University of Alberta

Web-assisted Anaphora Resolution

by

Yifan Li

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

Spring 2010
Edmonton, Alberta

Examining Committee

Petr Musilek, Electrical and Computer Engineering

Marek Reformat, Electrical and Computer Engineering

Ivan Fair, Electrical and Computer Engineering

Witold Pedrycz, Electrical and Computer Engineering

Richard Sutton, Computing Science

Slawomir Zadrozny, Polish Academy of Sciences

Abstract

This dissertation investigates the utility of the web for anaphora resolution. Aside from offering a highly accurate, web-based method for pleonastic *it* detection, which eliminates up to 4% of errors in pronominal anaphora resolution, it also introduces a web-assisted model for definite description anaphoricity determination and a prototype system of anaphora resolution that uses the web for virtually all subtasks.

The thesis starts with a thorough analysis of the relationship between anaphora and definiteness, a study that bridges the gap between previously reported empirical studies of definite description anaphora and the linguistic theories developed around the concept of definiteness. Various naturally-occurring definite descriptions found in the WSJ corpus are analyzed from both perspectives of familiarity and uniqueness, and a new classification scheme for definite descriptions is developed.

With the fundamental issues solved, the rest of the thesis focuses on the various ways the web can be exploited for the purpose of anaphora resolution. This thesis presents methods of high-precision, high-recall anaphoricity determination for both pronouns and definite descriptions. Evaluation results suggest that the performance of the pleonastic *it* identification module is on par with casually-trained human annotators. When used together with a pronominal anaphora resolution system, the module offers a statistically significant performance gain of 4%. The performance of the anaphoricity determination module for definite descriptions, which benefits from both the insight gained from the study on anaphora and definiteness and the significantly expanded coverage offered by the web, is also one of the highest among existing studies. The thesis also introduces a web-centric anaphora resolution system. Aside from serving as the information source for implementing selectional restrictions and discovering hyponym/synonym relationships, the web is additionally used for gender/number determination and many other auxiliary tasks, such as determining the semantic subjects of *as*-prepositions, identifying antecedents for certain empty categories, and assigning appropriate labels for proper names using information available from the text itself. With a design that specifically leaves room for the application of verb-argument and genitive co-occurrence statistics, the web-based features provide statistically significant gains to the system's performance.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

At its Ancient Greek root, the term **anaphora** literally means 'the act of carrying back'. Today, in the field of linguistics, it is used to denote a variety of phenomena where a word or phrase is associated with its previous mention. As Hirst (1981) defines it, anaphora is 'a device for making an abbreviated reference (containing fewer bits of disambiguating information, rather than being lexically or phonetically shorter) to some entity (or entities) in the expectation that the receiver of the discourse will be able to disabbreviate the reference and, thereby, determine the identity of the entity'. The abbreviated reference is denoted an **anaphor**, and the discourse element which its interpretation depends upon is identified as its **antecedent**. Anaphora is ubiquitous in natural languages. For example, the pronoun, one of the most important instruments of anaphora, is identified as universal across languages:

> Universal 42: All languages have pronominal categories involving at least three persons and two numbers. (Greenberg, 1963, page 96)

Indeed, without extensive use of pro-forms and other forms of anaphora, a discourse will soon spin out of control upon the need to completely specify all previously mentioned entities again and again. Speakers of the English language tend to make extensive use of anaphora. A close examination of selected news articles from the Wall Street Journal Corpus (WSJ; Marcus et al., 1993) reveals that up to 30% of nominal expressions are anaphoric. Due to the widespread use of anaphora, identifying the antecedent of an anaphor (the process of which is termed **anaphora resolution**) is crucial to any application that involves none-trivial natural language text understanding, most notably information extraction (e.g. Bonzi & Liddy, 1989; González & Rodríguez, 2000, and Sanchez-Graillet et al., 2006), automatic text summarization (cf. Kabadjov et al., 2005), and machine translation (cf. Mitkov, 1999). To illustrate, an information extraction system trying to keep track of corporate activities may find itself dealing with news pieces such as (1.1):

