Reference and Ellipsis in Ontological Semantics

Marjorie McShane and Sergei Nirenburg

MCCS-02-329

Computing Research Laboratory
Box 30001
New Mexico State University
Las Cruces, NM 88003-0001

The Computing Research Laboratory was established by the
New Mexico State Legislature
under the Science and Technology Commercialization Commission
as part of the Rio Grande Research Corridor.

1. Introduction

It is said that during Napoleon's march from Elba to Paris at the beginning of his "One Hundred Days" in 1815, the successive headlines in Paris newspapers ran something like:

> **The Corsican Monster** Lands at Toulon
> **The Usurper** Marches North
> **Bonaparte** Reaches Lyons
> **Ex-Emperor** in Fontainebleu
> Paris Welcomes **His Imperial Majesty**

It might be difficult to attempt to build a computer system that would emulate the nimbleness, inventiveness and political adroitness of the Parisian editors. However, it is within our reach to build a system that would determine that the text elements in boldface all refer to the same person, and that this person is Emperor Napoleon I Bonaparte.

In this report we describe the conceptual framework for our current work on reference and ellipsis within an ontological-semantic approach to natural language processing (NLP). We show the literature to be dominated by fractured approaches to what is really a singular, albeit complex, problem whose sufficient treatment requires knowledge-rich processing methods rather than stop gaps. Having recast subtypes of reference and ellipsis as variations on a theme, we describe the ontology-based environment that will support their treatment. Then we present a brief sketch of our preliminary plan of work. This research and development effort is certainly not the last word on this complex clutch of problems; rather, it is an ambitious first step in a new direction.

2. The Goal as We See It

We understand processing reference in NLP as finding all referring expressions in a text or a corpus and associating them with the representation of real-world entities or events. This definition implies that **co**reference—that is, the search for surface antecedents—is only a means to a more fundamental end. We are pursuing a novel approach to the treatment of reference in NLP that extends the current state of the art in both breadth and depth. It covers a broader array of phenomena and uses deeper – but available or attainable – sources of knowledge to power its heuristic algorithms than any extant approach. Our algorithms for treatment of reference will be incorporated into an existing natural language analysis system (Mahesh *et al.* 1997, Beale *et al.* 2002) that includes semantic analysis and produces meaning representations of input texts with the help of a formal ontology.

Our procedure for reference treatment addresses all the types of referring expressions and consists of two components, **detection** and **resolution**. Detection consists of the following three tasks: a) determining which objects and events have referential function (not all do, as in *My son is **a doctor***); b) categorizing the referential ones, of which there are many subclasses (as shown in Figure 1); and c) detecting elliptical references. Resolution then finds conceptual references for the expressions, possibly using textual antecedents as clues.

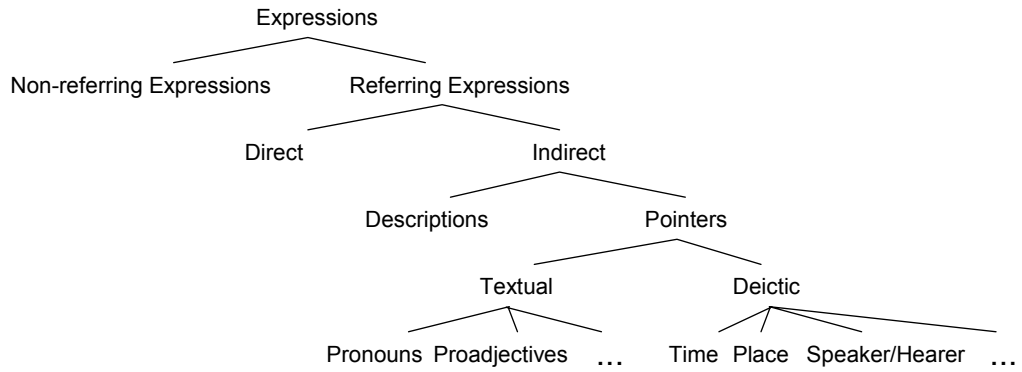

Figure 1. Types of expressions.

We will illustrate the **types** of referring expressions with examples drawn from the following text, taken at random from the CNN website (it is a typical text and it demonstrates how important it is to be able to treat reference adequately). In the text itself, for purposes of illustration, we marked just two of the many coreference chains in the text (the one referring to the Afghan foreign minister, in bold; the one referring to the Afghan people, in italics).

> WASHINGTON (CNN) -- **Afghanistan's interim foreign minister** expressed optimism Saturday that **his** nation can rebuild after more than two decades of conflict, provided that the international community remains committed to supplying support. "What *we* need is continued engagement from the United States, first of all, in the war against terror, which will help stability in Afghanistan and the whole region ... and also in the reconstruction efforts of *our* people," **Abdullah Abdullah** told CNN. "It is a major challenge. *We* are aware of it." "What is going on in the political process is a transition from war to peace. After 22 years of war, *we* have won the war, virtually, and *we* have to win the peace," **Abdullah** said. "It is rebuilding the state from scratch in all aspects of it – political, economical, from the infrastructure point of view, cultural, social. It is an enormous task. But I'm sure *the Afghans* will do it with the support of the international community," **he** said. **Abdullah** is in Washington to prepare for a visit by interim Afghanistan chairman Hamid Karzai, who is scheduled to meet with President Bush Monday, his first official meeting with Bush since assuming control after the fall of the Taliban regime. On Friday, **Abdullah** met with Secretary of State Colin Powell and National Security Advisor Condoleezza Rice. Powell, who visited Kabul, the Afghan capital, this month, vowed that the United States would stand by *the Afghan people*. **Abdullah** also gave the Council on Foreign Relations an outline of Afghanistan's reconstruction plan to rebuild the devastated country. **He** told the group that the interim administration is developing a constitution for Afghanistan and will make substantial efforts to include women and *the nation's* various ethnic groups in the government. Members of the commission that will organize the tribal council or Loya Jirga, whose task is to choose a transitional government at mid-year, were announced Friday. Women are included among the commission's members. "The opportunity is there," **Abdullah** said Saturday. "*We* were optimistic even before September 11 when there were no opportunities and *we* were trying hard, struggling hard, to create that opportunity," **he** said. "*We*, as *Afghans*, have to seize it, and have to seize it quickly, and *our* friends should support us. Together *we* can make it."

Direct reference is referring to an object or an event by its basic name. For people, this will typically be their full name (*Abdullah Abdullah, the Afghans*), their full name expanded by a description (*interim Afghanistan chairman Hamid Karzai, Secretary of State Colin Powell*) or a canonical abbreviation (*President Bush, Bush, Powell, Abdullah*). For organizations and places, this will typically be their full name (*the United States, the Council on Foreign Relations, Loya Jirga*) or a known acronym (*CNN, Washington [for Washington, DC]*). For events, this will typically be their full name (*the war against terror*) or a known abbreviation (*September 11*). One can view these expressions as keys for the database records for their referents. These referring expressions can in some cases be ambiguous (e.g., if the database contains more than one *Abdullah Abdullah*).

