracy with two distinct performance measures:

**precision** is the ratio of the number of correctly resolved anaphoric links to the total number of links that a system resolves;

**recall** is the ratio of the number of correctly resolved anaphoric links to the total number of anaphoric links in the annotated gold standard.

$$\text{Precision} = \frac{\#\text{correct}}{\#\text{resolved}} \qquad \text{Recall} = \frac{\#\text{correct}}{\#\text{wanted}}$$

A system that leaves potentially ambiguous cases unresolved would thus gain precision at the cost of recall, and, conversely, a system that tries to resolve as many potentially anaphoric noun phrases as possible would (potentially) gain recall at the cost of precision.

It is common to summarize precision and recall into a single evaluation measure, the harmonic mean of precision and recall, called **F-measure** ($\mathrm{F}_1$) following its introduction as evaluation measure for information retrieval by van Rijsbergen (1979). The harmonic mean of two numbers is well approximated by the arithmetical mean of these two numbers when they are close to each other; when the difference is large, the harmonic mean is closer to the minimum of the two numbers.

$$\begin{aligned} F &= \frac{2}{1/\text{Precision} + 1/\text{Recall}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \times \#\text{correct}}{\#\text{resolved} + \#\text{wanted}} \end{aligned}$$

### 5.2.2 Set-based Measures

Early versions of the MUC coreference task definition (up to MUC-6) did calculate precision and recall by comparing system and gold-standard links. But requiring a system to reproduce the very same links annotated in the gold-standard (and scoring links to other antecedents which are correct but not the closest as wrong) results in spurious errors. Consider an example such as

(58) [$_{A1}$ Peter] likes sweets.
[$_{A2}$ The boy] has always liked chocolate,
but at the moment, [$_{A3}$ he] is devouring it like mad.

and assume the gold annotation includes the links $\langle A_1{-}A_2,\ A_2{-}A_3\rangle$. Then a system that links *he*($A_3$) to *Peter*($A_1$) without including the link *The boy*($A_2$), producing the system response $\langle A_1{-}A_3\rangle$, would be scored as completely wrong, with no partial credit given (cf. Figure 2).

Evaluating a link as correct if the antecedent is *coreferent* with the anaphor, i.e. if they are both contained within the same equivalence class (i.e. the *same set*) of markables referring to the same entity, would solve this problem and count the link $\langle A_1{-}A_3\rangle$ as correct.[32]. However, this would still assign different scores to link structures that result in the same equivalence classes, as in the following example:

(59) [$_{A1}$ Peter] likes to have milk and cereal for breakfast, but was out of milk.
Fortunately, [$_{B1}$ the milkman] came by and
[$_{A2}$ he] was able to buy some milk.

Resolving both $B_1$(*The milkman*) and $A_2$(*he*) to $A_1$(*Peter*) would then get a better score than resolving $B_1$(*the milkman*) to $A_1$(*Peter*) and $A_2$(*he*) to $B_1$(*the milkman*), even though they both result in the same partition $\{\{A_1, A_2, B_1\}\}$ (cf. Figure 3).

Vilain et al. (1995) propose a solution to both of these problems by defining precision and recall statistics over equivalence classes. This measure was adopted for the MUC coreference task definition starting with MUC-6 and it has been later referred in the literature as the MUC evaluation measure. To compute the MUC score for a system's output, a partition of the markables in coreference chains is first viewed as a set of equivalence classes. For example, the gold-standard partition of figure 3 is the set $G$ of equivalence classes

$$G = \{\{A_1, A_2\}, \{B_1\}\}$$

i.e., the number of equivalence classes is $|G| = 2$. Then, the number of class members (the markables) is:

$$\left|\bigcup_{A\in G} A\right| = \Big|\{A_1, A_2, B_1\}\Big| = 3\ .$$

It is easy to see that the minimum number of anaphoric links required to specify a partition $S - \ell(S)$ (Figure 4) is the number of elements in a set minus the number of equivalence classes in $S$:

[32]This was the metric used in (Vieira and Poesio 2000)

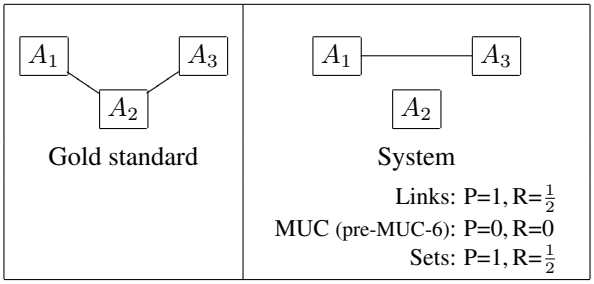



Figure 2: Link-based evaluation: Example (58) from Vilain et al. (1995).

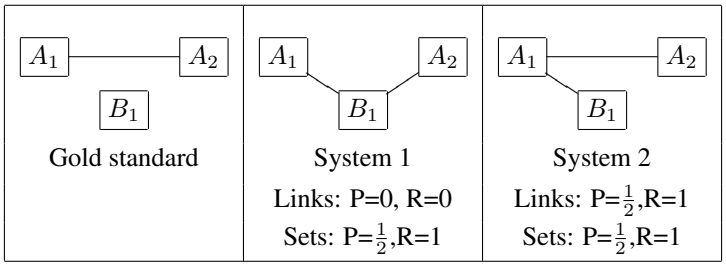



Figure 3: Link-based evaluation: Example (59).

$$\ell(S) := \left| \bigcup_{A \in S} A \right| - |S| = \sum_{A \in S} |A| - 1$$

which would give $3 - 2 = 1$ for the case of G, reflecting our intuition that in the gold standard in Figure 3 we have only one link, that which connects $A_1$ and $A_2$. Finally, let us define, for partitions $S_1$ and $S_2$, a partition $S_1 \cap S_2$ which contains the intersection of equivalence sets to $S_1$ and $S_2$:

$$S_1 \cap S_2 := \{A_1 \cap A_2 \mid A_1 \in S_1 \wedge A_2 \in S_2 \wedge A_1 \cap A_2 \neq \emptyset\} \ .$$

Given now a gold-standard partition $G$ and the partition $S$ induced by the system's output, we can define precision and recall for that system in a straightforward way:

$$P = \frac{\ell(G \cap S)}{\ell(S)} \qquad R = \frac{\ell(G \cap S)}{\ell(G)} \ .$$

So for instance, in the example in Figure 3, we have $G = \{\{A_1, A_2\}, \{B_1\}\}$ and $S_1 = S_2 = \{\{A_1, A_2, B_1\}\}$. This would give us $\ell(G) = 1$ and $\ell(S) = 2$ for both systems, and, with $G \cap S = \{\{A_1, A_2\}, \{B_1\}\}$, $\ell(G \cap S) = 1$, $P = \frac{1}{2}$ and $R = 1$ (see the "Sets" lines in Figures 2 and 3).

Note that the addition of singleton markables to a partition does not change its link count and, as a consequence, singletons in the gold-standard and system output have no effect on the precision and recall scores.

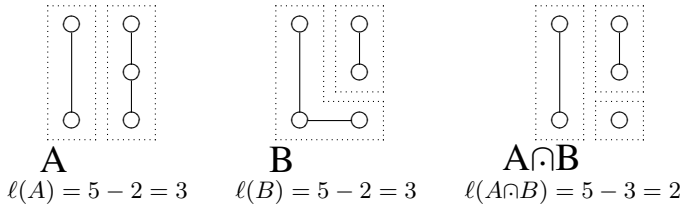



Figure 4: Examples for the $\ell$ (link count) function on partitions.

#### 5.2.3 Alignment-based Measures

The scoring method of Vilain et al. (1995) is designed to be an optimistic generalization of link-based measures to coreference sets, in the sense that the MUC scores are the best attainable scores for any decomposition into links of system and gold-standard partitions. This leads to counterintuitive effects on the small scale (misclassifying one markable into the wrong coreference set counts as one precision and one recall error, while completely merging two coreference sets counts as a single recall error), which are compounded when evaluating the system response on true (gold) mentions, where all singletons and non-referring mentions are removed. In this case, just merging *all* coreference chain simply incurs a number of precision errors of the number of coreference chain (minus one), whereas the number of correct links is evaluated as the total number of gold mentions (minus one), thus giving a trivial baseline with 100% recall and about 80% precision on the MUC-6 and MUC-7 datasets.

While this counter-intuitive result is not solely due to the MUC evaluation metric, there have been alternative proposals (Trouilleux et al. 2000; Luo 2005) to punish overmerging more strongly. Common to these proposals is the idea to induce an **alignment** between gold and system partitions by selecting those links such that (i) every coreference chain in the system's response corresponds to at most one coreference chain from the gold standard, and vice versa, and (ii) the highest weight among these assignments is reached.

Using a one-to-one alignment, overmerging (or any case where two coreference chains are completely merged, rather than having single elements misclassified) is punished more harshly in terms of precision (since the merged entity will only receive credit for covering one gold-standard coreference chain, and not the others), but also in terms of recall (since the merged entity will only be aligned to one gold-standard coreference chain and recall for the others will be zero).

Trouilleux et al. (2000) compute the weights of the alignment links as the weighted proportion between the sets of mentions that are in the intersection and the union of gold ($G_i$) and system ($S_j$) coreference chains (Dice coefficient):

$$\text{score}(G_i, S_j) = \frac{2|G_i \cap S_j|}{|G_i| + |S_j|} \ .$$

This score has the property that the sum of all possible links for one coreference chain is always at most one, also in the case where the score weights names, common noun phrases and pronouns differently – e.g. Trouilleux et al. (2000) propose a

weighting of 0.6 for names, 0.3 for common noun phrases, and 0.1 for pronouns. They then compare the number of "correct" links that are common to the aligned coreference chains with (i) the link count for the system's coreference chain, to get the precision, and (ii) the link count for the coreference chains in the gold standard, to obtain the number for recall.

In a similar fashion, the Constrained Entity-Alignment F-Measure (CEAF) metric proposed by Luo (2005) first computes an alignment and then compares the sets of mentions in the system's (for precision) or the gold standard coreference chains to the result of the alignment; to be counted as correct, a mention must be in both the system and gold-standard coreference chains linked by the alignment.

Let $\mathcal{G}$ be the set of entities (coreference chains) in the gold standard, and $\mathcal{S}$ the set of entities produced by a system. Let $G_m$ be the set of all alignments $g : \mathcal{G}_m \mapsto \mathcal{S}_m$ from subsets of size $m$ of the gold standard to subsets of size $m$ of the system's output, where $m = min(|\mathcal{G}|, |\mathcal{S}|)$. Then the best alignment between the gold standard and the system's output, according to Luo, is the alignment $g^*$ that maximizes the total similarity between coreference chains in the gold standard and coreference chains in the system's output–i.e., the assignment

$$g^* = argmax_{g \in G_m} \sum_{G \in \mathcal{G}_m} \phi(G, g(G))$$

where $\phi(G, g(G))$ is the similarity between coreference chain $G$ in the gold standard and the coreference chain in the system's output $g(G)$ to which $G$ is mapped by alignment $g$.

Defining the best similarity between gold standard and system output as

$$\Phi(g^*) = \sum_{G \in \mathcal{G}_m} \phi(G, g^*(G))$$

We can define precision and recall as follows:

$$P = \frac{\Phi(g^*)}{\sum_i \phi(S_i, S_i)} R = \frac{\Phi(g^*)}{\sum_i \phi(R_i, R_i)}$$

The ACE evaluation measure (NIST 2002) while being driven by application considerations and thus using a multitude of weighting factors that have changed between successive evaluations and that we cannot discuss here, uses weighting both for computing an alignment and for the computation of the score (unlike Trouilleux et al. (2000), who only use it for computing the alignment); Luo remarks that the weighting, which puts an emphasis on named entities and de-emphasizes pronoun, means that name matching is overemphasized (just matching names assures a result above 86% when evaluated on gold mentions), while progress on pronoun resolution is poorly reflected in the score.

### 5.2.4 Comparing the metrics

Set based metrics such as MUC credit a system if it identifies (part of) a coreference set, and penalize it if it misses (part of) it. By contrast, alignment-based methods take a global view and measure how well the system succeeded in discriminating between

the various coreference chains. In Figure 5, we compare the set-based MUC score with the CEAF alignment-based metric, as well as an evaluation metric used for evaluating document clustering systems, **purity** (Solomonoff et al. 1998), which computes the maximum of the number of markables at the intersection of a given system and gold-standard coreference chain alignment. The results show that CEAF punishes overmerging both in terms of recall and precision, where MUC only involves a slight decrease in precision, and purity has a larger loss in precision, but (like MUC) no loss in recall. In the case of a single misclassified mention ("System 2" in the figure), we see that the MUC measure punishes this more harshly than the other measures as it counts links instead of markables.

| | | MUC | Purity | CEAF |
|---|---|---|---|---|
| Gold | $[A_1\ A_2\ A_3]\ [A_4\ A_5]$ | P R $F_1$ | P R $F_1$ | P R $F_1$ |
| System 1 | $[A_1\ A_2\ A_3\ A_4\ A_5]$ | 3/4 3/3 0.86 | 3/5 5/5 0.75 | 3/5 3/5 0.60 |
| System 2 | $[A_1\ A_2]\ [A_3\ A_4\ A_5]$ | 2/3 2/3 0.67 | 4/5 4/5 0.80 | 4/5 4/5 0.80 |

Figure 5: Behaviour of MUC-score, Purity and CEAF on different examples.

# 6 Modern Computational Approaches

The shift to using increasingly larger quantities of real linguistic data led the researchers in coreference resolution to conclusions similar to those arrived at by researchers in other areas of CL (Klavans and Resnik 1996).

First of all, these researchers came to realize that linguistic and ontological information can be hard to get at, or even a significant source of errors, in a system that has to work with automatically generated analyses on unrestricted text. This realization led to so-called "knowledge-poor" methods, which try to forego expensive or unreliable information and instead try to rely on features that are easy and reliable to get at.

While earlier approaches relied heavily on either domain knowledge, or (e.g., in the case of Hobbs' naïve algorithm) deep syntactic analysis, the later approaches that were concerned about getting useful performance on unrestricted text had to find a way around problems that hindered the applicability of these techniques. Relying on domain knowledge means that an extraordinarily high effort has to be spent in analyzing and encoding relevant facts and assertions when adapting the system to a different domain, even more so for the newspaper texts used in empirical evaluations that were hardly restricted to a limited domain. Even using syntactic analyses proved to be a hard problem, given that the state of the art in automatic parsing in the mid-nineties was not accurate enough for this. For example, in the approach of Lappin and Leass (1994), which uses full parsing to provide information for pronoun resolution, although an automatic parser was used, sentences "*were edited slightly to overcome parse inaccuracies*" (footnote 9, p. 549), either with lexical substitutions or by simplifying constructions. This does not hinder the empirical evaluation of the anaphora resolver, but

also shows that the complete system (consisting of the slot grammar parser and Lappin and Leass' RAP resolver) would not have been robust enough for application on free text. As a result, researchers turned to the use of simpler information (i.e. mostly morpho-syntactic contextual features) and shallower methods (i.e. approximate models such as data-driven supervised learning) that gradually grew popular for solving other NLP tasks. However, in recent years, due to the availability of robust statistical parsing methods and resources annotated with semantic information, there have been many efforts to couple these shallow methods with more complex sources of information modeling syntactic heuristics, semantic preferences, as well as commonsense reasoning. This has let researcher focus (again) on re-introducing syntactic and semantic analysis into the models for coreference, this time as features for coreference classifiers automatically induced from linguistic data.

Working with large amounts of data also led to a second realization: that the traditional approach of encoding all rules by hand was very far from leading to satisfactory results. Getting the priorities among anaphoric resolution rules right, and finding ways of combining evidence, proved as difficult for anaphora resolution as for parsing and other aspects of CL, as testified, e.g., by the discussion in (Alshawi 1987) concerning the difficulty to set activation levels[33]. The solution was the adoption of machine learning techniques for learning such rules, and of probabilistic techniques for evidence combination, just as done in other areas of CL (Bod et al. 2003).

## 6.1 Resolution Architectures

Modern anaphora resolution algorithms are all concerned with the construction of equivalence sets of mentions of discourse entities, which in Computational Linguistics are generally known as **coreference chains**. Identifying such chains involves recognizing 'links' between mentions, or between mentions and entities, and clustering these into equivalence classes. Previous proposals differ on how to automatize this task along several dimensions.

**Hand-coded versus Machine Learning.** Virtually all anaphora resolution methods proposed in the last decade, starting with the seminal proposal of Soon et al. (2001), take advantage of machine learning techniques, but a substantial amount of hand-coding can still be found in the feature extraction components. Also, most such algorithms make use of supervised learning, that requires hand-annotated resources; a few proposals making use of unsupervised learning however have been made (Haghighi and Klein 2007; Ng 2008).

**Single versus Multiple Classifiers.** Whereas the early algorithms like the Hobbs algorithm (Hobbs 1978), or SPAR (Carter 1987) tend to either focus on a single type of NP or to include distinct algorithms for each type of NP like Sidner's algorithm (Sidner 1979), most machine learning systems tend to develop a single model for all types of

[33] We experienced similar problems in our own work using prioritized Default Logic for the interpretation of definites (Poesio 1993).

NP (Soon et al. 2001; Ng and Cardie 2002b, *inter alia*). However, some exceptions exist (Hoste 2005).

**Serial versus Parallel.** Many anaphora resolution algorithms consider one antecedent at a time, and choose the first that satisfies the constraints: examples include Winograd's algorithm, that considers antecedents going backwards from the anaphor; Sidner's algorithm, that considers the antecedents in the order dictated by Sidner's focusing rules; the Left-Right Centering algorithm (LRC) algorithm (Tetreault 2001); and Soon *et al.*'s algorithm, that goes backwards one antecedent at a time. This way of considering antecedents however makes it difficult both to compare alternatives and to consider global constraints like those we discussed with Carter's algorithm. The alternative are parallel or ranked algorithms, that consider several competing hypotheses at once and decide on the basis of their score or of other preferences.

Historically, the first theories of disambiguation both in psycholinguistics and in CL were serial, as this type of model can easily explain incremental interpretation effects such as garden paths; but most recent theories of semantic disambiguation in psycholinguistics are based on a ranked parallel model (see, e.g., (MacDonald et al. 1994)), and most disambiguation algorithms in CL are of this type as well, from classic algorithms using heuristically calculated weights such as the Hobbs / Shieber scope disambiguation algorithm (Hobbs and Shieber 1987) to abduction based resolution methods (Hobbs et al. 1993) to current statistical models. In anaphora resolution, as well, parallel models are becoming increasingly popular, as we will see in this Section. Examples include Carter's SPAR, the BFP algorithm (Brennan et al. 1987), and, more recently, the ranking algorithm proposed by Ng and Cardie (2002b), the tournament model (Iida et al. 2003a; Yang et al. 2003), as well as *global models* to handle the intricacies of the coreference resolution task. These include, among others, antecedent ranking models (Ng 2005; Denis and Baldridge 2007b; Rahman and Ng 2009), unsupervised models (Haghighi and Klein 2007), as well as document-level models of anaphoricity and antecedent selection (Culotta et al. 2007; Daumé III and Marcu 2005; Denis and Baldridge 2007a).