(1.1) Texas Instruments Japan Ltd.$_1$, a unit of Texas Instruments Inc., said *it*$_1$ opened a plant$_2$ in South Korea to manufacture control devices. *The new plant*$_2$, located in Chinchon about 60 miles from Seoul, will help meet increasing and diversifying demand for control products in South Korea, *the company*$_1$ said. *The plant*$_2$ will produce control devices used in motor vehicles and household appliances. WSJ 17:1-3[1]

This three-sentence article focuses on two entities, a company named *Texas Instruments Japan Ltd.* and the new plant that it recently opened in Chinchon. Each entity is mentioned three times – once in the beginning to introduce the entities into the discourse, with all subsequent mentions being anaphoric. Without associating the anaphors *it*$_1$, *The new plant*$_2$, and *The plant*$_2$, with their respective antecedents, *Texas Instruments Japan Ltd.*$_1$ and *a plant*$_2$, it would be impossible to determine the ownership of the plant or its location and purpose.

1.1 Motivations

Natural languages are instruments developed by human-beings and are used for the sole purpose of communication among human-beings. As researchers attempt to extend the audience to computers, one of the primary difficulties they face is that humans make various assumptions as they speak or write. Aside from the most basic assumption that the receiver is capable of the language per se, it is also assumed that he or she has enough background knowledge, or world knowledge, to decode and reason upon the presented message.

Since the process of anaphora resolution is largely one that attempts to re-attach information – previously stripped off by speakers in order to make their utterance more cohesive – to the anaphors, it follows naturally that world knowledge plays a vital role in this process. There are a number of syntactic and discourse theories that provide guidance to the anaphora resolution process, and many have been applied successfully in real-world systems. These theories usually prescribe the likelihood (or infeasibility) of an anaphoric relationship given a pair of words. Unfortunately, the constraints provided by the syntactic rules are often not enough to eliminate all the false candidates, while the remaining prescriptions are suggestive in nature. Without world knowledge, many such prescriptions are in a sense statistical, i.e. they reflect generic language use patterns, which may be overridden when content cues are strong enough[2]. Examples (1.2) and (1.3) each offer two sentences that have exactly the same surface structures while the antecedents of the anaphors vary according to the contents.

(1.2) a. Each child ate a biscuit. *They* were delicious.

b. Each child ate a biscuit. *They* were delighted.

Mitkov (2002, ex. 2.14-15)

(1.3) a. The soldiers shot at the women and *they* fell.

[1] Unless noted otherwise, all example sentences are selected from the WSJ corpus, with their original locations encoded in the format `article-id:sentence-id`. Continuous sentence blocks are marked with the begining and ending sentence numbers.

[2] In fact, virtually all syntactic constraints have counter examples.

b. The soldiers shot at the women and *they* missed.

Mitkov (2002, ex. 2.20-21)

In (1.2), the crucial knowledge for resolving the pronoun *they* is that a biscuit can be delicious and a child can be delighted, but not vice versa. In (1.3), it must be understood that in the event of shooting, only those who fired can miss, and those who were shot usually fell. In absence of world knowledge, existing theories would only be able to obtain one correct anaphoric relationship in each example. The usefulness of syntactic guidance diminishes quickly when associative anaphora is considered – as shown by examples (A.35) through (A.41), anaphors and antecedents in an associative relationship are restricted by few, if any, syntactic constraints. In fact, although not very common, associative anaphors can take the form of non-definite expressions, a sharp contrast to coreferential cases.

The utility of world knowledge in the domain of anaphora resolution is not limited to antecedent candidate selection. As illustrated by examples (1.4) and (1.5), it also plays a key role in distinguishing non-anaphoric mentions from the anaphoric ones.