All other referring expressions are **indirect**. They subdivide into descriptions and pointers. **Descriptions** denote their referents by mentioning some of their non-key properties. They can be definite (e.g., *Afghanistan's interim foreign minister, the international community, the Afghan capital*) or indefinite (*a transition from war to peace, continued engagement from the United States*). Unlike descriptions, **pointers** just contain enough information to allow hearers to reconstruct to which referring expressions they point. Pointers can be further subdivided into **textual pointers** (those that typically point to coreferents in the text itself, like *he*) and **deictic pointers** (pointing to objects in the "universe of discourse", that is, to some expected properties of facts – e.g., time, like *at mid-year*; space, like *here*; identity of the speaker and hearer, like *we*; etc.). People are adept at resolving references in well-constructed texts. Out task is to build a computer program that emulates that capability.

Detecting **ellipsis** involves locating syntactic gaps as well as semantically incomplete structures. In English, syntactic gaps include such things as elided verbs in gapping structures (*Mary likes politics but Bill ∅ only sports*), elided VPs (*Mary wants to watch CNN but Bill doesn't ∅*), and elided head nouns after modifiers (*Mary watched CNN for 40 minutes and Bill for only five ∅*). Semantically incomplete structures are found in phrases like *continued engagement from the United States,* where the full interpretation of the term *engagement* requires the addition of modifiers like *military* and *peace-keeping*.

The output of the detection step in reference treatment is, then, a list of all referring expressions, marked by their type, that were either overtly present in the text or were introduced in it through the detection of ellipsis.

Once all referring expressions are detected and classified, they must be **resolved**. We use the term 'resolve' in a broader sense than is typical in the literature: for us, resolution means that all referring expressions must ultimately be associated with representations of objects or events, not only put in a coreference relation with another text element. In our approach, the representations are stored in a fact database and the ontology (see below). Direct referring expressions are resolved through a direct link to the relevant database entry (or, if there is none such yet, to a newly created one), whereas indirect referring expressions require specialized processing by type. Definite and indefinite descriptions, e.g., *Afghanistan's interim foreign minister*, must be linked to the expression corresponding to a database key, e.g., *Abdullah Abdullah.* All pointers and ellipses must be expanded into their full referential form based upon the establishment of a coreference chain within the text (*he → Abdullah Abdullah*) or extra-textual information (*mid-year → the middle [perhaps May, June, July, August] of the year 2002*). Once expanded they, too, must be either linked to a database entry or initiate a new one.

3. Our Environment for the Treatment of Reference

To recapitulate, our model for processing reference treats referring expressions in text as referring to objects or events in the system's world model. More specifically, referring takes place either to *types* of objects or events (e.g., *capital cities*) or, more often, to their *instances* (*Washington, DC*). Processing reference means for us, basically, a) finding all referring expressions in a text; b) for each referring expression determining its **anchor**, the element of the world model to which it refers; and c) unifying the information in the referring expression with the description of its anchor. Note that in our approach the anchor is not a text element. It is, rather, an element of the knowledge stratum. This distinction underscores our accent on reference, not coreference.

Architecturally, our treatment of reference is being carried out within the framework of a general-purpose semantic analyzer developed at NMSU CRL and its associates. We will very briefly describe the basic semantic analyzer (a detailed description can be found in Nirenburg and Raskin 2002) and then present the work we have undertaken on reference.

Ontological-semantic processing for text analysis relies on the results of a battery of pre-semantic text processing modules (see Figure 2). The output of these modules provides input to and background knowledge for semantic analysis. The tokenizer module deals with any mark-up symbols in the input text, finds boundaries of sentences and words, and detects and recognizes dates, numbers, named entities and acronyms. Morphological analysis uses the results of tokenization.

A morphological analyzer accepts a string of word forms as input and for each word form outputs a record containing its citation form in the lexicon and a set of morphological features and their values that correspond to the word form from the text. Once the morphological analyzer has generated the citation forms for word forms in a text, the system can look them up in its lexicons, including the onomasticon (a lexicon of names), and thus activate the relevant lexical entries. The task of syntactic analysis in ontological semantics is, essentially, to determine clause-level dependency structures for an input text and assign syntactic valency values to clause constituents (that is, establish subjects, direct objects, obliques and adjuncts).

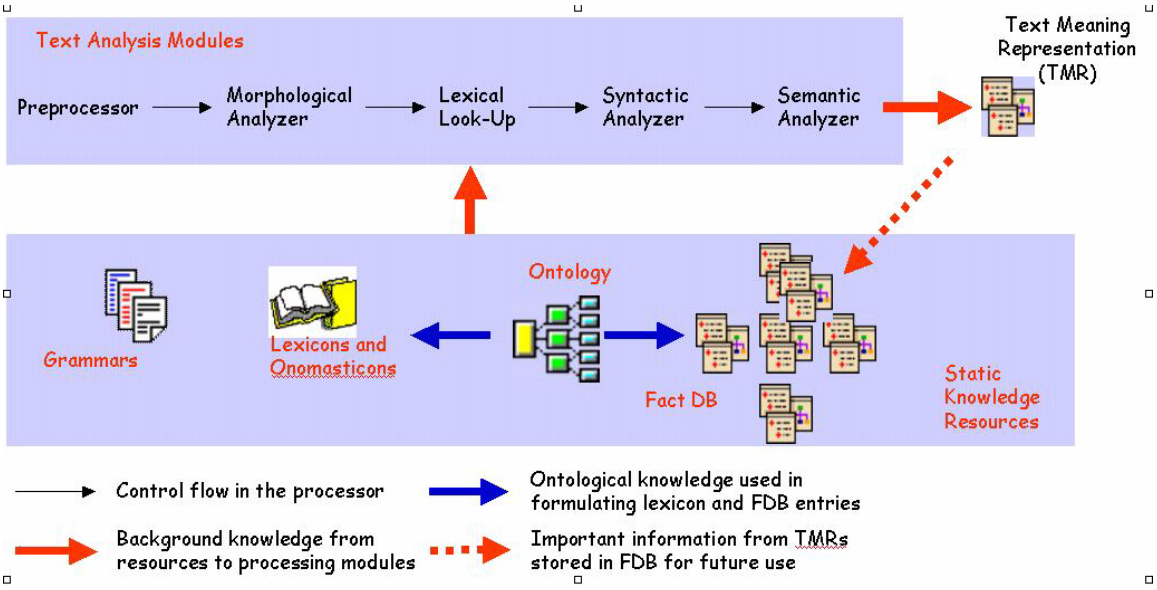

Figure 2. Ontological-semantic processing for text analysis.

Semantic analysis proper takes as input results from the earlier stages of processing and produces a **text meaning representation** (TMR). The central task for semantic analysis is to construct unambiguous propositional meaning by processing selectional restrictions, listed in the ontology and the semantic zones of lexicon entries. Other issues include treating such phenomena as aspect and modality, non-literal language (which, incidentally, is important for the treatment of reference as well) and building a discourse structure associated with the basic propositional structure of the text.