**Generate-Filter-Rank.** Most of the early algorithms that we have seen are of this type, including the Sidner algorithm (Sidner 1979), the BFP algorithm (Brennan et al. 1987), and RAP (Lappin and Leass 1994), but also many algorithms that we will discuss in this section, including (Mitkov 1998; Ng and Cardie 2002b). This class of algorithms is motivated by the distinction between constraints and preferences and involves three main steps:

- One or more **generators** are used to extract antecedent candidates, usually by extracting all noun phrases from the preceding text–but sometimes in a order dictated by syntactic or focusing preferences.
- **Filters** are used to exclude antecedents on the basis of *hard* linguistic constraints, such as binding constraints or agreement constraints[34].

[34]While hard linguistic constraints hold in almost all cases, one should also note that antecedents that do

- Finally, a **ranker** uses preferences – either based on syntactic and/or surface order, or other indicators of salience – to order the remaining antecedent candidates so that the most salient candidate is chosen. Given the unavailability of full-fledged document understanding components, ranking tend be performed based only on surface form and configuration information, e.g. position of a markable in a document, as well as a shallow approximation of the markables' semantic classes.

In the case where the ranking mechanism is very predictable (as in the case of Hobbs' algorithm), candidates can be generated in their ranking order and the first candidate that passes the filter is selected. For some models, such as many Centering-based approaches, which posit that all pronouns in an utterance are resolved simultaneously, the *generate-filter-rank* approach can be adapted by jointly filtering and ranking the decisions for several anaphoric mentions in a sentence. By contrast, almost all machine learning approaches do not make a distinction between constraints and preferences and treat both types of information as features.

**Clustering-based Approaches.** Most of the algorithms that we consider resolve anaphors in the order in which they appear in the text. **Clustering approaches**, instead, take a more global view in constructing coreference chains. These approaches are usually based on some sort of uncertainty reasoning, either using constraint propagation (Lin 1995; Klenner and Ailloud 2008) or a probabilistic approach (Kehler 1997; Culotta et al. 2007), under the assumption that not all decisions on single antecedents will be completely reliable and that taking into account the larger context allows to correct some of these decisions. These models can include the information used by generate-filter-rank models, sometimes by incorporating the decision of a resolver of that kind as a feature (as is the case with, e.g. Lin 1995 and Culotta et al. 2007, which use a rule-based pronoun resolver to include links for probable pronoun antecedents in their model).

## 6.2 Heuristic Approaches to Pronoun Resolution

The problem of developing robust anaphora resolution algorithms was at first tackled by developing heuristic approaches able to work with the poor-quality information that could be extracted from corpora back in the '90s.

**MARS.** On the first examples of this approach was developed by Mitkov (1998), who proposed an approach where heuristic rules are used to assign a score to antecedent candidates and subsequently select the highest scoring candidate. Mitkov uses technical manuals as evaluation material for his experiments. While the heuristics are tailored to the domain of technical manuals, using knowledge-intensive features was avoided.

not match the anaphor in number and gender are still possible: consider mentions of group entities such as e.g. police and rock bands with the pronoun *they*. Other errors at the filtering stage can be due to (1) incomplete or imperfect gender information, such as unknown or occasionally ambiguous gender (consider *nurse*, that is most often female, but can be male as well), or (2) in the case of syntactic constraints, errors from automatically generated phrase parses.

Compatible antecedent candidates are collected, and then the following set of heuristics is used to calculate a score as the sum of the heuristics' individual scores (the contribution of an indicator is 0 for the cases where it doesn't apply):

- *Definiteness*: Since definite noun phrases are more likely to be discourse-old, and thus salient, indefinite NP antecedent candidates get a $-1$ score.
- *Givenness*: The first NP in a sentence gets a score of $+1$, on the grounds that it is more likely to represent given information.
- *Indicating Verbs*: The objects of verbs such as *discuss*, *present*, *illustrate*, *summarise*, *examine* etc. are given a $+1$ score.
- *Lexical iteration*: if a noun phrase head occurs more than once within the paragraph, this is taken to be an indication that the entity is especially salient and the corresponding noun phrases are given a $+1$ (two occurrences in the paragraph) or $+2$ (more than two occurrences) score.
- *Section heading preference*: A noun phrase that occurs in the header to the current section gets a $+1$ score.
- *"Non-prepositional" noun phrases*: Noun phrases embedded in prepositions are dispreferred (on the grounds of grammatical salience) and given a $-1$ score.
- *Collocation pattern preference*: Noun phrases that occur as a subject/object of the same verb as the anaphor are preferred and get a $+2$ score.
- *Immediate reference*: In a coordinated construction of the form "$V_1$ NP and $V_2$ *it*", a resolution of *it* to the noun phrase in *NP* is preferred as it usually expresses strong parallelism. The noun phrase in parallel position (*NP*) gets a $+2$ score.
- *Referential distance*: This information source prefers nearby antecedent candidates over distant ones. In complex clauses, noun phrases in the previous clause get a $+2$ score. Otherwise, noun phrases one, two or more than two sentences back get scores of $+1$, $0$, or $-1$, respectively.
- *Term preference*: Candidate noun phrases are checked against a list of nouns that are part of the domain's terminology, and get a $+1$ score if they are such terms.

Tie breaking between candidates with the same score is done by considering the *Immediate reference*, *Collocation pattern preference*, and *Indicating verbs* scores (in that order) and selecting the first candidate that has a higher score for one of these individual features, and, failing that, selecting the most recent candidate.

Mitkov evaluates his approach on two technical manuals using hand-corrected gender, chunks and clauses, yielding 89.7% accuracy, compared to 75% for emulating Baldwin's approach (discussed below) by hand, or 66% for always selecting the most recent matching candidate.

**Heuristics for high-precision resolution.** Baldwin (1997) uses noun phrase and clause chunking to determine mentions and utterances and uses shallow patterns to identify a small number of cases that can be reliably resolved. Based on the notion that cases should be resolved only if the partial order established by the shallow information available (which includes approximate identification of subjects and objects, but not lexically determined control) yields a single preferred antecedent, the system implements the following high-precision rules:

- *Unique in Discourse*: If there is a single compatible antecedent in the prior discourse, resolve to that antecedent.
- *Reflexive*: Resolve reflexive to nearest possible antecedent.
- *Unique in Current+Prior*: If the preceding noun groups of the current sentence and those in the previous sentence yield exactly one compatible antecedent, resolve to that antecedent.
- *Possessive Pro*: In the case of a possessive pronoun in "*his* X", if the previous sentence contains one exact match "*his* X", resolve to that possessive pronoun as an antecedent.
- *Unique Current Sentence*: If there is a single compatible antecedent in the preceding noun groups of the current sentence, resolve to that antecedent.
- *Unique Subject/Subject Pronoun*: If the anaphor is the subject of the current sentence, and the subject of the prior sentence contains a single possible antecedent, then resolve to that antecedent. (In the case of coordinated noun phrases, Baldwin counts the conjuncts as multiple subjects).

To resolve all pronouns in the text, Baldwin proposes two additional rules:

- *Cb-Picking*: Motivated concepts from Centering Theory (Section 4.3), this rule resolves some cases that the *Subject/Subject* rule does not cover. If the anaphor is in a non-subject position and the subject of the utterance is a compatible pronoun (i.e., the Cb), pick that pronoun as the antecedent.
- *Pick most recent*: picks the most recent compatible antecedent.

On a corpus of three stories about two participants, hand-annotated with gender information, Baldwin obtains 92% precision, at 60% recall, using the six high-precision rules, compared to a resolution accuracy of 77.9% using the two additional rules to resolve all anaphora, and 78.8% for Hobbs' naïve algorithm.

A variant of this system was used in the context of the MUC-6 evaluation (Baldwin et al. 1995). The overall system performed gender determination by means of a WordNet look-up, and clause chunking using the Collins' parser (Collins 1997). Special treatment was added to process first-person pronouns in quoted speech, and the *Possessive Pro*, *Cb-Picking* and *Pick most recent* rules were removed and an additional Pattern which selects the subject of the immediately surrounding clause (*Subject same Clause*) was added. Using these settings, together with an automatic detector for non-referential *it* pronouns, the system achieved 75% recall and 73% precision on the pronouns from a subset of the MUC-6 data.

## 6.3 Early Machine Learning Models

In heuristic approaches such as Baldwin's and Mitkov's, the ordering of the selection heuristics and their weights are essential to produce the final clustering of markables into coreference chains. However, tuning the ordering and the weights of the single heuristics is a time-consuming and error-prone task. In contrast, using machine learning techniques allows to perform these tasks automatically by empirically learning both constraints and preferences from a set of training data. This in turn makes it possible to explore a much larger range of features than using a purely rule-based heuristic approach. A number of earlier approaches also focused on coreference resolution as a sub-task for information extraction, limiting themselves to the semantic types relevant for the information extraction task at hand. These systems have clearly influenced the subsequent domain-independent machine learning approaches to the full coreference task, and can also be seen as early models on how to use knowledge and/or reconcile conflicting information in the process.

**Aone and Bennett.** The system of Aone and Bennett (1995) resolves both anaphoric pronouns, anaphoric definite noun phrases and name coreference for persons and organizations in Japanese text using a machine-learning approach based on decision trees (Quinlan 1993). For the purpose of creating a training and test corpus, zero pronouns, anaphoric definites and discourse-old name mentions are marked up in the text by hand[35].

To train the decision tree classifier, positive examples are created by pairing each anaphor in the training data with all prior members of its coreference chain, whereas negative examples are generate by pairing those same anaphors with all previous mentions that are not coreferent. Feature vectors to model syntactic (e.g. NP form) and semantic information (e.g. semantic class) are then created for each instance pair and fed to the classifier to learn a model of the features' relevance. During testing, the resolution process works by pairing an anaphoric expression with each possible antecedent: feature vectors are again created for each candidate anaphor-antecedent pair and then classified by the decision tree previously learned from the training data. Among the possible antecedents that are classified as positive (i.e., coreferent), the one with the highest confidence is taken, possibly breaking ties by selecting the closest antecedent.

**RESOLVE.** McCarthy and Lehnert (1995) also developed a decision-tree-based coreference resolver, later called RESOLVE, for the MUC-5 information extraction task. Their system uses both domain-independent features (mention type, name substring, being in a common noun phrase, being in the same sentence) as well as domain-specific one (either or both mentions referring to a company created in a joint venture). An evaluation on hand-annotated texts from MUC-5 shows that the learnt decision tree gives significantly better recall than the previous rule-based system (described in Lehnert et al. 1992) at a very small cost in precision. McCarthy and Lehnert's system creates instances for all pairs of template-relevant noun phrases. Coreference sets are built

[35]The automatic identification of Japanese zero pronouns, is a difficult task in itself, see section 8.

by positing 'explicit' coreference links between mentions in pairs that are classified as positive by the decision tree.

A later version of RESOLVE was evaluated in the MUC-6 coreference task (Fisher et al. 1996), using features such as string match, being the most recent compatible subject, having the same semantic type, or being a role description for a person (using several patterns to match common forms of role descriptions). Blind evaluation results on MUC-6 achieved 44% recall and 51% precision, a rather low result compared to the performance on the development set and the best-performing systems in MUC-6 such as the rule-based system by Kameyama (1997), something that Fisher *et al.* attributed, among other things, to the focus on person and organisation mentions[36].

**Vieira and Poesio.** Vieira and Poesio (Vieira 1998; Vieira and Poesio 2000) focused on the resolution of definite noun phrases, but in unrestricted text. The resolution of nominals is the hardest part of anaphora resolution, much harder than either pronoun resolution or the resolution of proper names; and unrestricted domain resolution is harder than resolution of a restricted range of entities since the mentions that are identified as relevant for an information extraction task usually only represent a small portion of all mentions, but also because of the unavailability of a domain-specific representation. The Vieira and Poesio work is interesting as a first attempt to develop approximate solutions to the problem of using lexical and commonsense knowledge. It is also interesting as a comparison of hand-coded algorithms with machine-learned ones: Vieira and Poesio developed both hand-coded and machine-learned versions of their decision tree and compared them.

The first concern of Poesio and Vieira was to develop a typology of definite noun phrases (Poesio and Vieira 1998). Not all definite noun phrases are anaphoric (Loebner 1987; Fraurud 1990; Poesio and Vieira 1998): more than half of the definites in unrestricted corpora are discourse new descriptions like *the pope*, *the first man to land on the Moon*, or *the fact that Inter won Serie A* (see Section 2.5). In addition, there are associative descriptions that denote an object that is in itself discourse-new but associated to some introduced entity. Vieira and Poesio's system thus solves several tasks; one is the identification of discourse-old versus discourse-new descriptions, the other is finding compatible antecedent candidates for discourse-old descriptions and choosing one for resolution. A decision tree integrates heuristics relevant for both tasks, as unique descriptions can still be discourse-old.

For the resolution of **direct anaphora** such as "*a house*"…"*the house*", the system identifies all noun phrases that have the same head as the definite noun phrase; these candidate antecedents are then checked for compatibility using modification heuristics, as *the house on the left* cannot serve as an antecedent to a noun phrase *the house on the right*. For antecedents including premodifiers, the premodifiers of the definite description have to be a subset of the premodifiers of the antecedent; for antecedents that do not include a premodifier, the antecedent can have premodifiers, as these may be non-restrictive as in *a check…the lost check*.

Head-matching may also produce spurious antecedents where an earlier part of

[36] According to McCarthy (1996), only about 50% of the mentions in MUC-6 referred to persons or organizations.

discourse discussed one entity of a type and later on, a different entity of the same type is introduced by accommodation, which means that it can be beneficial to exclude potential antecedents that are too far away from a definite noun phrase by using adequate **segmentation heuristics**. Vieira and Poesio found that only considering the most recent same-head noun phrase on one hand and limiting the distance at which the antecedent can be works best; they developed a *loose segmentation* heuristic where the antecedent either has to be within a four-sentence window or either be *discourse-old*, or *identical* (including all modifiers and the definite article) to the definite noun phrase being resolved.

Vieira and Poesio's algorithm also includes several heuristics for detecting **discourse-new descriptions**–indeed, it was the first system developing methods for this purpose. (This then became an important area of research; subsequent results are discussed in Section 6.8.) An important source of information is syntax: Vieira and Poesio's algorithm looks for specific syntactic configurations such as the presence of restrictive postmodification by prepositional phrases and relative clauses (and indication of so-called **establishing relative clauses** (Hawkins 1978; Loebner 1987)) or whether the definite occurs in appositions or copula constructions (an indication of a predicative use). A second important source of information is the presence of functional heads like *father* or modifiers that make predicates functional, like ordinals or superlatives. Often definites with such modifiers are not anaphoric in that definiteness is licensed through semantic uniqueness instead of uniqueness in context (Loebner 1987).

Some kinds of **bridging descriptions**, as well, are within the reach of a system that uses semantic resources without aiming at full text understanding: while definite NPs that refer to an event introduced by a verb phrase (such as *Kadane oil is currently drilling two oil wells. The activity* ...) are currently out of reach, lexical resources such as WordNet allow to resolve some cases where hypernymy or synonymy with the antecedent's head indicates a possible coreference relationship (e.g., *dollar-currency*), or part-of relations that can underlie associative bridging (e.g., *door-house*). Lists of categorized named entities (gazetteers) are also helpful for resolving instance relations (as in *Bach/the composer*).

Vieira and Poesio combined the sources of information listed above for detecting discourse new descriptions and resolving anaphoric and bridging definites in two ways: using a hand-coded decision tree, and one learned using ID3 (Quinlan 1986). The hand-coded decision tree adopts a strategy similar to that adopted by Baldwin for COGNIAC: it first uses high-precision heuristics for discourse new detection (look for apposition/copula constructions and special predicates), then tries to find a same-head antecedent, and then uses lower precision information: apply the other heuristics for discourse-new detection and (if none of these applies) looking for either an coreferent or associative bridging antecedent. The machine learned decision tree switches the first two steps: first attempts same-head resolution, then the high-precision discourse new heuristics, then the rest. The system uses an incremental resolution strategy, building a file card for every noun phrase it encounters. When it encounters a definite nominal, it applies the decision tree to determine its classification, and possibly attempts to find an antecedent. In this case a serial resolution strategy is adopted, going right-to-left until the first suitable antecedent is found or the boundary of the segment reached.

The two decision trees (hand-coded and machine-learned) were trained / developed

using 20 texts from the Penn Treebank (Marcus et al. 1993), containing 6831 NPs in total and 1040 definite descriptions. The hand-coded system was evaluated on 14 texts from the Penn Treebank, containing 2990 NPs and 464 definite descriptions, using the hand-annotated mentions. The system achieved a recall of 53% and a precision of 76% for an overall F=63%.[37] A hand-coded version classifying all unresolved definites as discourse-new was compared with the automatically learned decision tree on a subset of the previous test set containing 200 definite descriptions, again using hand-annotated mentions. A version of the hand-coded system only attempting to distinguish between discourse new and discourse old definites achieved an F-measure of 77%, whereas the automatically learned decision tree an $F_1$ of 75%. Attempting to interpret bridging references improved recall but made precision much worse: this version of the system only achieved an F-measure of 62%.

## 6.4 Anaphora Resolution: A Probabilistic Formulation

The empirical turn of the last two decades pushed researchers to develop statistical formulations for the problem of anaphora resolution, just like in all other subfields of Computational Linguistics. Although not all machine learning methods used in anaphora resolution are probabilistic, formulating this problem from a probabilistic perspective has provided indeed a powerful framework to advance the stat-of-the-art.

Given mention $m_j$, anaphora resolution is the problem of finding entity $e_i$ belonging to the universe of discourse U for which it is most likely that $m_j$ is a mention of $e_i$. In probabilistic terms, this means finding entity $e_i$ such that the probability of $m_j$ being a mention of $e_i$ is maximal, given context the $C$ of $m_j$.

$$\text{argmax}_{e_i \in U} P(m_j \textbf{ mention-of } e_i | C)$$

A completely general formulation should also cover the possibility that $m_j$ is discourse new – i.e., that it introduces a new entity $e_{new}$ – or non-referring (i.e., an expletive). This can be done by allowing $m_j$ to be a mention of a new entity $e_{new}$ not included in U, and introducing a pseudo entity $\epsilon$: we write that $m_j$ is a mention of pseudo entity $\epsilon$ to mean that $m_j$ is not referring. This leads to the following more general formulation:[38]

$$\text{argmax}_{e_i \in E} P(m_j \textbf{ mention-of } e_i | C), \quad E = \{U \cup \{e_{new}\} \cup \{\epsilon\}\} \ .$$

The formulation above suggests that evidence combination techniques from probability could be used. E.g., viewing context C as a set of features $f_k$, applying Bayes' rule, and making the Naive Bayes assumption, we can compute the desired probability as follows:

[37] The metric used for computing recall and precision was whether the proposed antecedent belonged to the same coreference class as the annotated antecedent–see Section 5.2.2 for problems with this metric.