(1.4) But Sony ultimately . . . fired Mr. Katzenstein, after he committed the social crime of making an appointment to see the venerable Akio Morita, founder of Sony. *It*'s a shame their meeting never took place[†]. WSJ 37:34-35

(1.5) More and more corners of *the globe* are becoming free of tobacco smoke. WSJ 37:37

In (1.4), the crucial knowledge for identifying *It* as non-anaphoric is the combination of the sentence's syntactic structure and the matrix predicate *a shame*. In (1.5), the vital information is that *the globe* can be interpreted as referring to the planet where we live, a unique entity that is familiar to all. While the pronoun *it* is most-often used anaphorically, a rather significant portion of its uses belongs to the non-anaphoric category. An even larger portion of definite descriptions (noun phrases introduced by the definite article *the*) – roughly half – is also non-anaphoric. Obviously, being able to correctly identify the non-anaphoric cases contributes positively to the precision of an anaphora resolution system. Perhaps more importantly, not being able to do so results in wrong interpretations of the sentences, which may further propagate to the systems that rely on anaphora resolution. For example, if the non-anaphoric instance of *it* in (1.4) is not correctly identified and is linked to either one of the mentions of *Sony*[3], a text summarization system consuming this information may choose to collapse the coreference chain and produce '*Sony is a shame.*'

Once it is recognized that world knowledge is indispensable to the process of anaphora resolution, the remaining question is then how to give an artificial system access to it. Humans accumulate world knowledge through continuous learning; and one of the most effective ways of learning is to ask questions. Compared to the other sophisticated tools of learning that the humans are equipped

[3] In fact, the two mentions of *Sony* are the strongest candidates considering the syntactic structure of the sentence, recency, and gender/number information.

with, it is also relatively easy to implement in an artificial system – all that is needed is a searchable information source. Similar to the case of humans, a four-part process is involved in a query-based model of anaphora resolution:

1. Identifying the key questions that need to be answered in order to diagnose a specific discourse entity as (non-)anaphoric and to select the correct antecedent from a potentially large set of candidates.
2. Identifying the source of information capable of answering such questions.
3. Formulating the questions so that the information source can be best exploited.
4. Interpreting the answers and applying the newly-gained knowledge.

While many previous studies rely on manually-constructed knowledge sources such as word lists or thesauri, recent years have seen a growing interest in exploiting the web as a source of knowledge for different semantic relationships. Inspired by the success of these web-based approaches, this study adopts the web as the primary source of information. Naturally, the first research question following the decision is whether the web's utility is limited to the areas that have been already explored. In other words, can the web offer more for anaphora resolution? Compared to manually compiled corpora or thesauri, the web is also significantly more noisy, a shortcoming that is reflected in some previous studies as inferior precision figures. This leads to the second research question – is it possible to overcome the noise problem and obtain highly accurate results from the web, at least for some tasks? With a query-based process, the question directly corresponds to the third point outlined above: a 'good question' against the web must include some inherent features to suppress the influence of unwanted results.

The remaining research questions are not directly related to the web, but are also central to the issue of anaphora resolution: how do the fundamental properties of definite descriptions relate to the notion of anaphora? And more importantly, how does this relationship affect non-anaphoricity determination and anaphora resolution? Unlike the first two questions, they were not present at the inception of this work but have gradually grew more and more pestering as I gradually looked into the different aspects of definite description anaphora. Reviewing existing literatures reveals some confusing terms and practices, as well as contradicting factors proposed by different researchers. For example, the term 'discourse-new' frequently appears in studies on definite description anaphora. However, does it mean that 'discourse-old' entities are all anaphoric? If that is the case, why do we generally consider subsequent mentions of named entities non-anaphoric? And while we consider them non-anaphoric, why are they often put into the same group as the other anaphoric definite descriptions during classification? Another example is that while Bean (2004) consider ordinal number (subsumed under the category 'count terms') as a sign for anaphoric interpretation, Vieira (1998) include the term *first* in her 'special predicate' list for non-anaphoricity, which is processed first on the decision tree. In addition, some researchers also explicitly raised the concern that existing linguistic theories do not provide sufficient support for definite description annotation. All these ultimately point to the need to study the definite descriptions in more details under the context of anaphora resolution.