The major "static knowledge sources" for text analysis are: the **TMR language**, the **ontology**, the **fact database** (FDB) and a **lexicon** that includes an **onomasticon**. The ontology provides a metalanguage for describing the meaning of lexical units of a language as well as for the specification of meaning encoded in TMRs. The ontology contains specifications of concepts corresponding to classes of things and events in the world. Formatwise, the ontology is a collection of frames, or named collections of property-value pairs. For lack of space, we very briefly illustrate the ontology we intend to use in Figure 3. The ontology contains about 5,500 concepts, each of which has, on average, 16 properties defined for it. This ontology has been shown to be able to represent the meanings of over 35,000 entries in a Spanish lexicon. We also

have an English lexicon of about 16,000 entries and have developed an efficient methodology for the acquisition of the ontology and the lexicon (Nirenburg and Raskin 2002, Chapter 9). The Fact DB contains a list of remembered instances of ontological concepts. In other words, if the ontology has the concept CITY, the Fact DB may contain entries for London, Paris or Rome; if the ontology has the concept SPORTS-EVENT, the Fact DB may have an entry for the Salt Lake City Olympics (small caps are used to distinguish ontological concepts from English words). An entry in a Fact DB is illustrated in Figure 4.

The following BNF presents a very brief description of the version of the basic TMR that we are using.

TMR ::= proposition+ style coreference*
proposition ::= concept-instance | set aspect tmr-time modality* style
aspect ::= **aspect aspect-scope:** concept-instance
phase: begin | continue | end | begin-continue-end
iteration: integer | multiple
tmr-time ::= **time** **time-begin:** time-expr*
time-end: time-expr*
duration: concept-instance
time-expr ::= [<< | < | > | >> | >= | <= | == | !=] [YYMMDDHHMMSS | t_i]
modality ::= **modality** **modality-type:** modality-type
modality-value: (0,1)
modality-scope: concept-instance*
modality-type ::= epistemic | deontic | volitive | potential | epiteuctic | evaluative | saliency
set::= **set** **member-type:** concept | concept-instance
cardinality: [< | > | >= | <= | <>] integer
complete: boolean
excluding: [concept | concept-instance]*
elements: concept-instance*
subset-of: set
multiple: boolean
indeterminate: boolean
proper: boolean
boolean ::= true | false
coreference ::= concept-instance concept-instance+

Browsing the ontological concept NATION in CRL's KBAE

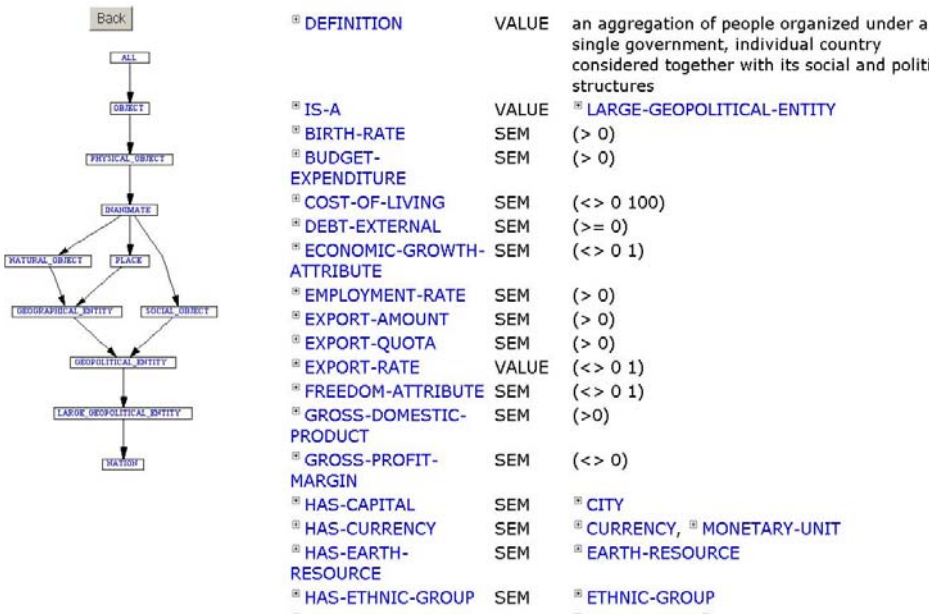

Figure 3. A view of the ontological concept NATION in CRL's Knowledge Base Acqusition Editor.

The ontological semantic lexicon contains not just semantic information, it also supports morphological and syntactic analysis. Semantically, it specifies what concept, concepts, property or properties of

A Fact Database entry for an instance of *nation*

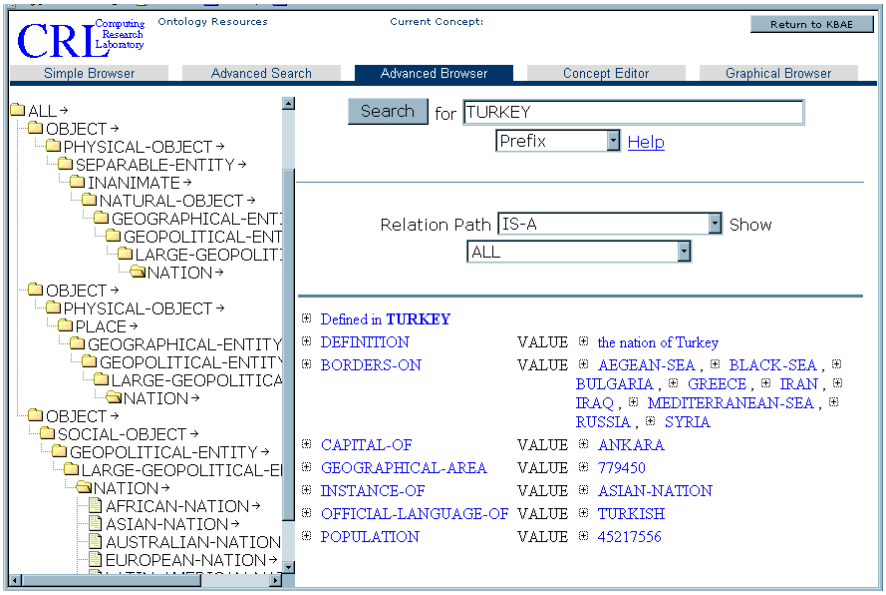

Figure 4. A view of the Fact DB entry for TURKEY in CRL's Knowledge Base Acquisition Editor.

concepts defined in the ontology must be instantiated in the TMR to account for the meaning of a given lexical unit of input.

The entries in the onomasticon directly point to elements of the Fact DB. Onomasticon entries are indexed by name (e.g., *New York*), while their corresponding entries in the Fact DB are named by appending a unique number to the name of the ontological concept of which they are instances (e.g., CITY-213). The following example illustrates the structure and content of the lexicon entry. The example shows the first verbal sense of the English lexeme *buy*:

```
buy-v1
        cat       v
        morph     stem-v   bought v+past
                           bought v+past-participle
        syn-struc
                  root     buy
                  subj     root     $var1
                           cat               n
                  obj      root     $var2
                           cat               n
                  oblique  root     from
                           cat               prep
                           opt               +
                           obj               root      $var3
                                             cat       n
        sem-struc
                  BUY
                           agent    value    ^$var1
                                             sem       HUMAN
                           theme    value    ^$var2
                                             sem       OBJECT
                           source   value    ^$var3
                                             sem       HUMAN
```

The above states that the verb *buy* takes a subject, a direct object and a prepositional adjunct; that its meaning is represented as an instance of the ontological concept BUY; that the AGENT of the concept BUY, which constitutes the meaning of the verb's subject, is expected to be a HUMAN; that the THEME of the concept BUY, which is the meaning of the verb's direct object, can be any OBJECT; and that the SOURCE of the concept BUY, which constitutes the meaning of the verb's prepositional adjunct, can be a HUMAN. The presence of variables (*$varN*) in the SYN-STRUC and SEM-STRUC zones of the entry is obviously intended to establish a kind of co-indexing. Indeed, it links syntactic arguments and adjuncts of the lexeme (if any) with the case roles and other ontological properties that the meanings (*^$varN* reads "the meaning of *$varN*") of these syntactic arguments and adjuncts fill.