[38] Notice that this formulation also covers discourse deixis – reference to abstract objects not explicitly introduced by nominals – and bridging references–references to entities which are new, but related by a 'sufficiently salient' relation (for the notion of 'sufficiently salient' see (Prince 1981; Barker 1991; Poesio 1994b)).

$$P(m_j \textbf{ mention-of } e_i|C) =$$

$$P(C) \times P(C|m_j \textbf{ mention-of } e_i) =$$

$$P(f_1) \times P(f_1|m_j \textbf{ mention-of } e_i) \times \ldots P(f_m) \times P(f_m|m_j \textbf{ mention-of } e_i) \ .$$

In practice, systems estimate the probability that an indicator variable $L$, which is 1 if $m_j$ is a mention of $e_i$ and 0 otherwise, is 1 (e.g., see (Yang et al. 2008)):

$$\text{argmax}_{e_i \in U} P(L|m_j, e_i) \ .$$

In the case of so-called *mention-pair models* (Section 6.6), this probability is approximated to classify links between mentions:

$$\text{argmax}_{m_i} P(L|m_j, m_i) \ .$$

## 6.5 Early Probabilistic Models

**Ge *et al.*** Ge et al. (1998) use a generative statistical model to add statistical gender identification, selectional preferences and a mention-count-based measure of saliency to Hobbs' naïve pronoun resolution algorithm. The way they compute a probability distribution over plausible antecedents $e_i$ could be reformulated in terms of the notation just introduced as[39]

$$P(m_j \textbf{ mention-of } e_i|C) \propto P(d_H|e_i)P(m_j \textbf{ is-pronoun}|e_i)\frac{P(e_i|h,t,l)}{P(e_i|t)}P(e_i|m_i)$$

where $P(d_H|e_i)$ is the 'Hobbs distance' (the distance between the pronoun and the last mention of $e_i$ computed using Hobbs' algorithm), $P(m_j \textbf{ is-pronoun}|e_i)$ is the probability that the mention take pronoun form given the antecedent, $\frac{P(e_i|h,t,l)}{P(e_i|t)}$ is selectional preference information, and $m_i = \#\{m_k|m_k \textbf{ mention-of } e_i\}$ is the number of times $e_i$ has been mentioned; all these probabilities are derived from corpus statistics.

For the distribution of Hobbs distances, Ge *et al.* ran the Hobbs algorithm repeatedly for each pronoun in the training corpus until 15 candidates were found and recorded the position in that list of candidates:

$$P(d_H = k|e_i) = \frac{\#\text{correct antecedents at position } k}{\#\text{correct antecedents}}$$

The relation between pronoun gender and the antecedent (e.g. $e_i =$ Marie Giraud, $m_j =$ *she*) is determined by looking at anaphor-antecedent pairs from the training data, backing off to the overall distribution of pronoun gender in the case of unseen $e_i$. Ge *et al.* also present an extension in which they use automatically resolved anaphor-antecedent pairs from a large corpus, which results in a small improvement.

[39] In the equation, $\propto$ means that an additional normalization factor is needed to obtain a probability distribution; for the purpose of choosing the antecedent with the greatest probability, however, this normalization need not be performed, since the relative magnitudes of the values remain the same.

The selectional preference information, represented as $\frac{P(e_i|h,t,l)}{P(e_i|t)}$ in the model, uses information from the statistical parsing model of Charniak (1997). For the statistic on mention count, $P(e_i|m_i, j)$ dependent on the mention count $m_i = \#\{m_k|m_k$ **mention-of** $e_i\}$ of the antecedent and the position in the text, represented as a sentence number $j$ (later mentions are bound to have been mentioned previously more often), a simple count of antecedent and non-antecedent $m_i, j$ pairs is performed.

Ge *et al.* test their model on texts from the Penn TreeBank, identifying mention counts by hand and also removing cases of expletive *it* pronouns; the full model achieves 84.2% accuracy, against 65.3% for the one using only the Hobbs distance (without any checking for gender compatibility), with the most important gains coming from the gender model (+10.4%) and the mention count (+5.0%).

**Kehler.** Kehler (1997) builds a maximum entropy (MaxEnt) classifier (Berger et al. 1996) to determine the probability that two mentions corefer. He then presents two approaches to convert these probabilities into a probability distribution over partitions of mentions (i.e., a set of coreference chains each corresponding to one entity). The first, called *evidential reasoning approach*, uses the pairwise classification (as coreferent or non-coreferent) of all pairs of mentions from the basic MaxEnt classifier, further normalized to account for the fact that the model would assign a non-zero probability distribution to inconsistent partitions (e.g., $A =_c B \wedge B =_c C \wedge A \neq_c C$). The other approach, called *merging decisions*, models the creation of a coreference set as a sequence of decisions where every mention is either merged to a previously existing set or put into a new set (when all classifier's responses for that mention are negative). The probability for the merging decision between a mention and an existing set of mentions is determined by the coreference probability between the mention itself and the closest one from the set. For example, given a partition of the previous mentions into sets $\{\{A, B\}, \{C, D\}\}$ a new mention $E$ could either be merged to the set $\{A, B\}$, with a probability proportional to $p(B =_c E) \times p(D \neq_c E)$, or to the set $\{C, D\}$, with a probability proportional to $p(B \neq_c E) \times p(D =_c E)$.

Training examples are generated depending on which resolution mechanism is used: for the evidential reasoning approach, a training example is generated for every pair in the training data, whereas for the merging decisions approach, each mention is paired with the most recent mentions of each of the (partial) coreference sets formed by previous mentions.

Among the features made available to the MaxEnt classifier, Kehler uses the template structure generated by an information extraction system, such as e.g.

$$\begin{bmatrix} \text{FACILITY} & \text{DEPOT} \\ \text{NUMBER} & 1 \\ \text{LOCATION} & \text{FAIRVIEW} \\ \text{TYPE} & \text{AMMUNITION} \end{bmatrix}$$

for the mention "*the ammunition depot in Fairview*". Using these slot-based representation of complex phrases, Kehler is able to model the compatibility between two mentions as a function of the two template representations, i.e. as either having *identical slot values*, or one template *properly subsuming* the other, or being *otherwise consistent*

(where both templates have a filled slot that is not filled in the other). Other features include the form of the noun phrase (*indefinite*, *definite* including pronouns, or *neither*, including bare nouns), the distance in characters between anaphor and antecedent, discretized into five classes corresponding to different ranges, as well as the result of a rule-based coreference module. This indicates whether a potential antecedent is *preferred*, i.e. it is the same as the one the rule-based module would choose, *possible*, if it is contained in its list of antecedents but not as the highest-ranked one, or *implausible*, in case it was not considered to be a suitable antecedent by the rule-based module.

Kehler reports that in the trained models, the indicators for two or more common slot fillers and for matching names receive a positive value, whereas the indicators for cases where the later template *properly subsumes* (i.e., is more specific than) the antecedent candidate or is *otherwise consistent* receive a negative weight; a positive value was also learned for instances that are the preferred antecedent of the rule-based system.

For the evaluation, Kehler reports cross-entropies on test data as well as the number of exact matches. While the cross-entropy for both the merging decision model and the evidential model were superior to the baseline model, the evidential reasoning is found superior both in terms of cross-entropy and of perfect matches to the merging decision model.

## 6.6 The Mention-Pair Model of General Coreference

Soon et al. (2001) and Ng and Cardie (2002b) brought about a shift away from models focused on a single NP type and a restricted domain, and towards general coreference in unrestricted domains. The so-called **mention-pair** model proposed by Soon *et al.* and developed by Ng and Cardie has become the standard statistical formulation of the anaphora resolution problem. According to this model, resolving anaphor $m_j$ can can be viewed as a classification task, the task of finding *mention* $m_i$ that maximizes the probability:

$$\text{argmax}_{m_i} P(L|m_j, m_i)$$

(See Section 6.4.) Their algorithms – including the choice of features, the training and decoding methods – have become the standard benchmarking baseline for anaphora resolution in the last decade, in a way similar to what the Hobbs algorithm represented for rule-based systems.

**Soon *et al.*** Soon et al. (1999, 2001) present a decision-tree-based system for coreference resolution that is evaluated on the MUC-6 and MUC-7 corpora. The proposal explicitly addresses the issue of preprocessing unrestricted text, i.e. in order to automatically identify those markables which are to be later analyzed by the coreference classifier. The preprocessing pipeline consists of a cascade of sequence taggers using standard statistical learning methods (Hidden Markov Models, HMM), for part-of-speech tagging, noun chunk identification and named entities (NE) recognition and classification. Since a mention can be both a named entity and a noun phrase, the spans found by both modules are merged and phrase boundaries adjusted as necessary. Two

additional modules extract possessive premodifiers (e.g., *Eastern* in *Eastern's parent*) and premodifying nouns (such as *wage* in *wage reductions*), which MUC-6 allows to co-refer with other mentions.

While the use of different modules for the identification of noun chunks and NEs, as well as an extra component for their merging, may appear unnecessarily complicated at first sigh, it allows the use of standard out-of-the-shelf components for these tasks (thus ensuring portability across languages and domains), as well as their combination to boost the recall of retrieving potentially coreferring candidates. Soon et al. (2001) mention that while the implementation of RESOLVE used for MUC (Fisher et al. 1996) fared much worse than their system, due to a preprocessing module that was mostly focused on persons and organizations, recreating the learning framework of McCarthy and Lehnert's RESOLVE with their preprocessing and features led to comparable results to the Soon *et al.* system.

The training examples in Soon *et al.*'s system are generated as follows: (i) positive examples are generate by pairing each markable in a gold-standard coreference chain (i.e. the *anaphor*) with its most recent antecedent in the chain; (ii) negative examples are generated by pairing the anaphor with all other markables found in the text between it and its most recent antecedent, i.e. those belonging to another coreference chain or no chain at all.

In order to train a decision tree classifier (Quinlan 1993), Soon *et al.* model the training and test instances in term of *feature vectors*. The system uses twelve features, listed in full in Table 3. Some of these features that are indicative of the **form of the noun phrase** (the anaphor/antecedent being a pronoun, the anaphor being a definite/demonstrative noun phrase, as well as a feature indicating that both mentions are proper names). Other features specify **agreement** in number, gender, or semantic class, **distance** in sentences between the two mentions, as well as **string matching** (stripping off determiners, i.e., essentially comparing the head and the premodifiers) and an **alias feature** that copes with name variation (allowing the matching of "Mr. Simpson" and "Bent Simpson"), common organizational suffixes, and abbreviations.

During testing, the list of previously identified mentions is sorted in document order and is processed from left to right: every mention is a potential anaphor and every mention that precedes it in the list is a potential antecedent. Accordingly, each mention is paired with any preceding one to generate a test instance. The resolution model is purely serial: as soon as a test instance is classified positively the algorithm stops. A feature vector based on the same features as for training instances (Table 3) is generated for the test instance and is given to the classifier, which decides whether the pair of mentions is coreferent or not. If the mention pair is classified as coreferent, the coreference resolution algorithm moves to the next candidate anaphor in the list, else it iteratively pairs the potential anaphor with the next preceding candidate antecedent until a pair labeled as coreferent is output, or the beginning of the document is reached. Finally, given the pairs of mentions found to be coreferent by the classifier, a partitioning is imposed on the document. The collection of mentions is viewed as a disjoint set and coreferent pairs are clustered into separate, non-overlapping (coreference) sets using a union-find algorithm. The resolution model of Soon *et al.* is thus simpler when compared against the approach of Aone and Bennett (1995) or McCarthy and Lehnert (1995): in terms of the generate-rank-filter model introduced earlier, this means that

| Feature | Value | Description |
|---|---|---|
| **Distance feature** | | |
| DIST | integer | the distance in sentences between $m_i$ and $m_j$ |
| **NP type features** | | |
| I_PRONOUN | boolean | 1 if $m_i$ a pronoun |
| J_PRONOUN | boolean | 1 if $m_j$ a pronoun |
| DEF_NP | boolean | 1 if $m_j$ a definite NP |
| DEM_NP | boolean | 1 if $m_j$ a demonstrative NP |
| **Agreement features** | | |
| STR_MATCH | boolean | 1 if $m_i$ and $m_j$ string match |
| ALIAS | boolean | 1 if $m_j$ an alias of $m_i$ |
| GENDER | boolean | 1 if $m_i$ and $m_j$ gender match |
| NUMBER | boolean | 1 if $m_i$ and $m_j$ number match |
| SEMCLASS | boolean | 1 if $m_i$ and $m_j$ match semantically |
| NUMBER | boolean | 1 if $m_i$ and $m_j$ number match |
| PROPER_NAME | boolean | 1 if $m_i$ and $m_j$ both proper names |
| **Syntactic position** | | |
| APPOSITION | boolean | 1 if $m_j$ in appositive position |

Table 3: The 12 features used in the system from Soon et al. (2001)

ranking is done purely based on recency, whereas filtering – performed by the decision tree classifier – must take into account both hard criteria (such as gender and number compatibility) and preferences.

Soon *et al.* present an in-depth analysis of which features are deemed as the most important ones by the coreference classifier, i.e. the ones found in the highest regions of the learned decision tree. With the decision tree learned in MUC-6, the system has a strong tendency to resolve mentions to the closest antecedent that either:

1. has the same surface form (*string match* feature);
2. is detected as a name *alias* of the anaphor;
3. is a gender-matching antecedent in the same sentence for a pronoun anaphor.

Overall, Soon *et al.* reported a MUC $F_1$ score of 62.6% on MUC-6, and of 60.4% on MUC-7. From a linguistic viewpoint, the fact that this system does better than a majority of the participants in the MUC-6 evaluation despite the simplicity of the features is rather surprising. One part of the explanation is that Soon *et al.*'s system performs well on an often-overlooked aspect of the coreference problem, namely the identification of mentions in text as a necessary preprocessing step. In contrast for instance to McCarthy and Lehnert (1995) and Aone and Bennett (1995), they explicitly assess the influence of the preprocessing component responsible for automatically identifying the markables to be classified as coreferent, and find that their preprocessing pipeline recovers 85% of the mentions in MUC-6. Even more importantly, they do not focus on a small number of semantic classes that are relevant for an information extraction system[40] but rather develop a system to process all markables in unrestricted text. For the other part, considering Kameyama's (1997) analysis, the system successfully reaps

[40]Aone and Bennett only evaluate their system on organizations, while McCarthy and Lehnert focus mostly on organizations and persons.

a number of low-hanging fruits in the textual domain used for MUC: it resolves same-sentence pronouns, which according to Kameyama constitute a seizable part of the pronouns in MUC-6 and are relatively easy to resolve, and it also does a relatively good job at resolving name coreference. At the same time, the string matching feature allows to resolve easy cases of common noun phrases (i.e., those that have the same set of premodifiers).

**Ng and Cardie.** Ng and Cardie (2002b) propose a system that improves over the system proposed by Soon *et al.* in two main respects:

**Best-first clustering.** Instead of stopping at the first antecedent for which $P(L|m_i, m_j)$ is greater than a given threshold (i.e. $> 0.5$), their system computes the probability for all antecedents and selects the the one with the highest coreference probability value from among all antecedents with coreference class values above 0.5.

**Feature set expansion.** The effects of using a much larger feature set are investigated in detail. This extension explores the effect of including 41 additional features to the original feature set from Soon *et al.*, which include a variety knowledge sources for the coreference resolution classifier such as lexical, grammatical, semantic features, as well as the result of a 'naïve' external pronoun resolver.

The system developed by Ng and Cardie achieves a MUC $F_1$ of 70.4% on MUC-6 and of 63.4% on MUC-7. One of the key results of Ng and Cardie is that the best results are obtained by coupling best-first clustering with a set of 27 features which are manually selected by discarding those features which lead to low-precision decision tree rules for common noun resolution. At least with this amount of data, the decision tree learning algorithm does not appear to be effective at feature selection. The 27 features include 9 of the original Soon *et al.* features (the exceptions being `DEF_NP`, `DEM_NP`, and `STR_MATCH`, this latter replaced by three more specialized string matching features for different types of NP), plus 18 new ones.[41]

## 6.7 Beyond Mention-Pair Models

In the years after the work by Ng and Cardie, researchers have developed models of the coreference task that incorporate a more sophisticated view of anaphora resolution than the original methods developed by Soon et al. (2001) and Ng and Cardie (2002b).

One key direction of research have been the proposals from Iida et al. (2003a) and Yang et al. (2003), who proposed an approach where a machine-learning classifier performs the ranking by tournament-based scoring.

Another key development has been the move away from *local models* which attempt to determine the probability of links between mentions back towards *global*

[41]The most systematic analysis of the impact of features on the performance of mention-pair based anaphoric resolvers in the MUC corpora was carried out by Uryupina (2006). Overall, Uryupina tested 351 nominal features, encoding surface information (122 features), syntactic properties (64 features), semantic information (29 features), and salience information (136 features).

*models* that instead model the probability that a mention refers to a given entity (**entity-mention model**), and are therefore closer to the discourse model-based theories of anaphora and anaphora resolution proposed in Linguistics and Psycholinguistics (see Section 2.4). This return to an entity-based view of anaphora resolution, maintained in work such as (Vieira and Poesio 2000; Kabadjov 2007) but abandoned in the work following Soon *et al.*, was primarily motivated by the observation that systems that resolve an anaphor to an antecedent without looking at the previous linking decisions involving that candidate antecedent are prone to make implausible errors such as resolving *she* to *Clinton* where *Clinton* has been previously found to be a subsequent mention of *Mr. Clinton* (Luo et al. 2004; Yang et al. 2008). The focus of the entity-mention models is therefore to enforce **global consistency** across anaphoric chains. However, this opens up a new problems:

- As observed by Kehler (1997), using *only* information about members of a coreference chain without the notion of antecedence blurs certain important notions such as recency.
- Inconsistencies in the coreference chains could derive by *any* decision in the sequence of those performed for a single document. This means that the algorithm has to keep track of multiple alternatives (and their scores) in a search space which increases exponentially with the number of markables in a document.

Working with coreference chains requires ensuring the global consistency of coreference sets. Two influential proposals in this respect have been made by Luo et al. (2004) and Daumé III and Marcu (2005); the most recent proposal in this direction, by Rahman and Ng (2009), combines a entity-based model with a ranking algorithm.

**The Tournament Model** The **tournament model** of Iida et al. (2003a), also known as the **twin-candidate model** Yang et al. (2003) uses a classifier where an anaphoric expression is paired with two preceding candidates and the outcome of the classifier expresses a *preference* over one of the two candidates; using a pre-identified set of candidates.

One crucial aspect for such a ranking-based approach is the initial selection of candidates to be presented to the coreference classifier. This is due to two main issues: first, because class imbalance in the training set can potentially lead the classifier to bias towards the first or second candidate; second, since the classifier could link noun phrases which are instead be discourse-new, i.e. non-anaphoric. Accordingly, training data are created by pairing every positive candidate with every negative candidate and using these as a training pair[42]. To avoid the class imbalance problem, test data are generate differently for different kinds of NPs. For pronouns, the number of negative examples is limited by considering the pronouns in the current and the last two sentences, and candidates not agreeing in number, gender, or person are discarded. For non-pronouns (i.e. common nouns and proper names), the algorithm to create the training data uses all non-pronominal antecedents as positive candidates, as well as plausible non-pronominal in the same, previous, and next sentences as negative candidates.

[42]In this approach, the ordering between candidates implicitly reflects surface order, that is, the more recent candidate is always to the right.