1.2 Main Contributions

This study offers both new theoretical insights for definite description anaphora and novel practical methods that exploit the web for better anaphora resolution.

First, the relationship between anaphora and definiteness is examined in details under the context of definite descriptions, and a new notion of definite noun phrase anaphora is developed. The new definition, which identifies anaphora as a device to satisfy the uniqueness and familiarity presuppositions of definiteness, is further used as a guide to design a novel classification scheme for definite descriptions. These new developments help bridge the gap between the empirical studies such as that of Poesio and Vieira (1998) and the linguistic theories developed around the concept of definiteness, one that have been noted for a long time but has not received a satisfactory answer. The analysis performed on the various naturally-occurring linguistic phenomena found in the WSJ corpus allows us to better understand the licensing conditions for non-anaphoric uses of definite descriptions and develop more accurate syntactic cues to identify such uses. As a side effect of the theoretical analysis, a number of issues related to current practices, such as the inadequacy of the notion of 'discourse-new' and some of the widely-used patterns proposed by Vieira (1998) for anaphoricity determination, are also identified.

Second, this thesis presents methods of high-precision, high-recall anaphoricity determination for both pronouns and definite descriptions. A common theme behind the two methods is that they are both based on careful studies of relevant linguistic theories and both exploit the richness of information offered by the web. Evaluation results suggest that the performance of the pleonastic *it* identification module is comparable to that of casually-trained human annotators. It is also relatively robust and works well on output of state-of-the-art parsers. The high performance (96% precision and 87% recall for parser-generated data on the test corpus) offered by the module makes it practically useful for anaphora resolution – when evaluated together with a pronominal anaphora resolution system, it offers a statistically significant performance gain of 4%. The performance of the anaphoricity determination module for definite descriptions is also one of the highest among similar studies. Again, the high precision it offers (97%) allows it to be used as a first step in definite description anaphora resolution. The success of the two modules indicates that given well-designed questions, the web can provide answers to linguistic questions beyond simple semantic relationships.

Third, a web-centric anaphora resolution system is introduced. Since one of the primary goals of the study is to identify potential areas where the web may be helpful, the system takes a rule-based approach instead of following the current trend of using machine-learning based approaches to anaphora resolution. The system seeks to combine previous rule-based methods, namely Hobbs' (1978) naive algorithm for pronominal anaphora resolution and Vieira and Poesio's (2000) algorithm for definite description anaphora, with information gained from the web. Aside from the well-known verb-argument selectional restrictions for pronouns and hyponym/synonym relationship determination for definite descriptions, the system also uses the web for gender/number determination and many other auxiliary tasks such as the determination of the semantic subjects of *as*-prepositions, antecedent identification for certain empty categories, and determining appropriate labels for proper names using information available from the text itself. With a design that specifically leaves room for the application of verb-argument and genitive co-occurrence statistics, the web-based features provide statistically significant gains to the system's performance.

1.3 Thesis Outline

The thesis is organized as follows. For readers not familiar with the various phenomena of anaphora, a brief overview is offered in Appendix A. Chapter 2 provides relevant background on the theories and implementations of anaphora resolution algorithms.