One important property of our ontology is that it contains specifications of complex events ("scripts") that will play an important role in the treatment of reference. Scripts provide a real-world situational context for reference resolution. For example, when the AIR-TRAVEL-EVENT script is activated, it will help the analyzer to resolve the definite reference *the captain* in *John ran to catch his flight. As soon as he got on the plane he heard the captain announce a 40-minute delay.* It will be the captain of the plane in this particular instantiation of the script. Scripts were introduced by Schank and his associates (e.g., Schank and Abelson 1977). Recently, some of their ideas have resurfaced in AI applications (see, e.g., Lin and Hovy 2000, who use ideas ascending to early work by DeJong, e.g., DeJong 1982) and also in some work by the University of Texas Rapid Knowledge Formation group (e.g., Clark and Porter 2000). The

following example of one of our scripts, in a simplifying pseudocode, is necessarily brief, with much detail omitted. The actual grain size will be determined by the needs of an application.

```
AIR-TRAVEL-EVENT
  IS-A          TRAVEL-EVENT
  AGENT         HUMAN
  PERSONAL-INFORMATION
  PURPOSE           HUMAN.GOAL
  SOURCE            LOCATION
  DESTINATION LOCATION
  FLIGHT-INFORMATION
        DEPARTURE-FROM LOCATION
        DEPARTURE-TIME TIME
        ARRIVAL-AT     LOCATION
        ARRIVAL-TIME   TIME
        CARRIER        CORPORATION
          FLIGHT-NUMBER INTEGER
  HAS-AS-PART
        BOOK-TRAVEL (AGENT: AIR-TRAVEL-EVENT.AGENT)
        TRAVEL (AGENT: AIR-TRAVEL-EVENT.AGENT; DESTINATION: AIRPORT)
        CHECK-IN (AGENT: AIR-TRAVEL-EVENT.AGENT)
        BOARD-AIRPLANE (AGENT: AIR-TRAVEL-EVENT.AGENT)
        FLY (AGENT: PILOT; THEME: AIRPLANE)
        CLAIM-BAGGAGE (AGENT: AIR-TRAVEL-EVENT.AGENT)
        TRAVEL (AGENT: AIR-TRAVEL-EVENT.AGENT; SOURCE AIRPORT)
```

4. The Lay of the Land, Briefly

Work on reference in NLP has been quite typically partial, with the purview of various systems being determined more by practical considerations or idiosyncratic interests than by a desire to develop a broad-coverage treatment. In practical terms, dealing with all types of reference in a single system is difficult to expect from systems that are ramped up quickly in response to an evaluation challenge and/or have another task as the main goal (e.g., information retrieval). Due to external requirements, certain subsets of referring expressions – e.g., resolving elided VPs, finding chains of coreferential NPs of the 'identity' type, resolving pronominal anaphora – have become standard in NLP circles. The specific approaches to dealing with each subset, however, do not necessarily have applicability to the rest of reference issues. Our approach, by contrast, creates a unified environment for treating all types of reference.

Recent MUC evaluations have fueled this cropping of research scope by stating clear guidelines for which aspects of reference would and would not be evaluated in MUC (for a description of the guidelines, see the guidelines themselves, the MUC Coreference Task Definition (Sundheim 1995), or Gaizauskas and Humphreys 1996). The results of even the best reference resolution systems evaluated in MUC, however, have been poor: Harabagiu and Maiorano (1999) report that "in the past two MUC competitions, the high scoring systems achieved a recall in the high 50's to low 60's and a precision in the low 70's". Moreover, Bagga (1998) argues that "the precision and recall statistics used by MUC-6 give little insight, both, into the actual performance of the coreference systems, and the difficulty of the coreference task," and proceeds to suggest a different framework for evaluating coreferences and systems that attempt coreference resolution. Moreover, the approaches catered for MUC or other limited tasks are often, as reported by the designers, not readily extensible to other domains or systems (e.g., Mitkov 1998; Cardie 1992). Thus, a quick-fix, "knowledge-lean" (Harabagiu and Maiorano 1999) approach to this very difficult problem is common and understandable if one is restricted by time and task (as for MUC), but it neither is entirely satisfying nor promises to advance the field in long run.

Most systems use knowledge-lean or statistical approaches to reference resolution (e.g., Kehler 1997; Cardie and Wagstaff 1999; Mitkov 1998; Yamamoto and Sumita 1998; Cristea, Ide, Marcu and Tablan 2000; Ge, Hale and Charniak 1998). If semantics are used at all, they tend to be surfacy or "light," usually just supplementing statistical or heuristics-based approaches (Muños and Palomar 2001; Cardie 1992). Attempts to use WordNet for semantic insights have largely proven unsuccessful (Barzilay and Elhadad 1997; Poesio, Schulte im Walde and Brew 1998; Poesio, Vieira and Teufel 1997). It is interesting that most of the workers in the field use essentially the same set of heuristics (e.g., text distance and gender and person constraints for pronominal anaphora), differing, essentially, only in the method used to make these heuristics work. Ge *et al.* (1998) use machine learning techniques and Klebanov (2001) uses latent semantic analysis. While the above methods are "imported," Lappin and Shih (1996), Baldwin (1997), Ravin and Kazi (1999) and Harabagiu and Maiorano (1999) develop their own algorithms. Harabagiu and Maiorano (who also include a brief survey of reference algorithms) use distinct and rather detailed heuristics for resolving coreferences of direct nominals and pronouns. Bagga and Baldwin (1998) use the vector space model to establish distances between sentences. However, all the above methods and heuristics are still essentially combinations of the few sources of knowledge available without access to ontological-semantic information. If such information is included, as we are doing, the expressive power of heuristics immediately grows very significantly, as it is possible to use expectations and constraints encoded in selectional restrictions on members of a proposition as well as the reference resolving potential in ontological scripts.

The work of the Sheffield group (e.g., Gaizauskas and Humphreys 1996, 1997; Azzam, Humphreys and Gaizauskas 1998; Wakao, Gaizauskas and Wilks 1996; Stevenson and Gaizauskas 2000) is the closest in spirit to our approach. They, like us, separate the general environment for the treatment of reference from the actual algorithms and heuristics. The LaSIE system (Humphreys *et al*. 1998) for information retrieval, used as a substrate for reference work, "should be seen not as embodying a fixed coreference algorithm, but rather as containing a base coreference algorithm on top of which various heuristics may be added or removed or combined to test their effectiveness" (Gaizauskas and Humphreys 1997). Their work, like ours, is based on a domain model, relies on a variety of types of heuristics and involves morphological and syntactic as well as semantic processing modules. However, our approach relies on a general-purpose ontology, lexicon and fact database, not a domain model developed specifically for a particular task. In addition, the coverage of reference phenomena in our work significantly extends the purview of the Sheffield work. Finally, for the first time in this field (though see Lin and Hovy 2000 for a discussion of using scripts), we will include among the sources of heuristics non-toy ontological scripts, thus creating capabilities of predictive analysis of reference on the basis of world-context priming.