Finally, for testing, a filtering of the candidates is performed using a Soon et al. (2001)-style single-candidate classifier for single candidates, and a ranking is induced among all candidates that have been previously classified positively by the single-candidate classifier.

The original model from Yang et al. (2003) is later extended in Yang et al. (2005) to include the identification of discourse-new, i.e. non-anaphoric, definite NPs in the general framework offered by the tournament model. In their extension of the original proposal, Yang *et al.* present an alternative way of determining non-anaphoric non-pronouns by integrating their classification into the tournament model used for ranking. Instead of choosing the left or right candidate (giving class label '10' and '01' as output, respectively), the classifier can also explicitly decide that both candidates are not suitable antecedents (thus outputting a '00').

In order to train such model, Yang *et al.* take discourse-new mentions from the gold-standard and pair them with randomly selected previous mentions; a sub-sample of these candidate-pair instances is then added to the training data with the appropriate classification label ('00'). During testing in terms of tournament classification, the score of the candidates is either increased for one and decreased for the other (in the case of 10 and 01), or it is decreased for both mentions (in the case of 00). Resolution to the best-scoring candidate is then only performed if its score is greater than 0 (i.e., more 'win' than 'lose' or 'neither one' classifications). Yang *et al.* find that, in comparison to the single-candidate model, this classification algorithm increases precision at a comparatively smaller cost in recall, yielding a substantial improvement in F-measure.

**Luo *et al.*** The earliest known proposal for a entity-based model in which training is carried out over clusters is due to Luo et al. (2004), whose resolution algorithm searches for the most probable partition of a set of mentions by structuring the search space as a Bell tree (Bell 1934) where each leaf contains a possible partition of the mentions. Their mention-entity model considers all mentions found in a document from left to right, and for each mention $m_k$, called *active mention*, it computes the probability of belonging to a previously generated partial entity $e_t$, referred as *in-focus* entity. Formally, the model estimates the probability

$$P(L|E_k, m_k, A_k = t) \tag{60}$$

where $L$ is an indicator variable equal to 1 if mention $m_k$ is to be linked to one of the partially-established entities $E_k$ to its left, and 0 otherwise, and $A_k$ is a random variable which expresses which entity $e_t \in E_k$ is in focus. In addition, the probability of $m_k$ of being non-anaphoric can be computed as $P(L = 0|E_k, m_k)$.

The size of the Bell tree, i.e. the number of ways of partitioning all mentions into non-empty disjoint subsets, is equal to the so-called Bell number, which increases factorially with the number of mentions. To explore such a huge search space, the entity-mention model is first approximated to assume that entities other than the one in focus have no influence on the linking decision, i.e. Equation (60) is approximated as $P(L = 1|e_t, m_k)$. In addition, several pruning strategies are used: these include removing very unlikely partial entity partitions, and limiting the number of hypotheses considered at one time by means of a beam search algorithm. Finally, the entity-

mention model is compared directly with a mention-pair model, which assumes that the entity-mention score can be obtained by the maximum mention pair score, i.e. here Equation (60) is instead approximated as $\operatorname{argmax}_{m \in e_t} P(L = 1|e_t, m)$.

Luo *et al.* train a binary classifier either on anaphor-antecedent pairs (for their *mention-pair* model) or on anaphor-coreference set pairs (for their *mention-entity* model), and similar to Ng and Cardie (2002b), they take the most highly scored candidate antecedent and resolve to it if its score is greater than some optimal threshold found on a held-out development data set.

For the entity-mention model, Luo *et al.* adapt the features that they use in the mention-pair model, including string matching, quantized edit distance, and surface distance; to achieve this, they take the *minimum* string distance across the mentions in a (partial) coreference chain, as well as the surface distance to the closest mention (i.e., the direct antecedent). They report that the entity-mention model performs slightly worse than the mention-pair model. However, crucial to their findings is the fact that not only the mention-pair model uses 20 times more features than the entity-pair model, but also that the entity-mention approach overcomes those errors deriving from a local modeling of the coreference problem, e.g. a male pronoun and female pronoun being clustered into the same entity.

**Daumé III and Marcu.** Daumé III and Marcu (2005) present a entity model for coreference resolution based on online learning. A beam search is used to overcome some types of non-optimal local decisions by keeping multiple partial solutions and discarding partial solutions when they are discovered to be inconsistent later on in the document.

In contrast to Luo et al. (2004), Daumé and Marcu do not modify the features to accommodate coreference chain information. Instead, their resolution algorithm aggregates the scores for pairing the anaphor with every antecedent in one (partial) coreference set, based on a variety of strategies such as: *max-link* (taking the highest score), *min-link* (taking the lowest score), *average-link* (taking the average score) or *nearest-link* (taking the score of the nearest element in that partial coreference set). In addition, they propose to use an aggregation method, called *intelligent-link*, which treats different mention types separately:

- Proper names are first matched to other names in the previous document, otherwise, against the last nominal, or, failing that, using the highest-scored link.
- Nominals are matched to the highest-scoring nominal in the previous chain, otherwise, against the most recent name; failing that, using the highest-scored link.
- Pronouns are resolved with an *average-link* approach against all pronouns or names, or, failing that, using the highest-scored link.

Linking pronouns only to pronouns or names makes sense in an ACE-type scenario (where only coreference chains for referents of a small set of classes are wanted). This is because most pronouns will either refer to persons or organizations (which are usually named), or to non-ACE mentions (in which case they do not have to be linked). Using mention clusters also allows Daumé and Marcu to include a "decayed density"

for each entity, somewhat similar to Lappin and Leass' (1994) salience measure, exploiting the fact that some entities are very central to a document and are referred to throughout the whole document, whereas other cases of pronominal coreference are only very local.

**Rahman and Ng** In the most recent contribution to the entity-mention model literature, Rahman and Ng (2009) propose a **cluster-ranking** algorithm combining a number of improvements over the early statistical models of anaphora resolution. In their algorithm, the most highly ranked coreference chain is chosen as the antecedent of a mention. Their model also jointly learns discourse-novel and anaphora resolution. The model outperforms (on ACE-05) mention-pair, mention-ranking, and entity-mention models both on true mentions and on system mentions and according to both the MUC and CEAF metrics.

## 6.8 Discourse-new Detection

As discussed in Section 2, not all definite noun phrases are anaphoric, and not all anaphoric noun phrases have a coreferring antecedent (witness the case of associative bridging or event noun phrases which take up a referent created by a verb phrase mention). This means that a system for coreference resolution can profit immensely from perfect or near-perfect information on which definite noun phrases need to be resolved to a coreferent antecedent and which do not, as techniques to use common-sense knowledge in the resolution of definite noun phrases (see Section 7) work well enough for resolving to an antecedent, but are of little help in deciding whether a definite noun phrase has to be resolved to an antecedent in the first place.

Besides the discourse-new/discourse-old distinction we are aiming at, a very important distinction is between uniquely specifying and truly anaphoric definite noun phrases (which can only be interpreted by considering information previously introduced in the discourse). Uniquely specifying noun phrases (such as *the man I saw yesterday* or even *the White House*) can felicitously occur as discourse-new mentions, and when they occur as a repeated mention, the variation in surface form between subsequent mentions – in most cases, name variation or aliasing – is of a different sort than in truly anaphoric definite noun phrases.

The first work in this area, (Vieira and Poesio 1997, 2000) uses mainly syntactic heuristics to distinguish between discourse-old and discourse new definite noun phrases: besides restrictive postmodification – both by prepositional phrases and restrictive relative clauses – they also use capitalization-based heuristics to recognize definite descriptions that are actually names (as in *the Pentagon*), and a hand-crafted list of special nouns such as *year* or *afternoon*, which usually are interpreted using information from the larger situation described in the text rather than a specific antecedent mention, and modifiers indicating uniqueness, such as *first*, *only*, *most* (see Section 6.3).

Bean and Riloff (1999) note that it is difficult to achieve full coverage with a hand-crafted list of nouns/descriptions and present an approach to automatically acquire such a list using unsupervised learning, exploiting the fact that uniquely specifying defi-

nite noun phrases (nearly) always occur with a definite article, whereas noun phrases that are anaphoric often occur in their indefinite variant (*the house*/*a house*, but *the weather*/??*a weather*).

One heuristic that Bean and Riloff use relies on the fact that mentions in the first sentence are (almost) always nonanaphoric since it is unlikely that it would have an antecedent. They use this to create a list of nouns that occur as definite in the first sentence of an article (the S1 list). Bean and Riloff then try to generalize this list to create patterns such as "*the* $\langle x^+ \rangle$ *government*", where the presence of a head noun (or nouns, in the case of compounds) together with premodifiers would be indicative of a matching noun phrase being uniquely referring. By expanding the patterns to the longest suffix of a noun phrase that occurs as often as the head (yielding *the* $\langle x^+ \rangle$ *national capitol* instead of *the* $\langle x^+ \rangle$ *capitol* for a noun phrase *the national capitol*), they improve the specificity of the patterns. Bean and Riloff call these *Existential Head Patterns* (EHP).

The other heuristic that Bean and Riloff use is the relative frequency of indefinite and definite variants of a noun phrase, as uniquely specifying noun phrases should only occur in definite form, whereas normally non-unique noun phrases should also occur in their indefinite variant. This is used in two ways: One is that heads and full noun phrases that occur at least five times in the training corpus and are definite in at least 75% (noun phrases) or 100% (heads) of the cases are used to create a list of 'definite-only' noun phrases (DO list), where nouns and noun phrases in the DO list are always classified as uniquely specifiying; for noun phrases that match the S1 or EHP lists, the definite/indefinite ratio of the noun phrase is compared to two thresholds: above the upper threshold, the noun phrase is always considered to be definite, whereas noun phrases with a definite/indefinite ratio that is between lower and upper threshold are classified as uniquely specifying if they occur within the first three sentences of the text.

More recent work has aimed to use larger-scale resources for the learning of definite-only noun phrases and the use of machine learning classifiers to integrate the information from syntactic and lexical heuristics: Ng and Cardie (2002a) present a machine learning classifier for discourse-new classification and show the results of the integration with their Soon *et al.*-style coreference system (Ng and Cardie 2002b). Ng and Cardie use features indicating the presence of a plausible antecedent (presence of a string-matching/head-matching/alias/hypernym antecedent candidate), but also pattern-based indicators of the form including pre- and postmodification, and also whether the mention is in the first sentence, first paragraph or in the header. For the integration into the coreference system, they find that not resolving a mention at all when the anaphoricity classifier indicates that the mention is discourse-new results in an appreciable gain in precision (mainly for common nouns, but partly also for proper nouns) that is however accompanied by a sharp drop in recall (again, mostly for common nouns). By forcing the system to resolve to a string-matching or alias antecedent even when the anaphoricity determination says otherwise, they are able to avoid most of the loss in recall while still having improved precision numbers.[43]

[43]Note that the results in Ng and Cardie (2002a) are below those of Ng and Cardie (2002b), presumably since the baseline resolver they use is the *all features* system in 2002b, which performs less well than either

Uryupina (2003) uses the complete noun phrase including premodifiers and the head noun to gather statistics from the web, which she uses as features in a decision list classifier including other information such as the presence of postmodification and the presence of and distance to a noun phrase with the same head. Uryupina uses two ratios: one of *the X* versus all occurrences of *X* and one versus indefinite *a(n) X*.

Poesio et al. (2004d, 2005) expand this work by including more features including the detection of superlatives, a web-based counterpart of Bean and Riloff's S1 list, and a text position feature that indicates whether the definite description occurs in the title, the first sentence, or the first paragraph. They also integrate the anaphoricity classifier into Guitar (Kabadjov 2007), a hybrid coreference resolver using the result of head-matching resolution as a feature in the anaphoricity classifier; they find that this scheme works well on the GNOME corpus used for testing.

In more recent work, in particular by Denis and Baldridge (2007a) and by Rahman and Ng (2009), discourse-new detection and anaphora resolution are treated as a joint inference problem. Denis and Baldridge (2007a) propose a particularly elegant solution based on the Integer Linear Programming methodology introduced by (Roth and Yih 2004).

## 6.9 Summary: What is the state of the art?

Making sense of the evaluation results reported in different evaluation settings and/or on different corpora is often difficult and requires careful attention to the exact evaluation setting; however, a comparison of results, where available, is often telling:

For example, the best participant entries in the original MUC-6 and MUC-7 evaluations (which are all rule-based) do better than the system of Soon et al. (2001), who adopted the most important features of successful approaches such as Kameyama's (1997), and only later approaches such as the one of Ng and Cardie (2002b) and Uryupina (2006), who use the possibilities offered by machine learning to introduce more informative features, outperform the best rule-based approaches.

The setting of "true" mentions (where only mentions that are part of a coreference chain in the gold-standard are considered for resolution) leads to a rather implausible evaluation setting when used with Vilain *et al.*'s MUC F-measure: Luo et al. (2004) remark that a baseline where all (gold-standard) mentions in a document are put into one coreference chain outperforms all published results for this setting by a large margin, whereas the result in Luo (2005) only produces 2-3 coreference chains for each document and is not meant to be a useful way of coreference resolution.

Results on the ACE-02 corpus are slightly harder to interpret due to large variation in settings and evaluation metrics used. The ACE-02 evaluation metric that was used in the shared task is difficult to interpret and has an application-oriented motivation rather than being geared towards simplicity or transparency. The ACE value is the only value for which published results include both settings with system-generated mentions and a setting with "true" mentions; it is apparent that even a very good system such as the one by Daumé III and Marcu (2005) that integrates the tasks of mention detection and coreference resolution produces results that are markedly below results

their 'duplicated Soon' baseline or their hand-selected feature set which yields the best results.

| MUC-6 | MUC-F |
|---|---|
| (Kameyama 1997)[a] | 64.8 |
| (Lin 1995)[a] | 63.0 |
| (Fisher et al. 1996)[a] | 47.2 |
| (Soon et al. 2001) | 62.6 |
| (Ng and Cardie 2002b) | 70.4 |
| **MUC-7** | **MUC-F** |
| (Humphreys et al. 1998)[a] | 61.8 |
| (Lin 1998b)[a] | 61.1 |
| (Soon et al. 2001) | 60.4 |
| (Ng and Cardie 2002b) | 63.4 |
| (Uryupina 2006) | 65.4 |
| **MUC-6 ("true" mentions)** | **MUC-F** |
| baseline: merge everything | 88.2 |
| (Harabagiu et al. 2001) | 81.9 |
| (Luo et al. 2004) | 83.9 |
| (Luo 2005) | 90.2 |
| (Haghighi and Klein 2007)[b] | 70.3 |

[a]: participants of the original shared tasks (no use of testing data for development) [b] unsupervised learning using a larger quantity of gold-mention information without coreference chains

Table 4: Evaluation Results reported on MUC-6/MUC-7.

for "true" mentions with the most basic feature set. Other metrics used on the ACE-corpus include the MUC measure and Luo's CEAF measure, for both of which different results were reported.

In Tables 4 and 5 we report some of the most prominent evaluation results reported on using the MUC-6/7 and ACE-02 corpora. As it can be seen from both tables, the variation between different proposals is very high. In practice, as discussed above, the systems themselves are hard to compare, mostly due to different preprocessing components and evaluation settings.

In order to make sense of these variations and limited comparability, Stoyanov et al. (2009) investigate the different subproblems involved in the coreference task and use a benchmarking system, called RECONCILE$_{ACL09}$ to identify the the impact of each subproblem on the overall performance of a coreference resolution engine. Their findings can be summarized as follows:

1. improving the performance of a named entity identifier is expected to have a limited impact on the overall performance of a coreference resolution system;

2. instead, that same performance can be boosted by a large margin by designing robust components to identify which markables in text are to be included in the coreference chains (*mention* identification), and whether these are anaphoric or not (*anaphoricity* identification);

| ACE-02 (system mentions) | MUC-F | CEAF | ACE-val |
|---|---|---|---|
| (Luo et al. 2004) | | | 73.4 |
| (Daumé III and Marcu 2005) | | | 79.4 |
| (Yang and Su 2007)[a] | 62.7–67.1 | | |
| ACE-02 ("true" mentions) | MUC-F | CEAF | ACE-val |
| (Luo et al. 2004) | | 73.2 | 89.8 |
| —, only string matching / alias | | 64.4 | 86.0 |
| (Daumé III and Marcu 2005) | | | 89.1 |
| (Ji et al. 2005)[b] | 83.7 | | |
| (Ponzetto and Strube 2006)[c] | 70.7 | | |
| (Ng 2007)[d] | 64.5 | 62.3 | |
| (Denis and Baldridge 2007a)[e] | 67.5–72.5 | | |
| (Haghighi and Klein 2007)[f] | 62.3–64.2 | | |

[a] Yang and Su only provide separate scores for different sections of the ACE-02 data set. [b] Ji *et al.* use a setting where links between non-key mentions are discarded. [c] Ponzetto and Strube use a setting where system mentions are aligned with the gold standard and non-aligned mentions are discarded. [d] Ng does not mention the exact setting in his paper; other papers suggest that his system delivers similar results for both gold-standard and system-generated mentions. [e] Denis and Baldridge only provide separate scores for different sections of the ACE-02 data set. [f] Haghighi and Klein report results on a subset of the training sets for an unsupervised approach.

Table 5: Evaluation Results reported on ACE-02.

3. as a natural consequence of (2), assuming the availability of "true" mentions immensely (therefore, unrealistically) simplifies the coreference resolution task.

As a result, their in-depth analysis suggest that, while many papers in the literature reported very high performance scores – implicitly suggesting that the coreference problem can be tackled with high accuracy in a data-driven setting – much work still lies ahead in order to develop high-performing coreference resolution systems which effectively learn the intricacies of the coreference problem inductively from linguistic data.

# 7 Linguistic, Commonsense and Encyclopedic Knowledge for Anaphora Resolution

Machine learning has led to performance rates for anaphora resolution systems fully competitive with rule based systems. Soon et al. (2001), for instance, show a comparison where the performance of their system is as good as the best performing systems from MUC-6 and MUC-7 (see *ibidem* Figures 3 and 4 at p. 532). Within the machine learning paradigm much effort has been spent in the very recent years on developing more robust approaches to handle the intricacies of the coreference resolution task: this includes antecedent ranking models (Ng 2005; Denis and Baldridge 2007b), unsupervised models (Haghighi and Klein 2007, 2010), as well as document-level global

models of anaphoricity and antecedent selection (Daumé III and Marcu 2005; Denis and Baldridge 2007a; Culotta et al. 2007). While these research efforts show that performance gains can be achieved by developing more robust statistical inference techniques, we note that current coreference resolution systems rely on rather shallow features, such as the distance between the coreferent expressions, string matching, and linguistic form[44]. These shallow features are not sufficient to correctly identify many of the coreferential relations between expressions in texts. As an example, consider the following fragment from the Automatic Content Extraction (ACE) 2002 data.