Chapter 3 focuses on the theoretical side of definite description and definite noun phrase anaphora. The first part of the chapter addresses the perceived 'need' for anaphoric interpretation, and develops a notion of anaphora based on the fundamental presuppositions of definiteness as coined by Roberts (2003). Compared to the other often-cited definitions, such as the one by Hirst (1981), quoted at the beginning of this chapter, the new notion bridges the gap between the practice of definite description anaphora and the theoretical works on the essence of definiteness. A number of closely related concepts, such as salience, indefinite description anaphora, and coreference, are also examined. The second part of the chapter discusses the various uses of definite descriptions observed in the corpus, examines their properties from both perspectives of familiarity and uniqueness, and develops a new classification scheme of definite descriptions. The scheme is applied to corpus annotation and the results are presented in the last part of the chapter.

Chapter 4 presents the design of an anaphora resolution system that uses the web as its primary source of information. All main components of the system – gender/number information acquirement, deep syntactic structure analysis, label assignment for named entities, pronominal anaphora resolution, definite description anaphoricity determination, and definite description anaphora resolution – depend heavily on the web. The pronominal anaphora subsystem implements Hobbs's (1978) naive algorithm on the dependency structure, but with some significant twists – the original algorithm is only used to provide salience scores based on syntactic-distance. The distance measure is grouped into tiers in order to provide room for the application of web-based tests for verb-argument and genitive relationships. The accuracy of this measure is also improved as the system attempts to assign antecedents to certain empty categories. Equipped with better understanding of the essence of definite description anaphora gained from Chapter 3, a set of more accurate syntactic cues are proposed for the purpose of identifying non-anaphoric definite descriptions. The performance of the anaphoricity detector is further boosted by three sets of queries targeting different syntactic constructs. The definite anaphora resolution subsystem is also rule-based. While many elements of the design are inspired by Vieira and Poesio's (2000) study, the fundamental philosophy is quite different. Instead of making the distinction between 'direct' (i.e. the anaphor and the antecedent share the same head noun) and 'indirect' anaphors, the system treats them equally but at the same time prescribe different treatments for simple definite descriptions (i.e. those devoid of additional descriptive contents) and the complex ones. This design is also motivated by the discussion about salience and definite descriptions offered in Chapter 3.

Chapter 5 introduces a novel, web-based pleonastic *it* detector. The extrapositional cases are identified using a series of queries against the web, and the cleft cases are handled with a simple set of syntactic rules. At the core of the system are two sets of query patterns – the *what*-cleft pattern, which transforms the potential extrapositional case into a pseudo-cleft, and the comparative expletiveness test pattern, which directly 'asks' the web about the feasibility of replacing the original *it* with other personal and relative pronouns and compares the result with that of the original construction. As shown by the evaluation results, the comparative study is an effective means to get highly accurate results from the web despite the fact that it is noisier than the manually compiled corpora.

Finally, Chapter 6 recapitulates the main findings of the thesis and suggests possible directions for future work.

Chapter 2

Background and Related Work

This chapter describes the state of the art in computational treatment of anaphora resolution. The past three decades – since Hobbs (1978) first demonstrated that it is possible for a computational approach to achieve non-trivial results in resolving pronominal references – have seen rapid proliferation of research in this field. However, their distribution among the different subfields is significantly skewed, with the vast majority focusing on resolving personal pronouns with nominal antecedents. In order to provide a balanced overview of the whole field, this chapter only discusses a few typical examples of such systems.[1] In general, the following guidelines are used in selecting the review targets: the older approaches are selected based on their relevance, both to this study and whether the same principles are still being widely used by other researchers; the more recent systems are selected on the same basis, plus their contribution to the state of the art. According to the anaphoric relationships and the type of anaphors they target, the reviewed approaches are organized into two sections: the personal pronoun coreference resolution systems (Section 2.3) and the definite description anaphora resolution systems (Section 2.4). The few approaches that process both pronouns and definite descriptions for coreference (but not associative anaphora) are organized in Section 2.3. In addition, there are also some specialized systems that do not belong to either category. They will be discussed in the ensuing chapters as appropriate.