Some approaches and their results have been too limited or idiosyncratic to promise broad applicability. For example, Harabagiu and Maiorano (2000) use parallel corpora from two languages to improve reference resolution results for identity coreference, but the availability of sufficiently large parallel corpora for all languages of interest is quite limited, and their creation, quite expensive. Yamamoto and Sumita (1998) discuss the resolution of elided arguments that have an extra-sentential antecedent, but make the simplifying assumptions that there already exists a module to detect the ellipsis to begin with (a very difficult problem in itself) and that the category must have an extra-sentential antecedent (which is a strong simplifying assumption). Kehler and Shieber (1997), Hobbs and Kehler (1997), Hardt (1999), and others have worked on the problem of strict/sloppy readings of anaphora contained in elided VPs: e.g., *John loves his mother and Jack does ∅ too* (Jack may love his own mother or John's mother). However, the strict/sloppy anaphora problem, which is extremely difficult even descriptively, is not, in our opinion, among the most pressing for NLP systems and, therefore, the close attention given it by the field is somewhat surprising, considering how many more immediate problems remain understudied.

To summarize the state of the field:

- Most research has a rather narrow purview, with approaches to one subset of phenomena generally not being extensible to others.
- Knowledge-lean approaches to reference resolution are more common than semantics-based ones, but their results are relatively poor and the programs are often not generalizable.
- Most systems that attempt semantic analysis either a) use WordNet as their ontology, which engenders significant problems, as WordNet does not provide sufficient semantic content (e.g., selectional restrictions – this insufficiency being the reason for the project described in Harabagiu, Miller, *et al*. 1999), b) use a small, experiment-specific ontology, which has a small domain and/or coarse grain size of descriptions, or c) have very light semantics of the semantic marker kind.

We do not, however, wish to imply that past work cannot inform our research—in fact, it can and will, in notable ways discussed in the next section.

5. What We Are Doing

In the sections below we describe a) how we are making progress in the detection and resolution of reference, b) how prior work described in the literature will speed our progress in certain areas, c) how our approach and methods differ from related precedents, and d) what the language processing system described above will look like once the planned reference modules have been incorporated. For orientation, this last point is illustrated in Figure 5.

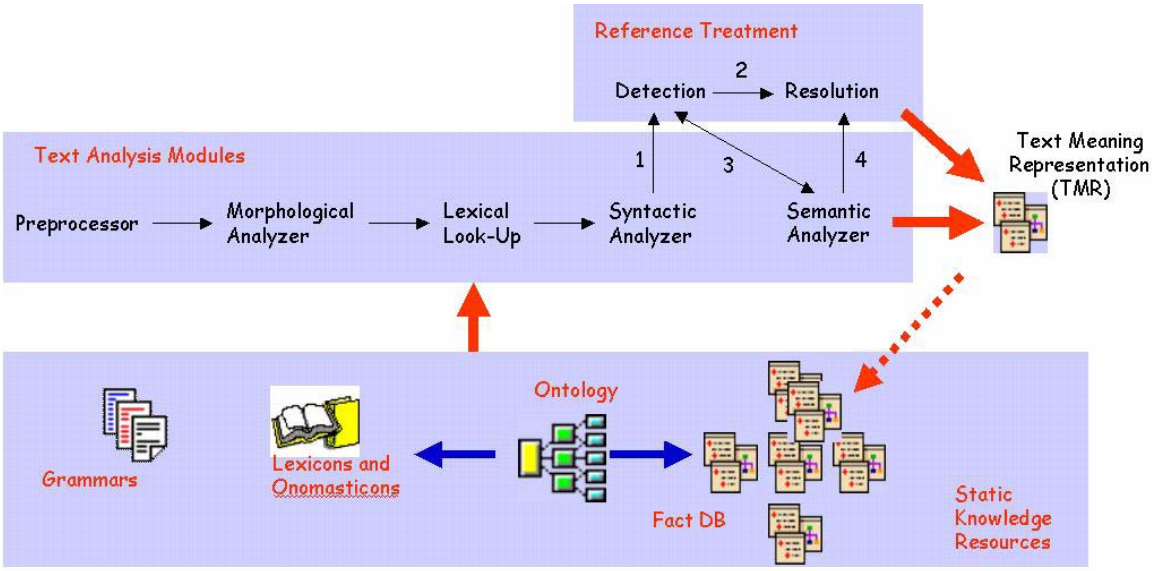

Figure 5. The environment for our treatment of reference. It is an extension of the basic semantic analyzer in Figure 2.

5.1 Detection

5.1.1 Distinguishing Referential and Non-Referential NPs and Verbs

Work on reference begins with separating non-referential NPs and verbs from referential ones. The computational literature hardly addresses this problem, so we are developing heuristics and algorithms from scratch, with help from the descriptive and theoretical literature. For example, the attributive usage of NPs is often found in copular sentences (*He is **a doctor***), so the syntactic structure NP_1- copula-NP_2, coupled with an indefinite NP_2, is a clue for the non-referential function of NP_2.

The literature has shown that one must not be too quick to classify NPs as referential or non-referential based on their form. For example, whereas one might assume that NPs with a definite article should belong to a chain of coreference, the percentage of definite NPs that, while referring to an object or an event, are not part of a coreference chain has been counted at 50% (Vieira and Poesio 1997), 63% (Bean and Riloff 1999), and 39% (Gundel, Hedberg and Zacharski 2002) for different corpora. In fact, many types of definite NPs do not corefer with previously mentioned objects or events in the text. The first type are those that are either definite by convention (*the fact that, the time when, the color blue, on the other hand*) or are definite because they refer to general world knowledge (*the international community*). Some such entities can be listed explicitly as a stoplist supplemented by a set of lexico-syntactic patterns (Seville and Ramsay 2000). However, this method will not cover everything, as noticed by Bean and Riloff (1999), who say that "existential NPs account for 63% of all definite NPs, and 24% of them could not be identified by syntactic or lexical means." They provide an algorithm for creating such a list on the fly that we are incorporating into our approach.

Other sources are also aiding our categorization of definite NPs (as referring or non-referring, part of a chain of reference or not, referring to an anchor in our database or not, etc.) by a) providing classification schemes and examples for their various types of definite NPs (Vieira 1998; Poesio, Vieira and Teufel 1997) and b) providing certain heuristics for different types of definite NPs, like syntactic and lexical clues (Bean and Riloff 1999; Seville and Ramsay 2000).

5.1.2 Categorizing Referential Objects and Events

Once candidate referring expressions are distinguished, they must be categorized. The tagging of closed-class parts of speech as personal pronouns, reflexive pronouns, deictic adverbs, etc., will be done using stoplists. Indefinite vs. definite descriptions will be determined based on their specifiers, if any (e.g., NPs with *the, this, Harry's* etc. are definite whereas those with *a, an,* or no determiner are indefinite). We are considering incorporating Mikheev's (1999) heuristics for dynamically inferring which nouns are proper and which are common in positions that always require capitalization.