(61) **Israel** will ask the United States to delay a military strike against Iraq until **the Jewish state** is fully prepared for a possible Iraqi attack with non-conventional weapons, the defense minister said in remarks published Friday. [...] **Israel** is equipping **its residents** with gas masks and preparing kits with antidotes. [...] **Israel**'s armed forces chief, Lt. Gen. Amnon Lipkin-Shahak, appeared on national television Friday night in attempt to reassure **a jittery public**. "When events are concrete and actually unfolding, we will include **the entire population** in the measures we decide are appropriate to take," Shahak said in an interview with Channel Two Television. (NWIRE/APW19980213.1305)

In order to correctly resolve the coreferent expressions highlighted in bold (which are all annotated as coreferent in the ACE data), lexical semantic and encyclopedic knowledge is required, i.e. that *Israel* is a *Jewish state*. To induce the correct coreference links in this example it seems therefore that one should connect those referring expressions which are strongly related to each other by some (possibly unspecified) relations[45]. Additional cases that highlight the need for lexical knowledge include finding *synonymy* between common nouns and *instance-of* relations between named entities and common nouns as in Examples 62 and 63 from ACE:

(62) After almost eight months of investigation, the Justice Department is preparing to challenge a plan by Murdoch's News Corp. and MCI Communications Corp. to merge their U.S. satellite TV operations with Primestar Partners, **the nation**'s second-largest satellite TV company, sources familiar with the inquiry say. [...] The slot is highly valuable because it is one of only three available in **this country** from which a satellite can beam TV programs across most of North America simultaneously. (NPAPER/9803.62)

(63) A new report reveals more problems at **the Internal Revenue Service**. A broad review of **the agency** found it used improper tactics in evaluating **IRS** employees at many **IRS** offices across the country. (BNEWS/CNN19980113.1600.1130)

To correctly resolve the coreference links in the two examples above it seems crucial to provide the system with information such as the fact that *nation* can be synonym with *country* (Example 62) and that the *IRS* is a (U.S. government) *agency* (Example 63).

[44] As noted, among others, by the feature analysis study presented in Bengtson and Roth (2008).

[45] In the ACE annotation the notion of coreference overlaps with that of *metonymy* – i.e. *Israel* can also refers to the country's population, rather than the country itself–so that completing the coreference chain in the example also requires identifying *residents* and *public* as mentions of Israel, a rather dubious assumption. We come back to this problem later in the evaluation section.

In practice, even with recent improvements in the learning framework, the accuracy of coreference resolution (both of pronouns and definite noun phrases) has remained imperfect due to the fact that there is no easy way to include the common sense knowledge that humans use. There are two main strands of research on using lexical and commonsense knowledge to help coreference on unrestricted text: one devoted to the use of information about the semantic structure of verbs and primarily concerned with improving precision in pronoun resolution; and a second one concerned with the use of hyponymy information to improve recall in the resolution of nominals.

## 7.1 Using Semantic Compatibility Information for Pronoun Resolution

For pronouns, several approaches aiming to use **selectional preferences** have been tried out, with varying success.

Dagan and Itai (1990) present an approach where automatic parses from a 60 million word corpus are used to extract statistics about subject-verb and object-verb cooccurrences, which are then used as a model of selectional preferences. Using a hand-selected sample of *it* pronouns where the antecedent as well as one or more other candidates compatible in number, gender and syntactic position were in the same sentence, Dagan and Itai found that in 64% of the cases, antecedent and candidates all occurred at least five times in the parsed corpus, and of these, 87% had the correct antecedent allowed by their selectional preference model, and in about half of these cases, the antecedent was the only one that fits the selectional preferences. While this approach clearly steers free of most problems that would hinder the use in a full coreference system – among others, noise in the determination of agreement features, classification of named entities, or treatment of infrequent words – it has certainly inspired further research that aims at using selectional preferences. For instance, Dagan et al. (1995) present a post-processor for a knowledge-rich, rule-based pronoun resolver (Lappin and Leass 1994, Section 4.3), which overrides the system's coreference decisions based on predicate-argument co-occurrence statistics, i.e. how many times a pronoun occurs as the argument of a certain predicate.

Ge et al. (1998) use a probabilistic model of selectional preferences that is derived from a statistical parser. While the difference in performance obtained by including selectional preferences is not statistically significant, the author argue that the inclusion of this information source gives a visible improvement.

(Kehler et al. 2004) discuss the integration of selectional preference features in a maximum-entropy based pronoun resolver; they find that in the absence of number or gender agreement features, selectional preference features give a very visible loss in accuracy, whereas otherwise they yield a small (but not statistically significant) improvement over a model with no selectional preferences applied to general text. By contrast, Bean and Riloff (2004) demonstrated that given sufficient amounts of data, selectional preferences inferred from corpora did result in significant improvements for coreference in general.

## 7.2 Using Semantic Compatibility Information for Nominal Resolution

State-of-the-art coreference resolution systems all make use of some kind of knowledge to quantify the degree of semantic compatibility holding between nominals. For instance, Soon et al. (2001) include a semantic class feature in their set of features (SEMCLASS, Table 3): this checks whether the first WordNet senses of the head nouns of the markables in a mention pair are semantically compatible, based on class relations in a simplified taxonomy (see *ibidem*, pages 524–525). Similarly, most current systems make use of a basic semantic typing system like that used by Soon *et al.*. These types, generally extracted from the output of a named entity tagger for proper nouns and from WordNet for nominals, can also provide some information about **animacy** – but cannot tell for instance whether a particular instance of organization is used as an animated agent or not. Machine learning methods for automatically determining animacy have been developed e.g. by Orasan and Evans (2007) among others. Another widespread semantic feature of today's systems concerns gender agreement, namely the fact that almost all current systems have extensive gender tables that specify the gender not only of 'obvious' names like *John* and *Mary* but also of more unusual and foreign names like *Andrea* or *Orhan*. These tables used to be compiled by hand but are now often extracted from large corpora and the Web (Ge et al. 1998; Bergsma 2005)

To resolve complex cases of anaphoric relations involving lexical, encyclopedic and world knowledge one needs information sources that will help identify *lexical relations* such either synonymity (*the suit . . . the lawsuit*), near-synonymity (*the house . . . the home*), hypernymy (*the mansion . . . the house*) or instance relations (*Bach . . . the composer*) – cf. also the examples from the previous section. However, coreference relations can be also triggered my more broad associative relations between an anaphor and its antecedent (*the house. . . the door*).

We can classify the methods which aim to use knowledge for the resolution of nominals based on the information sources they use. On one end of the spectrum, we find approaches which make use of traditional, hand-crafted knowledge bases – most commonly, wordnets such as Princeton WordNet (Fellbaum 1998) and gazetteer lists, which contain a large list of names belonging to one semantic category (such as persons, organizations, airports, etc.) Successful rule-based approaches at the MUC-6 and MUC-7 competitions already exploited path search in WordNet (Lin 1995) or gazetteer lists organized in a shallow taxonomy (Kameyama 1997). More recently, Vieira and Poesio (2000) and Harabagiu et al. (2001) explored the use of WordNet for different coreference resolution subtasks, such as resolving bridging references, *other*- and definite NP anaphora, and MUC-style coreference resolution. All of them present systems which infer coreference relations by means of WordNet search – to check whether the referring expressions are synonyms or in a hyponymy or hypernymy relation with each other.

On the other end of the scale, the limited coverage of existing lexicons, as well as the need of semantic relatedness for resolving ACE-style mentions, has also pushed researchers to investigate corpus-based approaches to mine semantic relations for coreference. (Gasperin and Vieira 2004; Markert and Nissim 2005; Garera and Yarowsky 2006).

Approaches using manually built knowledge bases rely on high-quality knowledge manually inputted by human experts at the cost of a limited coverage, whereas proposals making use of information automatically extracted from corpora achieve a higher coverage for a quality lower than that of humans. A proposal to stake out a middle ground between these two extremes is presented in Ponzetto and Strube (2006), who make use of the category structure of Wikipedia – which is not a taxonomy as such, but is generally organized according to specificity (Ponzetto and Strube 2007b,a).

Finally, there are approaches that aim to extract statistics about the behavior of words with a more direct focus on coreference, which becomes increasingly attractive with the growing size of available referentially annotated corpora. In practice, these approaches use the annotated data from coreferentially annotated coreference corpora to find common regularities. Ji et al. (2005) develop a two-stage approach where the probabilities output from a MaxEnt classifier are rescored by adding information about the semantic relations between the two candidate mentions. These relations are automatically output by a relation tagger, which is trained on a corpus annotated with the semantic relations from the ACE 2004 relation ontology. The approach of Ponzetto and Strube (2006) uses features based on semantic role labeling, which also aims at learning regularities from the annotated corpus itself. While such approaches are empirically successful, the exact nature of the relations learned is neither investigated nor discussed, which is regrettable since they could (potentially) uncover different insights or strategies from those realized by the approaches based on existing resources and/or unsupervised learning. Finally, Ng (2007) extracts statistics regarding the anaphoricity of noun phrases, and of likely coreference (allowing to exploit the fact that texts from one time period and country typically have the same antecedent for mentions such as "*the president*").

### 7.2.1 Using Knowledge Bases

As in many other subfields of CL a first solution to the need of external knowledge is to turn to traditional, hand-crafted knowledge resources such as the semantic lexicon provided by WordNet (Fellbaum 1998).

**Poesio et al. (1997)** investigate both coreferent and non-coreferent bridging relations between definite descriptions and mentions in the previous text. They break up the bridging descriptions into six classes, motivated mostly by processing considerations:

- lexical relations between the heads: synonymy, hypernymy, meronymy (e.g.: *new album*... *the record*)
- instance relations linking definite descriptions to proper names (e.g.: *Bach* ... *the composer*)
- modifiers in compound nouns (e.g.: *discount packages*... *the discounts*)
- event entities introduced by VPs (*Kadane Oil Co. is currently drilling*... *the activity*)
- associative bridging on a discourse topic (*the industry* in a text on oil companies)

- more complex inferential relations, including causal relations

In the 204 bridging definite descriptions from the corpus they analyzed, 19% had a lexical relation between common noun heads and 24% were definite descriptions referring to named entities. The system tries to resolve coreferent instances involving lexical relations using WordNet by looking for synonyms, hypernyms and coordinates sisters in the taxonomy. They find that an approach combining the relation information in WordNet with distance information (i.e., resolving to the nearest element that has a compatible distance) yields a precision between 36% (for synonyms) and 20% (for coordinate sisters) with an overall recall – namely the proportion of instances where a relevant relation between anaphor and antecedent could be found in WordNet – of 39%. For proper names, they assign a semantic type, based on appositive constructions and honorifics or other name parts, such as *Mr., Co., Inc.* as cues. With a mechanism for propagating the type to other entities (e.g. from *Mr. Morishita* to *Morishita*), they can correctly classify 69% of the names in the corpus, and are able to resolve 52% of the definite descriptions referring to named entities.

**Harabagiu et al. (2001)** go beyond synonymy and hypernymy and consider more general paths in WordNet that they find between anaphor-antecedent pairs found in the training data. To find candidate pairs, they filter out anaphoric expressions with an antecedent that can be found with knowledge-poor methods, such as string matching, appositions, name variation, or the most salient compatible antecedent.

For the remaining anaphoric definite expressions, they look for anaphor-antecedent pairs that are related by at most five of the following relation types in the WordNet graph:

- SYNONYM, ISA/R-ISA and HAS-PART correspond to synonymy and hypernymy and meronymy relations.
- GLOSS/DEFINES connect a word in a synset to the word used to define it.
- IN-GLOSS/IN-DEFINITION connects an element of the synset with one of the first words in its definition.
- MORPHO-DERIVATION connects morphologically related words.
- COLLIDE-SENSE connects synsets of the same word (homonyms or polysemous senses).

Harabagiu *et al.* use three factors to measure the confidence of a WordNet path to predict a coreference relation. The first factor is a binary-valued flag that is set to 1 if another coreference chain contains mentions in the same nominal as the anaphor and the antecedent – e.g. given *Massimo's son* and *his bicycle*, if *son* and *his* have been previously found to be coreferent, the factor for the former pair is set to 1, else to 0. The second factor prefers "stronger" relations where each WordNet relation type is assigned a weight ranging from 1.0 for SYNONYM over 0.9 for ISA and GLOSS down to 0.3 for IN-GLOSS). The weight is averaged over the relation types occurring in the path, with multiple occurrences of a relation weighted down by a factor corresponding to their

number of occurrences. Additionally, the total number of different relations is used to weight down longer paths. As an example, a path with one HASPART edge (weight $0.7$) and two ISA edges (weight $0.9$) would receive a weight of $\frac{1}{2} \cdot \left(\frac{0.7}{1} + \frac{0.9}{2}\right) \approx 0.57$, whereas a path with two ISA edges would receive a score of $\frac{1}{1} \cdot \frac{0.9}{2}$. Finally, the last factor is a semantic measure inspired by the *tf-idf* weighting scheme and it is determined by considering the search space built when considering at most five combinations of the semantic relations defined above, starting from either of the synset a nominal can be mapped to. The overall confidence of a path is given by a weighted harmonic mean of the three factors. Confidence scores are then used to iteratively select the paths with the highest confidence as rules of the system.

By exploiting lexical knowledge from WordNet in a flexible way, Harabagiu *et al.*'s proposal is able to achieve competitive results, i.e. a MUC F-measure of $81.9\%$. However the authors do not offer a motivation for the exact choice of the many weighting functions and threshold they use. As a result, it is not always clear whether the exact choice is due to an intuition about the problem to be solved or the result of ad-hoc modifications to a function until a satisfactory result is reached. However, at the same time, their work contains several noteworthy insights. First, Harabagiu *et al.* do not always use the direct antecedent, but instead allow their system to learn a relation to an antecedent that is further away. Second, they use WordNet to derive a *general distance measure*, including the definitions contained in the glosses, which yield a markedly different information source from Poesio et al.'s earlier approach (more focused on using the information in WordNet as it is and getting highly precise subsumption and synonymy predictions). Finally, they use *a global clustering-based model* that can make use of more reliable decisions (e.g. for possessive pronouns) to influence other decisions (for the possessed NPs) where the coreference between the possessors provides additional information.

**Ponzetto and Strube** Although it provides a well-structured and cognitively motivated semantic lexicon and it has been extensively leveraged for a variety of CL applications, WordNet suffers from limited coverage – cf. e.g. Schütze and Pedersen (1995) in the context of using word senses for information retrieval. The main focus on providing a complete sense repository for English common nouns has led to disregard much named entity information, that is, information about individuals. This is evidenced by the limited number of individuals it contains – only $9.4\%$ according to Miller and Hristea (2006) – and by the fact that an individual-specific relation such as instantiation has been introduced only recently with version 2.1. In order to overcome this limited coverage, Strube and Ponzetto (2006); Ponzetto and Strube (2006) propose to use the system of categories found in an online collaboratively generated encyclopedia, Wikipedia to compute semantic compatibility scores between nominals.

Wikipedia is a collaborative open source medium edited by volunteers and provides a very large domain-independent encyclopedic repository: the English version, as of 2 August 2010, contains more than 3,362,000 articles with tens of millions of internal hyperlinks. There are at least three main features which make Wikipedia a sound choice as a knowledge repository for AI and CL applications:

1. **high coverage**: it contains a large amount of information, in particular at the

instance level;

2. **multilingual**: it is available with a uniform structure for hundreds of languages;

3. **up-to-date**: it includes continuously updated content, which provides current information.

Since Wikipedia exists only since 2001 and has been considered a reliable source of information for an even shorter amount of time Giles (2005), researchers in CL have only begun recently to work with its content or use it as a resource. Wikipedia has been used successfully for a multitude of AI and NLP applications. These include both preprocessing tasks such as named entity (Bunescu and Paşca 2006; Cucerzan 2007) and word sense disambiguation (Mihalcea 2007; Ponzetto and Navigli 2010), text categorization (Gabrilovich and Markovitch 2006), computing semantic similarity of texts (Gabrilovich and Markovitch 2007; Milne and Witten 2008a) and keyword extraction (Csomai and Mihalcea 2008b; Milne and Witten 2008b), as well as full-fledged, end-user applications such as question answering (Ahn et al. 2004, 2005; Lo and Lam 2006, inter alia), topic-driven multi-document summarization (Nastase 2008), text generation (Sauper and Barzilay 2009) and cross-lingual information retrieval (Cimiano et al. 2009). In addition, since May 2004 Wikipedia has provided a thematic categorization scheme by means of its *categories*: articles can be assigned to one or more categories, which are further categorized to provide a so-called "category tree". In practice, this "tree" is not designed as a strict hierarchy, but allows multiple categorization schemes to coexist simultaneously.

The proposal of Ponzetto and Strube is to use the category tree from Wikipedia as an unlabeled semantic network and use it to compute semantic compatibility scores by means of taxonomy-based semantic distance measures previously developed for WordNet (Rada et al. 1989; Wu and Palmer 1994; Leacock and Chodorow 1998; Seco et al. 2004). The approach start with the baseline system from Soon et al. (2001) and extends it with features capturing different semantic knowledge sources. These features represent semantic distance scores computed from WordNet and Wikipedia. Performance variations with respect to the baseline coreference resolver indicate the quality of the information extracted from WordNet and Wikipedia, as well as provide an evaluation of their relative impact. The results obtained by using WordNet and Wikipedia are promising: on the ACE 2003 dataset, the authors find a large improvement in terms of recall on the broadcast news (whereas the results on the newswire section are modest) with Wikipedia-based scores performing on a par with WordNet. WordNet and Wikipedia features tend to consistently increase performance on common nouns. However, semantic relatedness is found not to always improve the performance on proper names, where features such as string matching and alias seem to suffice.

While measures of semantic relatedness computed from Wikipedia are competitive with those computed using WordNet, applications such as coreference typically need a tighter notion of semantic similarity. In order to compute semantic similarity, one needs a taxonomy, since approaches to measuring semantic similarity that rely on lexical resources use paths based on ISA relations only. However, the system of categories in Wikipedia is not a taxonomy with a full-fledged subsumption hierarchy, but only a thematically organized thesaurus. Ponzetto and Strube (2007a) accordingly

present a set of lightweight heuristics to generate a large scale taxonomy from the network of Wikipedia categories by using the syntactic structure of the category labels, the connectivity of the graph and lexico-syntactic patterns in very large corpora. Semantic similarity computed from the Wikipedia taxonomy is evaluated extrinsically by Ponzetto (2010), who uses them as features of a supervised coreference resolver in the same way as the previously used semantic relatedness scores. The evaluation on the ACE-2 data show that using relatedness works better than computing paths along the ISA hierarchy. Semantic relatedness always yields better results than using similarity scores: but while this is a counterintuitive result, the author argues that this behaviour is an artifact of the annotations in ACE. The use of the Geo-political entities (GPE) class in the ACE data allows for coreferential links such *Beijing ... China* and mixes therefore metonymy with coreference phenomena – cf. also our discussion in Section 5.1 and the link between *Israel* and *its residents* in the Example 61. To generate these coreference links one needs indeed a more permissive notion of semantic compatibility, i.e. semantic relatedness. Using ISA relations only is in fact expected to work better for data modeling coreference as identity instead.

### 7.2.2 Acquiring Knowledge from Text

Hand-crafted taxonomies typically have problems with respect to their limited coverage. Accordingly, in the last decades, many research efforts in CL have been concentrated in the field of knowledge acquisition, which aims at automatically extracting structured information from unstructured text sources. Perhaps even more crucially, automatically extracting knowledge from text allows to acquire information which is expected to be relevant in context, cf. the observation from Markert and Nissim (2005) that some relations between mentions, such as *age* being a *risk factor* are only relevant in specific contexts – thus being undesirable in a general-purpose ontology – which implies that extracting such relations from text could be more suitable for a coreference system.

**Poesio et al. (1998)** discuss the use of a distributional similarity measure similar to HAL (Lund et al. 1995) by using second-order co-occurrences of words with a fixed-size context window on the same data set as Poesio, Vieira and Teufel (1997). Using the count vector of words co-occurring with a given word and different vector distances, they use the British National Corpus (Clear 1993) to learn an association measure of words using different combinations of window sizes and similarity metrics, as well as a variant with lemmatization and part-of-speech marking. In the best configuration they achieve 22.2% precision for synonymy/hyperonymy/meronymy cases overall, against 39% with the WordNet-based approach. The more complex inferential relations are found to be the only area where the association measure outperforms the more precise, i.e. taxonomy-based, methods.

**Gasperin and Vieira (2004)** use a word similarity measure from (Gasperin et al. 2001) very similar to the one introduced by Lin (1998a). In contrast to Poesio, Schulte im Walde, and Brew's work, they do not resolve to the semantically closest noun, but

instead build lists of globally most similar words (a so-called *distributional thesaurus*), and enable the resolution to antecedents that are in the most-similar list of the anaphoric definite, where the antecedent has the anaphoric definite in its most-similar list, or where the two lists overlap. Working on Portuguese data, Gasperin and Vieira find that they reach similar levels of resolution accuracy to the results of Poesio, Schulte im Walde and Brew's (1998) with a window-based association metric.

Both Poesio et al. (1998) and Gasperin and Vieira (2004) aim at exploiting **association measures** such as HAL or **distributional similarity metrics**. The methods rely therefore on a distributional similarity hypothesis: similar concepts occur in similar contexts, which means that using vectors of frequencies, probabilities or association strengths to represent the distribution of the contexts a lemma occurs in and using numerical similarity measures over these vectors allows one to use these vectors as a coarse approximation to conceptual similarity. An alternative to making use of metrics of unstructured, associative knowledge, as provided by the cooccurence of words in large corpora, is offered instead by mining the occurrence of **lexico-syntactic patterns** within large corpora. The occurrence of these patterns is taken to be indicative of particular lexical relations, such as the patterns introduced by Hearst (1992) for hypernymy (e.g., *Ys* such as *X*, *X* and other *Ys*) or by Berland and Charniak (1999) for part-of relations (*Y*'s *X*, *X* of *Y*). These semantic relations can be then used to help identify strongly related mention pairs as coreferent.

**Poesio et al. (2002)** use patterns including '(the) *X* of *Y*' and '*Y*'s *X*' to acquire indicators for meronymy in a partially parsed version of the British National Corpus, and find that selecting antecedents with the mutual information statistic using these associations results in much better recall for meronymy relations than either WordNet or the vector-based method. The pattern-based approach requires large corpora to achieve a reasonable recall: this is because patterns occur rarely in corpora. Accordingly, researchers in CL turned in the last years to the Web as a very large resource of linguistic data and developed a variety of knowledge acquisition methodologies (typically using weakly supervised techniques) to mine this large repository of text. Markert and Nissim (2005) for instance, do a study similar to Poesio et al. (2002) for hypernym relations where they also use the counts returned by a web search engine. They find that, due to the much greater size of the World Wide Web, the results using web searches are much better than using a corpus such as the British National Corpus. In a similar fashion, Poesio, Mehta, Maroudas, and Hitzeman (2004b) use a multilayer perceptron with features including simple graph distance in WordNet (indicating the number of nodes between the anaphor and the potential antecedent) and a feature based on the raw count of matches for a search engine query using a meronymy pattern. To express salience, Poesio *et al.* include the sentence distance to the anaphor, but also whether it is in first-mention position, or if any preceding mention of the entity had been in first-mention position.

**Hybrid approaches.** An integrated approach is presented in Daumé III and Marcu (2005), who use several classes of features. Besides including WordNet graph distance and WordNet information for preceding/following verbs (in an attempt to let the

coreference resolver learn approximate selectional preferences in a supervised way), they also use name-nominal instance lists mined from a large news corpus (Fleischman et al. 2003), as well as similar data mined from a huge (138GB) web corpus (Ravichandran et al. 2005). They also used several large gazetteer lists of countries cities, islands, ports, provinces, states, airport locations and company names, as well as a list of group terms that may be referenced with a plural term. Bunescu (2003) proposes to use discourse-based patterns in conjunction with web queries to resolve bridging anaphora: To resolve an associate definite description to an antecedent, he embeds anaphor and antecedent noun phrases in a pattern "*Y*. The *X verb*", where verb is subsequently filled with a list of auxiliary and modal verbs, and results are scored using pointwise mutual information. On a test set of associative bridging anaphora sampled from the Brown corpus section of the Penn Treebank, Bunescu's approach reaches a precision of 53% at a recall of 22.7%. A very similar approach is presented by Garera and Yarowsky (2006), who investigate the use of an unsupervised model to extracts hypernym relations from cooccurrence statistics for resolving definite nominals. The method aims at exploiting association metric scores to find likely categories for named entities: using the English Gigaword corpus as source of textual data, they evaluate on a hand-selected sample and show that, when using the same corpus, their association measure performs better than Hearst-style patterns.

### 7.2.3 Inducing Knowledge from Anaphorically Annotated Corpora

An alternative to knowledge-lean approaches leveraging existing resources and unsupervised approaches extracting structured knowledge from unstructured textual resources is to learn semantic regularities directly from the same coreferentially annotated corpora used to train supervised coreference resolvers.

**Ji et al. (2005)** use heuristics to integrate constraints from relations between mentions with a coreference resolver. The methodology consists of a two-stage approach where the probabilities output from a MaxEnt classifier are rescored by adding information about the semantic relations between the two candidate mentions. These relations are automatically output by a relation tagger, which is trained on a corpus annotated with the semantic relations from the ACE 2004 relation ontology. Given a candidate pair 1.B and 2.B and the respective mentions 1.A and 2.A they are related to in the same document, they identify three lightweight rules to identify configurations informative of coreference:

1. If the relation between 1.A and 1.B is the same as the relation between 2.A and 2.B, and 1A and 2A don't corefer, then 1.B and 2.B are less likely to corefer.

2. If the relation between 1.A and 1.B is different from the relation between 2.A and 2.B and 1.A is coreferent with 2.A, then 1.B and 2.B are less likely to corefer.

3. If the relation between 1.A and 1.B is the same as the relation between 2.A and 2.B and 1.A is coreferent with 2.A, then 1.B and 2.B are more likely to corefer.

While Ji *et al.* argue that the second rule usually has high accuracy independently of the particular relation, the accuracy of the other two rules depends on the particular relation. For example, the chairman of a company, which has a EMP-ORG/Employ-Executive relation, may be more likely to remain the same chairman across the text than a spokesperson of that company, which is in the EMP-ORG/Employ-Staff relation to it. Accordingly, the system retain only those rule instantiated with a *specific* ACE relation which have a precision of 70% or more, yielding 58 rule instances. For instances that still have lower precision, they try conjoining additional preconditions such as the absence of temporal modifiers such as "current" and "former", high confidence for the original coreference decisions, substring matching and/or head matching. In this way, they can recover 24 additional reliable rules that consist of one of the weaker rules plus combinations of at most 3 of the additional restrictions. They evaluate the system, trained on the ACE 2002 and ACE 2003 training corpora, on the ACE 2004 evaluation data and provide two types of evaluation: the first uses Vilain et al's scoring scheme, but uses perfect mentions, whereas the second uses system mentions, but ignore in the evaluation any mention that is not both in the system and key response. Using these two evaluation methods, they get an improvement in F-measure of about 2% in every case. In the main text of the paper, Ji *et al.* report an improvement in F-measure from 80.1% to 82.4%, largely due to a large gain in recall. These numbers are relatively high due to the fact that Ji *et al.* use a relaxed evaluation setting disregarding spurious links. A strict evaluation on exact mentions is able instead to yield an improvement in F-measure from 62.8% to 64.2% on the newswire section of the ACE corpus.

**Ng (2007)** includes an ACE-specific semantic class feature that achieves superior results to Soon et al.'s method using WordNet by looking at apposition relations between named entities and common nouns in a large corpus to find better fitting semantic classes than using WordNet alone. In addition, he uses a semantic similarity feature similar to the one introduced by Gasperin *et al.* (indicating if one NP is among the 5 distributionally most-similar items of the other), and two features that are learnt from an held-out subset of the training data:

- a *pattern-based* feature, encoding the the span between mentions by means of a variety of patterns, e.g. as sequences of NP chunk tokens;
- an *anaphoricity* feature which encodes how often an NP is seen as a discourse-old noun phrase in the corpus;
- a *coreferentiality* feature modeling the probability that two noun phrases are coreferent, estimated by looking at pairs occurring in the corpus.

Training on the whole ACE-2 corpus, Ng is able to improve the MUC score from 62.0% on the ACE-2 merged test set to 64.5% using all the features except the pattern-based one.

**Yang and Su (2007)** present an approach to select patterns as features for a supervised coreference resolver. Starting from coreferent pairs found in the training data

such as "*Bill Clinton*" and "*President*" (or, due to the annotation scheme of the ACE corpora, "*Beijing*" and "*China*", cf. the example 61), they extract patterns from Wikipedia where a pattern is defined as the context that occurs between the two mention candidates – e.g. "(Bill Clinton) *is elected* (President)". To select those patterns that identify coreferent pairs with a high precision, the method filters out in a first step those that extracts more non-coreferent pairs than coreferent ones in the training data. In a subsequent step, patterns are ranked – based either on raw frequency or on a reliability score – and the 100 top-ranking patterns are kept. In the case of the frequency-based approach, a feature is created for each pattern that indicates the frequency of that particular word pair with the pattern in the Wikipedia data. For the other approaches, they calculate a *reliability metric* for each pattern (determined by summing the pointwise mutual information values between a pair of noun phrases and the pattern, over all coreferent pairs from the training data). The score for a given pattern and a pair of fillers is then determined as the value of the reliability of that pattern multiplied by the positive mutual information between positive mention pairs. Yang and Su apply these features in a coreference resolution system similar to the one described by Ng and Cardie (2002b) on the ACE-2 corpus. Using the reliability-based single relatedness feature for proper names (the setting they found to work best) results in an improvement from 64.9% F-measure to 67.1% on the newswire portion, 64.9% to 65.0% on the newspaper portion, and from 62.0% to 62.7% on the broadcast news part.

# 8 Conclusions

In this paper we presented a survey of computational approaches to anaphora resolution. In the first part we presented the available linguistic and psychological evidence about anaphora and the resolution of anaphora, and current linguistic theories about the interpretation of anaphoric expressions. We also discussed early work on anaphora resolution in which these theories about anaphora were directly encoded. Next, we concentrated on the so-called 'data-driven' revolution, namely the creation of modern, substantial corpora of anaphoric information, which allowed both an extensive empirical study of the phenomenon and the development of statistical methods for this phenomenon. We concluded by presenting a variety of approaches which aim to acquire and make use of lexical and encyclopedic knowledge for anaphora resolution.

The aim of this survey was to provide a comprehensive survey of the last thirty years of work on the computational treatment of anaphora resolution by framing it within the context of other fields as well (linguistics and psycholinguistics, in particular). Our presentation shows that previous research efforts have made it possible to achieve a fair understanding of the phenomenon of anaphora and the main factor affecting it. However, from a computational perspective we crucially argue that there are still a variety of open issues to achieve robust models of anaphora resolution.

**New and More Data.** Addressing this challenge will require first to develop methods to annotate this type of anaphoric information, and either to create larger annotated resources or to develop better ways of using the annotation, such as active learning. We argue that the recent creation of the OntoNotes corpus (Hovy et al. 2006) represents a

substantial step in this direction; a further contribution to the problem of creating larger resources may come from the **games for a purpose** approach pioneered by von Ahn (2006) and applied to anaphoric annotation through the development of the PhraseDetectives game (Chamberlain et al. 2008).

**Better Methods for Nominal Anaphora Resolution.** Second, it will be necessary to develop much more sophisticated models of anaphora resolution. An encouraging development is the fact that 'coreference' is beginning to attract the attention of the machine learning community, which has been developing much more sophisticated models of the task (Culotta et al. 2007; Klenner and Ailloud 2008; Haghighi and Klein 2010, *inter alia*). However, we argue that current coreference resolution systems still mostly rely on rather shallow features which are not sufficient to correctly identify many of the coreferential relations between expressions in texts. Accordingly, a key challenge for the future is to improve the use of lexical and commonsense knowledge, both in terms of better (i.e. fully unsupervised, large-scale and multilingual) methods of knowledge acquisition and ways of deploying that knowledge effectively for coreference resolution.

**Modeling complex anaphoric relations.** The empirical shift in anaphora resolution has had many beneficial effects on the field, resulting in a much clearer picture of the phenomenon and its complexity. But while anaphora is a pervasive phenomenon in natural language, modern CL research has focused mostly on those anaphoric relations for which a substantial amount of annotated data exist: that is, nominal anaphora to antecedents introduced with nominal expressions. It is therefore fair to say that statistical models of anaphora resolution so far have only scratched the surface of the phenomenon, and the contributions to the linguistic understanding of the phenomenon have been few. One major challenge for the next decade will be to expand the range of anaphoric phenomena considered and accordingly to go beyond nominal anaphora – e.g., develop models able to deal with reference to abstract objects, bridging and ellipsis – as well as develop robust methods to cope with variations in the domain and genre (as shown in (Müller 2008), current results with spoken dialogue are very mediocre).

**Applications.** Last but not least, future work will need to address the utility of coreference resolution for end-user applications. Very encouraging results have come from summarization (Steinberger et al. 2007) and opinion mining Jakob and Gurevych (2010) but in other types of application the improvements deriving from anaphora resolution have been less pronounced (Morton 2000; Sanchez-Graillet et al. 2006).

Other Languages

Most of the work reported in the previous chapters works with English data, in all probability due to the presence of high-quality preprocessing tools, large annotated corpora as well as a sufficient number of researchers working on this topic within the larger area of computational linguistics. At the same time, **corpora** for other domains (e.g., medical text) and other languages already exist (see Table 2). Some of these corpora, such as the Catalan and Spanish ANCORA corpus, the Dutch COREA corpus, the

German TüBa-D/Z, the Italian LiveMemories corpus, or the Japanese NAIST Text Corpus, rival their English counterparts in size, whereas others, such as Zeisler's Tibetan corpus, are relatively small and mostly intended for theoretical corpus studies in these resource-poor languages. Nevertheless, recent community efforts, such as the SemEval 2010 task on multilingual coreference resolution (Recasens et al. 2010), concentrated on enabling and encouraging the development of coreferent resolution systems for a variety of languages.

While it is not possible to present all the computational work that has been done on these languages, we will present the results for three languages, German, Dutch and Japanese, to illustrate the state of the art in computational anaphora resolution for these languages but also in order to illustrate some new ideas about anaphora resolution arising from this work.

**German.** A smaller but still significant amount of work has been done, some concentrating only on the resolution of pronominal anaphora (Stuckardt 2004; Schiehlen 2004; Kouchnir 2004; Hinrichs et al. 2005a), or only on the resolution of names and definite noun phrases (Versley 2006) but some also tackling the full task of coreference resolution for German (Hartrumpf 2001; Strube et al. 2002; Klenner and Ailloud 2008).

Hartrumpf (2001) uses a statistical backoff model to choose between antecedent candidates that have been identified by manually designed rules. Sentences are processed by a word-expert-based parser (Helbig and Hartrumpf 1997), with a chunking parse being put in place where parsing fails. The rules identifying antecedent candidates rely on extensive knowledge in the semantic lexicon HaGenLex (Hartrumpf et al. 2003) to provide information on group-denoting nouns (in the case of coreference with non-matching number), semantic class (in the case of coreferent bridging) and synonymy information, in addition to sentence distance information. The resolver then ranks candidates using a statistical backoff model that backs off to subsets of the candidates until matching examples are found. Additionally, global semantic compatibility over a coreference set is enforced; for this, a beam search over all possible partitions is carried out, pruning solutions with lower overall scores.

Strube et al. (2002) adapted the coreference algorithm of Soon et al. (2001) to German data: in addition to features like grammatical function and a coarse semantic classification (both of which were added to the data by hand), they use minimum edit distance between mentions to improve the recall on definite descriptions and names. Some elements of Strube *et al.*'s experiments, notably the pre-annotation of perfect (coarse) semantic class and grammatical function values, as well as perfect markable boundaries, would be difficult to achieve in practice.

Klenner and Ailloud (2008) use an easiest-first approach to globally optimize the consistency of coreference sets, using a constraint propagation algorithm similar to earlier rule-based approaches (Lin 1995), and the classifications of a a memory-based learning classifier on syntactic and distance features in conjunction with head matching. Like in Lin's approach, they start by the most confident decisions on coreference or non-coreference (easiest-first strategy), but by keeping the 16 best-scoring partial solutions, this approach is also successful in avoiding a larger number of search errors.

**Dutch.** Hoste (2005) develops a machine learning approach to coreference resolution for Dutch also based on the mention-pair model but that differs from the approach by Soon *et al.* in that separate classifiers are trained for each type of NP – i.e. a different classifier for pronouns, common nouns and proper names. In addition, her work explores the effect of a variety of decisions concerning the machine learning architecture, including 'lazy' vs. 'eager' learning, different ways of handling the skewness in the distribution of training instances – i.e., the preponderance of negative instances in the set of training examples given to the coreference classifier – as well as the optimization of the parameters of the coreference learner.

**Japanese.** While the differences between English and German – synthetic compounds, freer word order, and morphological variation, are small enough that systems perform in generally similar fashion even though some features that are very simple in one language are more complicated in the other (as is the case with string matching as an approximation to modifier and head equality), the situation is markedly different in Japanese and similar languages, where arbitrary verb arguments may be realized as a zero pronoun (i.e., completely left out), and the detection of zeroes is as challenging a task as their resolution.

While earlier, theoretical work focuses on the choice of antecedents (see e.g., (Kameyama 1985; Walker et al. 1994), for work discussing the application of the centering model to Japanese), and some recent practical work also exclusively treats the resolution without going into the detection of zero pronouns, for example the work of Iida et al. (2003b), who present a tournament model for the resolution of zero pronouns and incorporate a centering-based salience model as well as chain length and selectional preferences using both a hand-crafted lexicon and noun-verb pairs extracted from a large unannotated corpus.

Seki et al. (2002) propose a model in which preferences for filled arguments are extracting from a large unannotated corpus by noting the co-occurrence of overt NPs as arguments for a verb, and using these as indicators that a non-filled argument slot should be filled by (the antecedent of) a zero pronoun. This probabilistic detection model is then combined with other features including distance, selectional preferences, and information about embedding, to yield a probabilistic model for the integrated detection and resolution of zero pronouns. In contrast to Seki *et al.*, Iida et al. (2007a) use a supervised approach to learn syntactic patterns using tree boosting, an approach that learns a weighting on informative parts from tree structures given in the training data.