5.1.3 Detecting Ellipsis

Different types of ellipsis will be detected using syntactic, lexical, and semantic clues. Sample syntactic clues, well-known in the literature, include:

- A dangling wh-word signals VP ellipsis: *We wanted to invite someone but we didn't know who ∅.*
- An auxiliary or to with no VP complement signals VP ellipsis: *They asked me to help because John couldn't ∅.*
- A latter conjunct with no verb signals gapping: *Mary married Bob and Sue ∅ Peter*.

Ellipsis can also be triggered by certain lexical items that are known to replace fully specified categories, for example:

- A determiner, a quantifier, *one*, *some*, etc., with no head noun signals that the head noun must be reconstructed: *These books are good but those ∅ are boring; A good book should be read but a bad one [=book] should be ignored*.
- The lexical item *do* must also be reconstructed in certain configurations (not VP-elliptical ones): *The girls swam faster than we expected they would do*.
- *Not* can signal ellipsis (so-called Stripping as well as other types): *Dan follows politics but not me.*

Another type of semantic ellipsis will be detected through lexical underspecification (see, e.g., Viegas and Nirenburg 1995; Pustejovsky 1995). For example, *I forgot my keys*, means, abductively, *I forgot to take my keys*. Similarly, *Bill started a book* should, in the general case, be abductively interpreted as *Bill started to read a book*; however, if the Fact DB contains information that Bill is an author, then the abductive inference will conclude that the real meaning is *Bill started to write a book.*

Our work on detecting ellipsis consists of a) constructing a full the list of heuristics for ellipsis, based on existing resources supplemented by our own research, b) constraining their applicability, as necessary, based on system trials, and c) enhancing the English lexicon to include, e.g., a call to an ellipsis-resolution procedure in entries for special words, like *begin, one,* etc.

5.2 Resolving Referring Expressions

5.2.1 Direct Referring Expressions

Under our approach, resolving direct referring expressions means linking them to an existing anchor in the TMR, FDB or ontology, or—if there is no such anchor—establishing them as a new anchor.

Publicly known personalities, organizations, etc., are expected already to have an entry in the FDB and/or onomasticon, and variants of their "anchor" names will generally be available. Therefore, identity referents with the same head noun (*President Bush, George W. Bush*) will trivially be linked to that already existing anchor. In addition, processes like those described in Ravin and Kazi (1999) will help to determine even more variants of direct referring expressions by analyzing candidate expressions across documents. Direct referring expressions that are not listed in our resources will trigger establishment of a new anchor. So-called identity referents to that anchor (*Thomas Gallagher ~ Tom Gallagher, Dr. Gallagher*) will be associated with that anchor, as well. In the cases of ambiguity (e.g., *Eduardo Frei* can refer both to a recent Chilean president and to his father, a president of Chile in the 1940s), mutually exclusive properties in the FDB or TMR will be used. It is entirely realistic to have no local disambiguating clues, in which case all the reference candidates will be carried on for later disambiguation.

5.2.2 Indirect Referring Expressions

Indefinite indirect referring expressions (***A man*** *walked in*) will always initiate a new anchor, barring exceptions we have yet to encounter (remember that the non-referring indefinite expressions will have been already detected by this point in the process).

Definite indirect referring expressions (*The man, The first day of winter*) are considerably more complex to resolve and much literature has been devoted to them. Many such expressions refer to an already-established discourse topic or, in our terms, anchor. The lexical and semantic relationship between them may, however, take many forms (Bergler and Knoll (1996) pursue this issue using their own categories for a Wall Street Journal corpus). The definite indirect referring expression might:

- contain the same head as its anchor (*Manhattan Ridge Bank ~ the failing bank*), in which case we will use a head-matching program plus feature comparison heuristics. The literature, however, reveals complications regarding the latter process: e.g., Kehler (1997) discusses the analysis of a text containing the successive strings *Kingston Military Rail Depot (title), a rail depot, the*

ammunition depot and *the depot*—the last of which must corefer to the first two but not the third, even though it is most recent; thus, special rules will have to be created to tighten the reference search;

- be a hypernym of the anchor (*Manhattan Ridge Bank ~ the financial institution*), in which case we will measure the ontological distance between potentially coreferring elements to determine if they link to the same anchor;
- be a hyponym of the anchor (*Manhattan Ridge Bank ~ the 6th Street Branch of Manhattan Ridge Bank*), which will be treated similarly as a hypernym;
- be a description of the anchor (*Manhattan Ridge Bank ~ the tall brick building on 6th and Main*), in which case we will search the FDB to see if we have enough previously gathered information about the anchor to suggest linking the given element to it;
- be an event describing the NP anchor or vise versa (*I **had** a Greek salad and chocolate cake. **That meal** was just what the doctor ordered*), which will be identified if possible by the complex events, or scripts, in the ontology;
- be part of an ontological *script* that the anchor triggers; for example, if the RESTAURANT script is triggered, the definite references in text will be checked against expectations created by the existence in the script of potential anchors for referring expressions: *I went to a restaurant and **the table** was dirty, **the waitress** was slow and **the food** was bad.*

All of these types of correspondence except the first (where the heads match) are called 'bridging' constructions. These constructions have been discussed in many sources (e.g., Poesio, Schulte im Walde and Brew 1998; Poesio, Vieira and Teufel 1997; Asher and Lascarides 1996) and attempts to process bridging phenomena have been reported in some of them. Developing ontology-based heuristics for resolving bridging phenomena is a central component of our work.

In searching for anchors for **textual pointers**, we will use not only ontological-semantic knowledge but are also considering incorporating non-semantic heuristics presented in the literature. These heuristics rely primarily on feature matching (gender, number), the distance between referents (with closer antecedents/postcedents having priority), and the number of prior repetitions of candidate coreferring elements (with more repetitions having priority) (see, e.g., Morton 1999; Muñoz and Palomar 2001; Mitkov 1998; Ge, Hale and Charniak 1998). For some classes of textual pointers—e.g., reflexive pronouns—syntactic constraints indeed quite definitively indicate the antecedent/postcedent. For many others, however, deeper analysis is necessary. This is where ontological semantics should give us an advantage. For example, in addition to gender and number, we also have access to the results of the "unilateral" application of selectional restrictions, which are stored in the ontological-semantic lexicon in entries for the textual elements whose meanings are members of the same semantic dependency structure as the referring expressions. Thus, in the sentence *It idled*, the anchor for *it* is selectionally constrained to ENGINE based on the selectional constraint on the INSTRUMENT of the ontological concept IDLE in our ontology. The reasoning in the reference resolution algorithm will be as follows. The case role INSTRUMENT is used because AGENT in the current ontology is semantically constrained to sentient entities and forces, such as gravity. The use of the pronoun *it* rules out an animate agent; and no forces are selectionally compatible with the concept IDLE. Overall, we will use an unprecedented combination of clues when searching for anchors for textual pointers.