Sasano et al. (2008) present a more refined model for unsupervised zero pronoun detection and resolution where statistics about both the arguments in a caseframe and selectional preferences for the fillers are acquired from a corpus and used in a statistical model that serves both to select the correct caseframe using the overt arguments and and to provide selectional preferences for the antecedent. In addition to this information, they also encode the syntactic relation (path between anaphor and antecedent) into one of twelve classes, and compute a score for cumulative decayed salience similar to the one proposed by Lappin and Leass (1994).

Off-the-shelf Anaphora Resolvers

Several anaphora resolution tools for English coreference are publically available.

**GUITAR.** GuiTAR (Kabadjov 2007; Steinberger et al. 2007)[46] is a highly modular and easily modifiable system developed in Java that comes with an implementation of Vieira and Poesio's (2000) algorithm for the resolution of definite descriptions, Mitkov's heuristic algorithm for resolving pronouns Mitkov (1998), and Bontcheva *et al.*'s heuristic algorithm for resolving proper names.

**BART.** BART (Versley et al. 2008)[47] is a generic toolkit that comes with the approach of Soon et al. (2001) as a default resolver, but makes it easy for a determined user to implement new features and/or resolution algorithms within its framework.

**OpenNLP.** The OpenNLP toolkit[48] comes with an implementation of the resolver of Morton (Morton 2000), using a similar to the one by (Ratnaparkhi 1999) for syntactic preprocessing.

**JavaRAP.** Finally, JavaRAP (Qiu et al. 2004)[49] is an implementation of the approach of Lappin and Leass (1994), and resolves only pronouns.

# References


Ahn, David, Valentin Jijkoun, Gilad Mishne, Karin Müller, Maarten de Rijke, and Stefan Schlobach. 2004. Using Wikipedia at the TREC QA track. In *Proceedings of the Thirteenth Text REtrieval Conference, Gaithersburg, Md., 16–19 November*.

Ahn, Kisuh, Johan Bos, James R. Curran, Dave Kor, Malvina Nissim, and Bonnie Webber. 2005. Question answering with QED at TREC-2005. In *Proceedings of the Fourteenth Text REtrieval Conference, Gaithersburg, Md., 15–18 November*.

Almor, A. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review* 106:748–765.

Alshawi, H. 1987. *Memory and Context for Language Interpretation*. Cambridge: Cambridge University Press.

Alshawi, H. 1990. Resolving quasi-logical forms. *Computational Linguistics* 16(3):133–144.

Alshawi, H., ed. 1992. *The Core Language Engine*. The MIT Press.

Anderson, A., S. Garrod, and A. Sanford. 1983. The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology* 35:427–440.


[46] `http://dces.essex.ac.uk/research/nle/GuiTAR`
[47] `http://www.sfs.uni-tuebingen.de/~versley/BART`
[48] `http://www.opennlp.org`
[49] `http://www.comp.nus.edu.sg/~qiul/NLPTools/JavaRAP.html`

Aone, Chinatsu and Scott W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Mass., 26–30 June*, pages 122–129.

Ariel, M. 1990. *Accessing Noun-Phrase Antecedents*. Croom Helm Linguistics Series. Routledge.

Arnold, J. E., J. G. Eisenband, S. Brown-Schmidt, and J. C. Trueswell. 2000. The immediate use of gender information: eyetracking evidence of the time-course of pronoun resolution. *Cognition* 76:B13–B26.

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

Asher, N. and A. Lascarides. 1998. Bridging. *Journal of Semantics* 15(1):83–13.

Baldwin, Breck. 1997. Cogniac: high precision coreference with limited knowledge and linguistic resources. In *ACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Baldwin, Breck, Jeff Reynar, Michael Collins, Jason Eisner, Adwait Ratnaparkhi, Joseph Rosenzweig, and Anoop Sarkar. 1995. University of Pennsylvania: Description of the University of Pennsylvania system used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

Barker, C. 1991. *Possessive Descriptions*. Ph.D. thesis, University of California at Santa Cruz, Santa Cruz, CA.

Barwise, J. and J. Perry. 1983. *Situations and Attitudes*. The MIT Press.

Bean, David and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Md., 20–26 June*.

Bean, D. and E. Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proc. of NAACL*.

Bell, E.T. 1934. Exponential numbers. *Amer. Math. Monthly* pages 411–419.

Bengtson, Eric and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25-27 October , pages 294–303.

Berger, A., S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–72.

Bergsma, S. 2005. Automatic acquisition of gender information for anaphora resolution. In *Proc. of 18th Conf. of the Canadian Society for Computational Studies of Intelligence*, pages 342–353. Victoria, B.C., Canada.

Berland, Matthew and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Md., 20–26 June*, pages 57–64.

Beun, R. and A. Cremers. 1998. Object reference in a shared domain of conversation. *Pragmatics and Cognition* 6(1/2):121–152.

Bod, R., J. Hay, and D. Jannedy, eds. 2003. *Probabilistic Linguistics*. MIT Press.

Boguraev, B. 1979. *Automatic Resolution of Linguistic Ambiguities*. Ph.D. thesis, University of Cambridge, Cambridge, UK.

Boyd, A., W. Gegg-Harrison, and D. Byron. 2005. Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In *In Proceedings of the ACL Workshop on Feature Selection for Machine Learning in NLP*, pages 40–47. Ann Arbor.

Bransford, J.D., J. R. Barclay, and J. J. Franks. 1972. Sentence memory: a constructive vs. interpretive approach. *Cognitive Psychology* 3:193–209.

Brennan, S.E., M.W. Friedman, and C.J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford, Cal., 6–9 July*, pages 155–162.

Broadbent, D. E. 1973. *In Defence of Empirical Psychology*. Methuen.

Bunescu, Razvan. 2003. Associative anaphora resolution: A web-based approach. In *EACL 2003 Workshop on the Computational Treatment of Anaphora*.

Bunescu, Razvan and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April*, pages 9–16.

Byron, Donna K. 2001. The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics* 27(4):569–577.

Byron, Donna K. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn., 7–12 July*, pages 80–87.

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

Carlson, G. N. 1977. *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts, Amherst.

Carter, D. M. 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK: Ellis Horwood.

Chamberlain, J., M. Poesio, and U. Kruschwitz. 2008. Phrase detectives - a web-based collaborative annotation game. In *Proc. of I-Semantics*. Graz.

Charniak, E. 1972. *Towards a Model of Children's Story Comprehension*. Ph.D. thesis, MIT. Available as MIT AI Lab TR-266.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence, Providence, R.I., 27–31 July*.

Cheng, Hua. 2001. *Modelling Aggregation Motivated Interactions in Descriptive Text Generation*. Ph.D. thesis, Division of Informatics, the University of Edinburgh, Edinburgh.

Chinchor, Nancy A. 1998. Overview of MUC-7/MET-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Barriers*. Cambridge, MA: MIT Press.

Cimiano, Philipp, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit vs. latent concept models for cross-language information retrieval. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence, Pasadena, Cal., 14–17 July*, pages 1513–1518.

Clark, H. H. 1977. Bridging. In P. N. Johnson-Laird and P. Wason, eds., *Thinking: Readings in Cognitive Science*, pages 411–420. London and New York: Cambridge University Press.

Clark, H. H. and C. R. Marshall. 1981. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, eds., *Elements of Discourse Understanding*. New York: Cambridge University Press.

Clear, J.H. 1993. The British national corpus. In P. Delany and G. P. Landow, eds., *The Digital Word: text-based computing in the humanities*, pages 163–187. Cambridge, Mass.: MIT Press.

Clifton, C. Jr. and F. Ferreira. 1987. Discourse structure and anaphora: some experimental results. In M. Coltheart, ed., *Attention and Performance XII: the psychology of reading*, pages 635–654. Hove, UK: Lawrence Erlbaum.

Cohen, P. R. 1984. The pragmatics of referring and the modality of communication. *Computational Linguistics* 10(2):97–146.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July*, pages 16–23.

Cooper, R. 1996. The role of situations in generalized quantifiers. In S. Lappin, ed., *Handbook of Contemporary Semantic Theory*, chap. 3, pages 65–86. Blackwell.

Cornish, F. 1986. Anaphoric pronouns: under linguistic control or signalling particular discourse representations? *Journal of Semantics* 5(3):233–260.

Crawley, R. J., R. A. Stevenson, and D. Kleinman. 1990. The use of heuristic strategies in the comprehension of pronouns. *Journal of Psycholinguistic Research* 19:245–264.

Csomai, A. and R. Mihalcea. 2008a. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems* Special issue on Natural Language Processing for the Web.

Csomai, Andras and Rada Mihalcea. 2008b. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems* 23(5):34–41.

Cucerzan, Silviu. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning, Prague, Czech Republic, 28–30 June*, pages 708–716.

Culotta, Aron, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April, pages 81–88.

Dagan, Ido and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics, Helsinki, Finland, 20–25 August*, vol. 3, pages 330–332.

Dagan, Ido, John Justeson, Shalom Lappin, Herbert Leass, and Ammon Ribak. 1995. Syntax and lexical statistics in anaphora resolution. *Applied Artificial Intelligence* 9(6):633–644.

Dale, R. 1992. *Generating Referring Expressions*. Cambridge, MA: The MIT Press.

Dalrymple, M., S. M. Shieber, and F. C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy* 14(4):399–452.

Daumé III, Hal and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., Canada, 6–8 October*, pages 97–104.

Denis, Pascal and Jason Baldridge. 2007a. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April, pages 236–243.

Denis, Pascal and Jason Baldridge. 2007b. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January, pages 1588–1593.

Dowty, D. R. 1986. The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and Philosophy* 9(1).

Dwivedi, V. D., N. A. Phillips, M. Laguë-Beauvais, and S. R. Baum. 2006. An electrophysiological study of mood, modal context, and anaphora. *Brain Research* 1117:135–153.

Eckert, M. and M. Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics* .

Ehrlich, K. and K. Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior* 22:75–87.

Evans, R. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing* 16(1):45–57.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Fisher, David, Stephen Soderland, Joseph McCarthy, Fangfang Feng, and Wendy Lehnert. 1996. Description of the UMass system as used for MUC-6. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*.

Fleischman, M., E. Hovy, and A. Echihabi. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July, pages 1–7.

Fraurud, K. 1990. Definiteness and the processing of NPs in natural discourse. *Journal of Semantics* 7:395–433.

Gabrilovich, Evgeniy and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July, pages 1301–1306.

Gabrilovich, Evgeniy and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January, pages 1606–1611.

Gardent, C. and K. Konrad. 2000. Interpreting definites using model generation. *Journal of Language and Computation* 1(2):193–209.

Gardent, Claire and Hélène Manuélian. 2005. Création d'un corpus annoté pour le traitement des déscriptions d éfinies. *Traitement Automatique des Langues* 46(1):115–140.

Garera, Nikesh and David Yarowsky. 2006. Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *Proceedings of the 10th Conference on Computational Natural Language Learning,* New York, N.Y., USA, 8–9 June, pages 37–44.

Garnham, A. 1982. On-line construction of representations of the content of texts. Reproduced by Indiana University Linguistics Club.

Garnham, A. 2001. *Mental models and the interpretation of anaphora*. Psychology Press.

Garnham, A., J. V. Oakhill, M. F. Ehrlich, and M. Carreiras. 1995. Representation and process in the interpretation of pronouns. *Journal of Memory and Language* 34:41–62.

Garrod, S. C. 1993. Resolving pronouns and other anaphoric devices: The case for diversity in discourse processing. In C. Clifton, L. Frazier, and K. Rayner, eds., *Perspectives in Sentence Processing*. Lawrence Erlbaum.

Garvey, C. and A. Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry* 5:459–464.

Gasperin, Caroline, Pablo Gamallo, Alexandre Agustini, Gabriel Lopes, and Vera Strube De Lima. 2001. Using syntactic contexts for measuring word similarity. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*.

Gasperin, Caroline and Renata Vieira. 2004. Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Ge, Niyu, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora,* Montréal, Canada, pages 161–170.

Gernsbacher, M. A. and D. Hargreaves. 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* 27:699–717.

Geurts, B. 1997. Good news about the description theory of names. *Journal of Semantics* 14(4):319–348.

Giles, Jim. 2005. Internet encyclopedias go head to head. *Nature* 438:900–901.

Givon, T., ed. 1983. *Topic continuity in discourse : a quantitative cross-language study*. Amsterdam and Philadelphia: J. Benjamins.

Givon, T. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics* 30.

Gordon, P. C., B. J. Grosz, and L. A. Gillion. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science* 17:311–348.

Gordon, P. C. and R. Hendrick. 1997. Intuitive knowledge of linguistic coreference. *Cognition* 62:325–370.

Gordon, P. C., R. Hendrick, K. Ledoux, and C. L. Yang. 1999. Processing of reference and the structure of language: an analysis of complex noun phrases. *Language and Cognitive Processes* 14(4):353–379.

Gordon, P. C. and K. A. Scearce. 1995. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory and Cognition* 23:313–323.

Grishman, Ralph and Beth Sundheim. 1995. Design of the MUC-6 evalutation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

Groenendijk, J.A.G. and M.J.B. Stokhof. 1991. Dynamic Predicate Logic. *Linguistics and Philosophy* 14:39–100.

Grosz, B.J., A.K. Joshi, and S. Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, Cambridge, Mass., 15–17 June*, pages 44–50. Cambridge, MA.

Grosz, B. J. 1977. *The Representation and Use of Focus in Dialogue Understanding*. Ph.D. thesis, Stanford University.

Grosz, B. J., A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):202–225.

Grosz, B. J. and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Gundel, J. K. 1974. *The Role of Topic and Comment in Linguistic Theory*. Ph.D. thesis, University of Texas at Austin. Reprinted by Garland Publishing, New York and London, 1988.

Gundel, J. K. 1998. Centering theory and the givenness hierarchy: Towards a synthesis. In M. A. Walker, A. K. Joshi, and E. F. Prince, eds., *Centering Theory in Discourse*, chap. 10, pages 183–198. Oxford University Press.

Gundel, J. K., N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2):274–307.

Gundel, J. K., N. Hedberg, and R. Zacharski. 2002. Pronouns without explicit antecedents: How do we know when a pronoun is referential? In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*.

Haghighi, Aria and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June, pages 848–855.

Haghighi, Aria and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistics,* Los Angeles, Cal., 1–6 June, pages 385–393.

Halliday, M. A. K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Harabagiu, S. and D. Moldovan. 1998. Knowledge processing on extended WordNet. In C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, pages 379–405. MIT Press.

Harabagiu, Sanda M., Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, Penn., 2–7 June*, pages 55–62.

Hardt, D. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 23(4):525–541.

Hartrumpf, Sven. 2001. Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the 3rd Conference on Computational Natural Language Learning,* Toulouse, France, 6–7 July 2001, pages 137–144.

Hartrumpf, Sven, Herrmann Helbig, and Rainer Osswald. 2003. The semantically based corpus HaGenLex - structure and technological environment. *Traitement automatique des langues* 44(2):81–105.

Hasler, Laura, Constantin Orasan, and Karin Naumann. 2006. Nps for events: Experiments in coreference annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May*.

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

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics, Nantes, France, 23-28 August*, pages 539–545.

Heeman, P. A. and J. F. Allen. 1995. The TRAINS-93 dialogues. TRAINS Technical Note TN 94-2, University of Rochester, Dept. of Computer Science, Rochester, N.Y.

Heim, I. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

Heim, I. 1983. File change semantics and the familiarity theory of definiteness. In R. Bauerle, C. Schwarze, and A. von Stechow, eds., *Meaning, Use and Interpretation of Language*. Berlin: de Gruyter.

Helbig, Hermann and Sven Hartrumpf. 1997. Word class functions for syntactic-semantic analysis. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*.

Hendrickx, Iris, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for dutch. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May – 1 June*.

Hinrichs, Erhard, Katja Filippova, and Holger Wunsch. 2005a. What treebanks can do for you: Rule-based and machine-learning approaches to anaphora resolution in German. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05)*.

Hinrichs, Erhard, Sandra Kübler, and Karin Naumann. 2005b. 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.

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

Hirschman, Lynette, Patricia Robinson, John Burger, and Marc Vilain. 1997. Automating coreference: The role of automated training data. In *Proc. of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Hirst, G. 1981. *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science 119. Berlin: Springer-Verlag.

Hitzeman, J. and M. Poesio. 1998. Long-distance pronominalisation and global focus. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August*, pages 550–556.

Hobbs, J. R. 1978. Resolving pronoun references. *Lingua* 44:311–338.

Hobbs, J. R. 1979. Coherence and coreference. *Cognitive Science* 3:67–90.

Hobbs, J. R. and A. Kehler. 1997. A theory of parallelism and the case of vp ellipsis. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July*, pages 394–401.

Hobbs, J. R. and S. M. Shieber. 1987. An algorithm for generating quantifier scopings. *Computational Linguistics* 13(1-2):47–63.

Hobbs, J. R., M. Stickel, P. Martin, and D. Edwards. 1993. Interpretation as abduction. *Artificial Intelligence Journal* 63:69–142.

Hoste, V. 2005. *Optimization Issues in Machine Learning of Coreference*. Ph.D. thesis, University of Antwerp.

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* New York, N.Y., 4–9 June.

Hudson-D'Zmura, S. and M. K. Tanenhaus. 1998. Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In M. A. Walker, A. K. Joshi, and E. F. Prince, eds., *Centering in Discourse*, pages 199–226. Oxford University Press.

Humphreys, K., R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. 1998. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Proceedings of MUC-7*.

Iida, Ryu, Kentaro Inui, and Yuji Matsumoto. 2007a. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing (TALIP)* 6(4):1–22.

Iida, R., K. Inui, H. Takamura, and Y. Matsumoto. 2003a. Incorporating contextual cues in trainable models for coreference resolution. In *Proc. EACL Workshop on the Computational Treatment of Anaphora*.

Iida, Ryu, Kentaro Inui, Hiroya Takamura, and Yuji Matsumoto. 2003b. Incorporating contextual cues in trainable models for coreference resolution. In R. Dale, K. van Dempter, and R. Mitkov, eds., *Proceedings of the EACL-03 Workshop on the Computational Treatment of Anaphora*. Budapest.

Iida, Ryu, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007b. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the ACL-07 Linguistic Annotation Workshop*.

Jakob, Niklas and Iryna Gurevych. 2010. Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July, pages 263–268.

Ji, Heng, David Westbrook, and Ralph Grishman. 2005. Using semantic relations to refine coreference decisions. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., Canada, 6–8 October*, pages 17–24.

Jurafsky, D. and J. H. Martin. 2009. *Speech and Language Processing*. Prentice Hall, 2nd edn.

Kabadjov, M. A. 2007. *Task-oriented evaluation of anaphora resolution*. Ph.D. thesis, University of Essex, Dept. of Computing and Electronic Systems, Colchester, UK.

Kameyama, M. 1985. *Zero Anaphora: The case of Japanese*. Ph.D. thesis, Stanford University, Stanford, CA.

Kameyama, Megumi. 1997. Recognizing referential links: an information extraction prespective. In *ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Kamp, H. 1979. Events, instant and temporal reference. In R. Bauerle, U. Egli, and A. von Stechow, eds., *Semantics from Different Points of View*, pages 376–417. Springer-Verlag.

Kamp, H. 1981. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, eds., *Formal Methods in the Study of Language*. Amsterdam: Mathematical Centre.

Kamp, H. and U. Reyle. 1993. *From Discourse to Logic*. Dordrecht: D. Reidel.

Kaplan, D. 1977. Demonstratives. an essay on the semantics, logic, metaphysics and epistemology of demonstratives and other indexicals. Unpublished manuscript, University of California, Los Angeles.

Karamanis, N. 2003. *Entity coherence for descriptive text structuring*. Ph.D. thesis, University of Edinburgh, Informatics.

Karamanis, N., M. Poesio, J. Oberlander, and C. Mellish. 2009. Evaluating centering for information ordering using corpora. *Computational Linguistics* 35(1):xx–yy.

Karttunen, L. 1976. Discourse referents. In J. McCawley, ed., *Syntax and Semantics 7 - Notes from the Linguistic Underground*, pages 363–385. New York: Academic Press.

Kehler, Andrew. 1997. Probabilistic coreference in information extraction. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, Providence, R.I., 1–2 August*, pages 163–173.

Kehler, Andrew, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Boston, Mass., 2–7 May*, pages 289–296.

Kehler, Andrew, Laura Kertz, Hannah Rohde, and Jeffrey Elman. 2008. Coherence and coreference revisited. *Journal of Semantics* 25(1):1–44.

Kelleher, J.D., F. Costello, and J. van Genabith. 2005. Dynamically updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence* 167:62–102.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August*, vol. 1, pages 113–118.

Kibble, R. and R. Power. 2000. An integrated framework for text planning and pronominalization. In *Proc. of the International Conference on Natural Language Generation (INLG)*. Mitzpe Ramon, Israel.

Klapholz, D. and A. Lockman. 1975. Contextual reference resolution. *American Journal of Computational Linguistics* .

Klavans, J. and P. Resnik, eds. 1996. *The Balancing Act*. MIT Press.

Klenner, Manfred and Étienne Ailloud. 2008. Enhancing coreference clustering. In *Second Bergen Workshop on Anaphora Resolution (WAR II)*.

Knott, A., J. Oberlander, M. O'Donnell, and C. Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, eds., *Text representation: linguistic and psycholinguistic aspects*, pages 181–196. Amsterdam and Philadelphia: John Benjamins.

Kouchnir, Beata. 2004. A machine learning approach to German pronoun resolution. In *Proceedings of the ACL'04 Student Research Workshop*.

Kripke, S. A. 1972. Naming and necessity. In D. Davidson and G. Harman, eds., *Semantics of Natural Language*, pages 253–355. Dordrecht: Reidel.

Landragin, F., A. De Angeli, F. Wolff, P. Lopez, and L. Romary. 2002. Relevance and perceptual constraints in multimodal referring actions. In K. van Deemter and R. Kibble, eds., *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 395–413. CSLI.

Langacker, R. 1969. Pronominalization and the chain of command. In D. Reibel and S. Schane, eds., *Modern Studies in English*. Englewood Cliffs, N.J.: Prentice-Hall.

Lappin, S. and H. J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–562.

Lasnik, H. 1976. Remarks on coreference. *Linguistic Inquiry* 2(1).

Leacock, Claudia and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, ed., *WordNet. An Electronic Lexical Database*, chap. 11, pages 265–283. Cambridge, Mass.: MIT Press.

Lehnert, Wendy, Claire Cardie, Joseph McCarthy, Ellen Riloff, and Steven Soderland. 1992. University of massachusetts: Description of the CIRCUS system as used for MUC-4. In *Proceedings of the fourth Message Understanding Conference (MUC-4)*.

Lin, Dekang. 1995. University of Manitoba: Description of the PIE system used for MUC-6. In *Proceedings of the 6th Message Understanding Conference*.

Lin, Dekang. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August*, pages 768–774.

Lin, Dekang. 1998b. Using collocation statistics in information extraction. In *Proceedings of MUC-7*.

Linde, C. 1979. Focus of attention and the choice of pronouns in discourse. In T. Givon, ed., *Syntax and Semantics 12*. Academic Press.

Lo, Ka Kan and Wai Lam. 2006. Using semantic relations with world knowledge for question answering. In *Proceedings of the Fifteenth Text REtrieval Conference, Gaithersburg, Md., 14–17 November*.

Lockman, A. and A. D. Klappholz. 1980. Toward a procedural model of contextual reference resolution. *Discourse Processes* 3.

Loebner, S. 1987. Definites. *Journal of Semantics* 4:279–326.

Lund, Kevin, R. A. Atchley, and Curt Burgess. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., Canada, 6–8 October*, pages 25–32.

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July, pages 136–143.

LuperFoy, Susann. 1992. The representation of multimodal user interface dialogues using discourse pegs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, Newark, Del., 28 June – 2 July*, pages 22–31.

MacDonald, M. C., N. J. Pearlmutter, and M. S. Seidenberg. 1994. Lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4):676–703.

Magnini, B., E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, and R. Sprugnoli. 2006. I-CAB: the italian content annotation bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May*.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2):313–330.

Markert, Katja and Malvina Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics* 31(3):367–402.

Matthews, A. and M. S. Chodorow. 1988. Pronoun resolution in two-clause sentences: effects of ambiguity, antecedent location, and depth of embedding. *Journal of Memory and Language* 27:245–260.

May, R. 1985. *Logical Form in Natural Language*. The MIT Press.

McCarthy, Joseph F. 1996. *A Trainable Approach to Coreference Resolution for Information Extraction*. Ph.D. thesis, University of Massachusetts.

McCarthy, Joseph F. and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montréal, Canada, 20–25 August*, pages 1050–1055.

Mihalcea, Rada. 2007. Using Wikipedia for automatic Word Sense Disambiguation. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April, pages 196–203.

Miller, George A. and Florentina Hristea. 2006. WordNet nouns: Classes and instances. *Computational Linguistics* 32(1):1–3.

Milne, David and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July*, pages 25–30.

Milne, David and Ian H. Witten. 2008b. Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management, Napa Valley, Cal., 26–30 October*, pages 1046–1055.

Miltsakaki, E. 2002. Towards an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics* 28(3):319–355.

Minsky, Marvin. 1975. A framework for representing knowledge. In P. H. Winston, ed., *The Psychology of Computer Vision*, pages 211–277. New York: McGraw-Hill.

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August*, pages 869–875.

Mitkov, Ruslan. 2000. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Third Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC3)*.

Mitkov, R. 2002. *Anaphora Resolution*. Longman.

Morton, T. 2000. Coreference for NLP applications. In *Proc. of the 38th ACL*, pages 173–180. Hong Kong.

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

Muskens, R.A. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy* 19:143–186.

Nastase, Vivi. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Waikiki, Honolulu, Hawaii, 25-27 October , pages 763–772.

Ng, Vincent. 2005. Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June, pages 157–164.

Ng, Vincent. 2007. Shallow semantics for coreference resolution. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India, 6–12 January, pages 1689–1694.

Ng, V. 2008. Unsupervised models for coreference resolution. In *Proc. of EMNLP*.

Ng, Vincent and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 24 August – 1 September*.

Ng, Vincent and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn., 7–12 July*, pages 104–111.

Nicol, J. and D. A. Swinney. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research* 18:5–19. Special Issue on Sentence Processing.

NIST. 2002. The ACE 2002 evaluation plan. `ftp://jaguar.ncsl.nist.gov/ace/doc/ACE-EvalPlan-2002-v06.pdf`.

Orasan, C. and R. Evans. 2007. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research* 29:79–103.

Partee, B. H. 1972. Opacity, coreference, and pronouns. In D. Davidson and G. Harman, eds., *Semantics for Natural Language*, pages 415–441. Dordrecht, Holland: D. Reidel.

Partee, B. H. 1973. Some structural analogies between tenses and pronouns in English. *Journal of Philosophy* 70:601–609.

Partee, B. H. 1995. Quantificational structures and compositionality. In E. Bach, E. Jelinek, A. Kratzer, and B. H. Partee, eds., *Quantification in Natural Languages*. Kluwer.

Passonneau, R. J. 1993. Getting and keeping the center of attention. In M. Bates and R. M. Weischedel, eds., *Challenges in Natural Language Processing*, chap. 7, pages 179–227. Cambridge University Press.

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

Poesio, M. 1993. A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, eds., *Situation Theory and its Applications, vol.3*, chap. 12, pages 339–374. Stanford: CSLI.

Poesio, M. 1994a. *Discourse Interpretation and the Scope of Operators*. Ph.D. thesis, University of Rochester, Department of Computer Science, Rochester, N.Y.

Poesio, M. 1994b. Weak definites. In M. Harvey and L. Santelmann, eds., *Proceedings of the Fourth Conference on Semantics and Linguistic Theory, SALT-4*, pages 282–299. Cornell University Press.

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

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

Poesio, Massimo, Mijail Alexandrov-Kabadjov, Renata Vieira, Rodrigo Goulart, and Olga Uryupina. 2005. Does discourse-new detection help definite description resolution? In *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*.

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

Poesio, M. and R. Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 26 May – 1 June*.

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

Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, 29–31 May*, pages 1220–1225.

Poesio, M. and M. A. Kabadjov. 2004. A general-purpose, off the shelf anaphoric resolver. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May*, pages 653–656.

Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004b. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July, pages 143–150.

Poesio, M. and N. N. Modjeska. 2005. Focus, activation, and this-noun phrases: An empirical study. In A. Branco, R. McEnery, and R. Mitkov, eds., *Anaphora Processing*, pages 429–442. John Benjamins.

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

Poesio, Massimo, Sabine Schulte im Walde, and Chris Brew. 1998. Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*.

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

Poesio, Massimo, Olga Uryupina, Renata Vieira, Mijail Alexandrov-Kabadjov, and Rodrigo Goulart. 2004d. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of the ACL Workshop on Reference Resolution*.

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

Poesio, M., R. Vieira, and S. Teufel. 1997. Resolving bridging references in unrestricted text. In R. Mitkov, ed., *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6. Madrid. Also available as HCRC Research Paper HCRC/RP-87, University of Edinburgh.

Pollard, C. and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, CA: University of Chicago Press.

Ponzetto, Simone Paolo. 2010. *Knowledge Acquisition from a Collaboratively Generated Encyclopedia*, vol. 327 of *Dissertations in Artificial Intelligence*. Amsterdam, The Netherlands: IOS Press & Heidelberg, Germany: AKA Verlag.

Ponzetto, Simone Paolo and Roberto Navigli. 2010. Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July, pages 1522–1531.

Ponzetto, Simone Paolo and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* New York, N.Y., 4–9 June, pages 192–199.

Ponzetto, Simone Paolo and Michael Strube. 2007a. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence,* Vancouver, B.C., Canada, 22–26 July, pages 1440–1445.

Ponzetto, Simone Paolo and Michael Strube. 2007b. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research* 30:181–212.

Pradhan, Sameer, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*.

Prince, E. F. 1981. Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*, pages 223–256. New York: Academic Press.

Prince, E. F. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, eds., *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

Qiu, Long, Min-Yen Kan, and Tat-Seng Chua. 2004. A public reference implementation of the RAP anaphora resolution algorithm. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May*.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. San Mateo, Cal.: Morgan Kaufman.

Quinlan, Ross. 1986. Induction of decision trees. *Machine Learning* 1(1):81–106.

Rada, Roy, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics* 19(1):17–30.

Rahman, A. and V. Ng. 2009. Supervised models for coreference resolution. In *Proc. of EMNLP*.

Ratnaparkhi, Adwait. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning* 34:151–178.

Ravichandran, Deepak, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mich., 25–30 June, pages 622–629.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Uppsala, Sweden: Association for Computational Linguistics.

Recasens, M. and M. A. Martí. 2009. Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation* .

Reichman, R. 1985. *Getting Computers to Talk Like You and Me*. Cambridge, MA: The MIT Press.

Reinhart, T. 1976. *The Syntactic Domain of Anaphora*. Ph.D. thesis, MIT, Cambridge, MA.

Reinhart, T. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1). Also distributed by Indiana University Linguistics Club.

Reinhart, T. and E. Reuland. 1993. Reflexivity. *Linguistic Inquiry* 24:657–720.

Roberts, C. 1989. Modal subordination and pronominal anaphora in discourse. *Linguistics and Philosophy* 12:683–721.

Roberts, C. 2003. Uniqueness presuppositions in english definite noun phrases. *Linguistics and Philosophy* 26(3):287–350.

Rodriguez, K. J., F. Delogu, Y. Versley, E. Stemle, and M. Poesio. 2010. Anaphoric annotation of Wikipedia and blogs in the LiveMemories corpus. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* Valletta, Malta, 19–21 May 2010.

Rooth, M. 1987. Noun Phrase Interpretation in Montague Grammar, File Change Semantics, and Situation Semantics. In P. Gärdenfors, ed., *Generalized Quantifiers*, pages 237–268. Dordrecht, The Netherlands: D. Reidel.

Roth, D. and W.-T. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proc. of CONLL.*

Runner, J. T., R. S. Sussman, and M. K. Tanenhaus. 2003. Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye movements. *Cognition* 81:1–13.

Sag, I. A. and J. Hankamer. 1984. Toward a theory of anaphoric processing. *Linguistics and Philosophy* 7:325–345.

Sanchez-Graillet, O., M. Poesio, M. Kabadjov, and R. Tesar. 2006. What kind of problems do protein interactions raise for anaphora resolution? - a preliminary analysis. In *Proc. of SMBM*. University of Jena, Jena.

Sanford, A. J. and S. C. Garrod. 1981. *Understanding Written Language*. Chichester: Wiley.

Sasano, Ryohei, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, U.K., 18–22 August*, pages 769–776.

Sauper, Christina and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July*, pages 208–216.

Schiehlen, Michael. 2004. Optimizing algorithms for pronoun resolution. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 390–396. Geneva, Switzerland.

Schütze, Hinrich and Jan O. Pedersen. 1995. Information retrieval based on word senses. In *Proceedings of the Fourth Symposium on Document Analysis and Information Retrieval,* Las Vegas, Nev., 24–25 April , pages 161–175.

Seco, Nuno, Tony Veale, and Jer Hayes. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 23–27 August*, pages 1089–1090.

Seki, Kazuhiro, Artsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 24 August – 1 September*.

Sheldon, A. 1974. The role of parallel function in the acquisition of relative clauses in english. *Journal of Verbal learning and Verbal behavior* 13(272–281).

Sidner, C. L. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Smyth, R. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research* 23:197–229.

Solomonoff, A., A. Mielke, M. Schmidt, and H. Gish. 1998. Clustering speakers by their voices. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 757–760.

Soon, Wee Meng, Hwee Tou Ng, and Chung Yong Lim. 1999. Corpus-based learning for noun phrase coreference resolution. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, pages 285–291.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4):521–544.

Stede, Manfred. 2004. The Potsdam Commentary Corpus. In *ACL'04 Workshop on Discourse Annotation*.

Steinberger, J., M. Poesio, M. Kabadjov, and K. Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management* 43:1663–1680. Special issue on Summarization.

Stevenson, R. J., R. A. Crawley, and D. Kleinman. 1994. Thematic roles, focus, and the representation of events. *Language and Cognitive Processes* 9:519–548.

Stevenson, R. J., A. W. R. Nelson, and K. Stenning. 1995. The role of parallelism in strategies of pronoun comprehension. *Language and Cognitive Processes* 38:393–418.

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, 2–7 July*, pages 656–664.

Strube, M. 1998. Never look back: An alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada, 10–14 August*, pages 1251–1257.

Strube, M. and U. Hahn. 1999. Functional centering–grounding referential coherence in information structure. *Computational Linguistics* 25(3):309–344.

Strube, Michael and Simone Paolo Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July, pages 1419–1424.

Strube, Michael, Stefan Rapp, and Christoph Müller. 2002. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, Penn., 6–7 July*, pages 312–319.

Stuckardt, Roland. 2004. Three algorithms for competence-oriented anaphor resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2004)*.

Suri, L. Z. and K. F. McCoy. 1994. RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics* 20(2):301–317.

Tetreault, J. R. 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics* 27(4):507–520.

Trouilleux, François, Éric Gaussier, Gabriel G. Bies, and Annie Zaenen. 2000. Coreference resolution evaluation based on descriptive specificity. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece, May*.

Uryupina, Olga. 2003. High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*.

Uryupina, Olga. 2006. Coreference resolution with and without linguistic knowledge. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 22–28 May*.

Vallduvi, E. 1993. Information packaging: a survey. Research Paper RP-44, University of Edinburgh, HCRC.

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

van Rijsbergen, C. J. Keith. 1979. *Information Retrieval*. Butterworths.

Versley, Yannick. 2006. A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.

Versley, Yannick. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation* 6(3–4):333–353.

Versley, Yannick, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Companion Volume to the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, 15–20 June, pages 9–12. Demo session paper.

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

Vieira, Renata and Massimo Poesio. 1997. Processing definite descriptions in corpora. In S. Botley and M. McEnery, eds., *Corpus-based and Computational Approaches to Discourse Anaphora*. UCL Press.

Vieira, Renata and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics* 26(4):539–593.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann.

von Ahn, Luis. 2006. Games with a purpose. *Computer* 39(6):92–94.

Wagner, Andreas and Bettina Zeisler. 2004. A syntactically annotated corpus of Tibetan. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May*.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium.

Walker, M. A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, B.C., Canada, 26–29 June*, pages 251–261.

Walker, M. A. 1998. Centering, anaphora resolution, and discourse structure. In M. A. Walker, A. K. Joshi, and E. F. Prince, eds., *Centering in Discourse*, chap. 19, pages 401–435. Oxford University Press.

Walker, M. A., M. Iida, and S. Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2):193–232.

Walker, M. A., A. K. Joshi, and E. F. Prince, eds. 1998. *Centering Theory in Discourse*. Clarendon Press / Oxford.

Webber, B. L. 1979. *A Formal Approach to Discourse Anaphora*. New York: Garland.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2008. Ontonotes release 2.0. LDC2008T04, Philadelphia, Penn.: Linguistic Data Consortium.

Wilks, Y. A. 1975. An intelligent analyzer and understander of english. *Communications of the ACM* 18(5):264–274. Reprinted in *Readings in Natural Language Processing*, Morgan Kaufmann.

Winograd, Terry. 1972. *Understanding Natural Language*. Academic Press.

Woods, W. A., R. Kaplan, and B. Nash-Webber. 1972. The lunar sciences natural language information system: Final report. Report 2378, BBN, Cambridge, Mass.

Wu, Zhibiao and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, N.M., 27–30 June*, pages 133–138.

Yang, Xiaofeng and Jian Su. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic, 23–30 June, pages 528–535.

Yang, X., J. Su, J. Lang, C. L. Tan, T. Liu, and S. Li. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proc. of ACL*, pages 843–851. Columbus.

Yang, Xiaofeng, Jian Su, and Chew Lim Tan. 2005. A twin-candidate model of coreference resolution with non-anaphor identification capability. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing, Jeju Island, South Korea, 11–13 October*, pages 719–730.

Yang, Xiaofeng, Guodung Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July, pages 176–183.