We are also considering using the results of some of the statistics-based approaches as a contribution to coreference resolution (e.g., Kehler 1997), but this will be more of a "last-resort" heuristic for choosing between variants than a fundamental part of the procedure. In this project, we do not plan to pursue separate discourse modelling component of the type used, e.g., in centering theory (e.g., Grosz and Sidner 1998). The practical benefits of such modelling for reference resolution have been mixed and, besides, our TMRs actually include (through included instances of discourse relations) a discourse model.

Deictic pointers are a subclass of closed-class lexical items. Each will be provided with a procedure that is fired during the process of reference resolution. The literature contains little work on this type of reference resolution, so heuristics will need to be developed from scratch based on findings in test corpora. A first approximation of the procedure appended to the lexicon entry for a word like *here* is: find the most recent reference to location in the text; if none, search for clues [x, y, …] in the pragmatic context or speech situation; etc.

6. Anticipated Challenges and Evaluation of Results

The work we have undertaken is scientific research that carries risk. We must prove that ontological semantic analysis improves quality levels of reference resolution, though we have every reason to believe that it will. We also have not yet determined how best to integrate all the knowledge sources and different types of algorithms and heuristics that will be fundamental to the approach, and whether we will use one or more of the existing methods of combining evidence or develop a custom method specific to the problem of reference resolution. The process of integration, in fact, will be in large part a matter of evaluating the results of training runs of the system to determine which types of rules with what priority produce the best outcome.

The best evaluation for any operational NLP system is to put it in a real application and measure the overall performance improvement. This type of evaluation is expensive and does not allow for testing system components separately. For this reason, a new regimen of evaluation has been introduced in the field in the late 1980s that compares the performance of an NLP system component on a narrowly defined task outside a real application with results obtained by humans on the same task. We will evaluate our system by comparing the results of reference resolution on ten texts of average newspaper article length carried out by 1) a person, 2) an automatic reference resolution module that uses all the "traditional" algorithms and heuristics (that is, all the algorithms we develop minus ontological semantic analysis), and 3) automated reference resolution using our full inventory of algorithms, heuristics and ontological semantic analysis. Evaluation will take place at the level of the TMR, where links to anchors as well as coreference chains can be analyzed. The evaluation will center around what types of phenomena we can and cannot adequately cover and the reasons why we cannot cover the missed ones. Since we are not seeking full lexicographic coverage but, rather, progress in a more programmatic sense, we do not expect the system at the end of this project to offer full coverage of reference for any application in any realm. However, our environment (including the ontology, the FDB, lexicons, etc.) is constantly being expanded, improved, and applied to projects ranging from machine translation to knowledge extraction. The progress in reference that we expect to make in this project will feed into all such applications. Task-oriented evaluation must, of course, follow, but that will be part of those other application-oriented projects.

7. Further Implications of this Work

Our work, we believe, will fundamentally change the path of research and development in the treatment of reference in NLP, which currently represents a substantial logjam in all areas of application, from machine translation to question answering to automatic text summarization. We are raising the bar of expectations by stating that knowledge-poor NLP is not adequate for the task and that knowledge-rich NLP is attainable (i.e., it is *not* too expensive to build full-sized ontologies linked to large lexicons, FDBs, etc.). Our task is to show that the knowledge-rich environment and methods we are developing are superior to all other current methods and represent the correct direction for future research and development.

The importance of high-quality NLP to modern-day society cannot be overstated. The more reliable NLP-based systems become, the more people will be able to use them with confidence in the workplace, the academic arena, the military, and to foster their private interests. Although multilinguality is not an

express focus of the preliminary research and development, the architecture within which we will be working has been and continues to be used in multilingual projects. Moreover, our plan to conceptualize phenomena in terms of cross-linguistically valid parameters and values (see Section 5) means our reference results for English can be directly applied to other languages in later projects. As our society becomes more globally oriented, access to things like reliable English-language summaries of texts from another language and high-quality translation systems will become ever more a part of our everyday lives. Systems that inspire such confidence, however, remain to be built, and our goal is to do this with a full understanding of the challenges and with long-term success as the goal.

Our research into reference resolution using ontological semantics pushes the envelope of linguistics, computational linguistics, information technology and knowledge engineering. Ontological semantics seeks practical solutions for age-old linguistic problems, like reference and coreference, and makes progress in the field of linguistics itself while applying the results to NLP systems.

References

Asher, Nicholas and Alex Lascarides. 1996. 'Bridging.' R. van der Sandt, R. Blutner and M. Bierwisch, eds., *From Underspecification to Interpretation, Working Papers of the Institute for Logic and Linguistics.* IBM Deutschland, Heidelberg.

Azzam, Saliha, Kevin Humphreys and Robert Gaizauskas. 1998. 'Extending a Simple Coreference Algorithm with a Focusing Mechanism.' *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC2)*, Lancaster, 15-27.

Bagga, Amit. 1998 'Evaluation of Coreferences and Coreference Resolution Systems.' *Proceedings of the First Language Resource and Evaluation Conferene*, May 1998.

Bagga, A. and B. Baldwin. 1998. 'Entity Based Cross-Document Coreferencing Using the Vector Space Model.' *36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, 79-85.

Baldwin, Breck. 1997. 'CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources.' Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, Madrid, Spain, 38-45.

Barzilay, Regina and Michael Elhadad. 1997. 'Using Lexical Chains for Text Summarization.' *Proceedings of the Intelligent Scalable Text Summarization Workshop, ACL*, Madrid, Spain.

Beale, Stephen and Sergei Nirenburg.h 2002. 'An Ontological-Semantic Analyzer.' *Submitted to Coling-02*.

Bean, David L. and Ellen Riloff. 1999. 'Corpus-Based Identification of Non-Anaphoric Noun Phrases.' *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).*

Bergler, Sabine and Sonja Knoll. 1996. 'Coreference Patterns in the Wall Street Journal.' Carol Percy, Charles Meyer, and Ian Lancashire, eds., *Synchronic Corpus Linguistics*. *Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora*, Toronto 1995.

Cardie, Claire. 1992. 'Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics.' *Proceedings of the 30th Annual Meeting of the ACL*, University of Delaware, Newark, DE.

Cardie, Claire and Kiri Wagstaff. 1999. 'Noun Phrase Coreference as Clustering.' *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Association for Computational Linguistics, 82-89.

Clark, Peter and Bruce Porter. 2000. '$RESTAURANT Re^{n}-visited: A KM Implementation of a Compositional Approach.' Technical Report, AI Lab, University of Texas at Austin.

Cristea, Dan, Nancy Ide, Daniel Marcu and Valentin Tablan. 2000. 'Discourse Structure and Co-Reference: An Empirical Study.' *The 18th International Conference on Computational Linguistics, COLING 2000*, Luxembourg, July 31-August 4.

DeJong, Gerald. 1982. 'An Overview of the FRUMP System.' Lehnert, Wendy G. and Martin H. Ringle, eds., *Strategies for Natural Language Processing,* 149-176.

Gaizauskas, Robert and Kevin Humphreys. 1996. 'Quantitative Evaluation of Coreference Algorithms in an Information Extraction System.' Botley, S. and T. McEnery, eds., *Corpus-Based and Computational Approaches to Discourse Anaphora.* UCL Press and the Centre for Computational Linguistics, UMIST.

Gaizauskas, Robert and Kevin Humphreys. 1997. 'Using a Semantic Network for Information Extraction.' Department of Computer Science, University of Sheffield, Technical Report CS-97-03.

Ge, N., J. Hale and E. Charniak. 1998. 'A Statistical Approach to Anaphora Resolution.' *Proceedings of the Sixth Workshop on Very Large Corpora,* 161-171.

Grosz Barbara and Candace Sidner. 1998 'Lost Intuitions and Forgotten Intentions.' Marilyn Walker, Aravind Joshi, and Ellen Prince, eds., *Centering Theory in Discourse*, 89-112. Clarendon Press. Oxford.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 2002. Forthcoming. 'Cognitive Status and Definite Descriptions in English: Why Accommodation is Unnecessary.' *Journal of English Language and Linguistics*.

Harabagiu, Sanda M. and Steven J. Maiorano. 1999. 'Knowledge-Lean Coreference Resolution and its Relation to Textual Cohesion and Coherence.' *Proceedings of the ACL Workshop on the Relation of Discourse/Dialogue Structure and Reference*, 29-38.

Harabagiu, Sanda M. and Steven J. Maiorano. 2000. 'Multilingual Coreference Resolution.' Proceedings of the Language Technology Joint Conference on Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics ANLP-NAACL 2000, May, Seattle WA

Harabagiu, Sanda M., George A. Miller and Dan I. Moldovan. 1999. 'WordNet 2 – A Morphologically and Semantically Enhanced Resource.' *Proceedings of the SIGLEX Workshop*.

Hardt, Daniel. 1999. 'Ellipsis and the Structure of Discourse.' *Workshop on Ellipsis and Information Structure*. Berlin ZAS.

Hirshman, L. and N. Chinchor. 1998. 'MUC-7 Coreference Task Definition. Version 3.0.' *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Applications International Corporation. http://www.muc.saic.com/: Science.

Hobbs, Jerry R. and Andrew Kehler. 1997. 'A Theory of Parallelism and the Case of VP Ellipsis.' *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, July, 394-401.

Humphreys, Kevin, Robert Gaizauskas and Saliha Azzam. 1997. 'Event Coreference for Information Extraction.' *Proceedings of the ACL/EACL '97 Workshop: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.* July 1997, Madrid, Spain.

Humphreys, K., R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham and Y. Wilks. 1998. 'Description of the LaSIE-II System as Used for MUC-7.' *MUC-7*.

Kehler, Andrew. 1997 'Probabilistic Coreference in Information Extraction.' *Empirical Methods in Natural Language Processing*, Volume 2.

Kehler, Andrew and Stuart Shieber. 1997. 'Anaphoric Dependencies in Ellipsis.' *Computational Linguistics*, 23(3): 457-466.

Klebanov, Beata. 2001. 'Using Latent Semantic Analysis for Pronominal Anaphora Resolution.' MSc Dissertation, Division of Informatics, University of Edinburgh.

Lappin, Shalom and Hsue-Hueh Shih. 1996. 'A Generalized Reconstruction Algorithm for Ellipsis Resolution.' *16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, 687-692.

Lin, C. and E. H. Hovy. 2000. 'The Automated Acquisition of Topic Signatures for Text Summarization.' *Proceedings of the COLING Workshop on Text Summarization*. August, Strasbourg, France.

Mahesh, Kavi, Sergei Nirenburg and Stephen Beale. 1997. 'If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation.' *Proceedings of Theoretical and Methodological Issues in Machine Translation (TMI-97)*, Santa Fe, NM, 1-9.

MUC-6. Proceedings of the Sixth Message Understanding Conference. 1995. (MUC-6). Morgan Kaufmann.

MUC-7. Proceedings of the Seventh Message Understanding Conference. 1998. (MUC-7). Morgan Kaufmann.

Mikheev, Andrei. 1999. 'A Knowledge-Free Method for Capitalized Word Disambiguation.' *Proceedings of ACL'99.* Maryland, June 1999.

Mitkov, Ruslan. 1998. 'Robust Pronoun Resolution with Limited Knowledge.' *Proceedings of ACL '98*, 869-875.

Mitkov, Ruslan. 2000. 'Anaphora Resolution: The State of the Art.' Wolverhampton: School of Languages and European Studies, University of Wolverhampton.

Morton, Thomas S., 1999. 'Using Coreference in Question Answering.' *ACL Workshop on Coreference and Its Applications*.

Muñoz, Rafael and Manuel Palomar. 2001. 'Semantic-Driven Algorithm for Definite Description Resolution.' *Proceedings of Recent Advances in Natural Language Processing, RANLP 2001*, Tzigov Chark, Bulgaria. September, 180-186.

Nirenburg, Sergei and Victor Raskin. 2002. Submitted. *Ontological Semantics.*

Poesio, Massimo, Renata Vieira and Simone Teufel. 1997. 'Resolving Bridging References in Unrestricted Text.' *Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, Madrid, July 1997, 1-6.

Poesio, Massimo, Sabine Schulte im Walde and Chris Brew. 1998. 'Lexical Clustering and Definite Description Interpretation.' *Proceedings of the AAAI Spring Symposium on Learning for Dscourse.* Stanford, CA, March 1998, 82-89.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Ravin, Yael and Zunaid Kazi. 1999. 'Is Hillary Rodham Clinton the President? Disambiguating Names across Documents.' *Proceedings of the ACL 1999 Workshop on Coreference and Its Applications*. Maryland, USA, June.

Schank, Roger and Robert Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures.* New York: L. Erlbaum Associates.

Seville, Helen and Allan Ramsay. 2000. 'Making Sense of Reference to the Unfamiliar.' Kay, M. ed., *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Universit at des Saarlandes, 775-781.

Stevenson, Mark and Robert Gaizauskas. 2000. 'Improving Named Entity Recognition Using Annotated Corpora.' *Proceedings of the LREC Workshop 'Information Extraction Meets Corpus Linguistics.'* Athens, Greece.

Sundheim, Beth. 1995. 'The MUC coreference task definition v. 3.0.' *Proceedings of the 6th Message Understanding Conference*.

Viegas, Evelyne and Sergei Nirenburg. 1995. 'The Semantic Recovery of Event Ellipsis: Its Computational Treatment.' *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montréal, Québec, Canada.

Vieira and Poesio 1997. 'Corpus-Based Processing of Definite Descriptions.' Botley and McEnery, eds., *Corpus-Based and Computational Approaches to Discourse Anaphora*. Amsterdam and Philadelphia: J. Benjamins.

Vieira, Renata. 1998. 'A Review of the Linguistic Research on Definite Descriptions.' Available at http://citeseer.nj.nec.com/vieira98review.html.

Wakao, Takahiro, Robert Gaizauskas and Yorick Wilks. 1996. 'Evaluation of an Algorithm for the Recognition and Classification of Proper Names.' *Proceedings of the 16th International Conference on Computational Linguistics (COLING96)*, Copenhagen, 418-423.

Yamamoto, Kazuhide and Eiichiro Sumita. 1998. 'Feasibility Study for Ellipsis Resolution in Dialogues by Machine-Learning Technique.' *Proceedings of COLING-ACL'98*, 1428-1435.