2.1 An Anatomy of the Anaphora Resolution Task

Anaphora resolution is not a monolithic process but rather a collection of well-defined subtasks as outlined below[2]:

Text Modeling

Text modeling is the process of identifying the basic elements (eg. words and sentences) of the input text and the various syntactic and semantic properties associated with them. Much of the process is usually assisted by a syntactic parser.

[1] Mitkov's (2002) book offers a very comprehensive overview of coreference resolution systems.

[2] A generic algorithm for anaphora resolution, GENERIC-RESOLVE, is also proposed by Ng (2003).

Anaphor Identification

In this process, systems identify the anaphoric expressions they target. This can be as simple as selecting all expressions pertaining to a specific syntactic category, such as pronouns (for a pronominal coreference resolution system) or definite descriptions. Many systems also include an additional process called anaphoricity determination to filter out non-anaphoric expressions belonging to the same syntactic categories.

Antecedent Determination

The antecedent determination process is applied to each anaphor to associate it with its antecedent. It typically involves identifying the candidates for the antecedent and subsequently selecting the best match among them. What should be included as a possible antecedent is largely based on the system's scope. For a pronominal coreference resolution system that does not consider abstract entities, it could be the entire collection of nominal expressions in the article, up to and including the sentence containing the anaphor. For practical reasons, systems usually include additional syntactic constraints, such as number and gender matching, and define a window of sentences to limit the number of possible antecedents.

Virtually all anaphora resolution systems follow the same meta-procedures while substantiating them differently. Since antecedent determination is the most important step in anaphora resolution, it is also the place where the key differences among the systems are often found.

2.2 Common Constraints and Preferences for Antecedent Determination

In coreference resolution, various forms of guidance are available in determining the degree of match between an anaphor and a possible antecedent. The rarely-violated rules are commonly used as constraints to eliminate false candidates, and the rest are often used as preferences to facilitate selection of the best match. Below is a list of some constraints and preferences commonly used in coreference resolution:

Number, Gender, and Person Agreement

An anaphor and its antecedent usually match in number, i.e. if an anaphor is in singular or plural form, the antecedent is expected to be in the same form. A pronominal anaphor usually also agrees with its antecedent on gender. This rule works best on antecedents for which the gender information is readily available. If an antecedent does not carry gender information, it cannot always be safely eliminated on the basis that the pronominal anaphor does so. Consider the following example in which the pronoun *he* introduces the gender as new information into the discourse:

(2.1) Also, Big Board Chairman Phelan said *he* would support SEC halts of program trading during market crises but not any revival of a "collar" on trading. WSJ 178:9

In addition, the pronoun *it* usually cannot be used to refer to a human, with the exception of a baby. The reverse may not be true, though, since countries and ships, among other things, can be referred to using gender-marked pronouns.

Binding Requirements

Chomsky's binding theory provides another set of syntactic rules for intrasentential nominal anaphora. At the core of the theory are the three principles of binding, which put constraints on the referents of reflexive pronouns, personal pronouns, and definite descriptions, respectively. As Chomsky (1993, page 188) puts it, the principles are:

A. An anaphor is bound in its governing category

B. A pronominal is free in its governing category

C. An R-expression is free[3]

The rigorous definitions of the principles depend on a number of related concepts. Instead of repeating all the details, the effects of the three principles can be (very) roughly described as follows:

A. The antecedent of a reflexive pronoun or reciprocal expression (e.g. *each other*) can usually be obtained by traveling up the parse tree, searching for the closest clause or noun phrase that has a subject. If the match is successful, the subject is its antecedent.

B. Exactly to the opposite of reflexives, a personal pronoun cannot corefer to any entity residing within the clause or noun phrase as identified using the previous rule.

C. A noun phrase cannot be considered as coreferential with a definite description anaphor if its parent phrase also contains the anaphor.

Centering

Centering (Grosz et al., 1995) is a theory concerning local discourse coherence. The basic idea of the theory is that in order to have a coherent discourse, at least one of the entities mentioned in a successive utterance[4] should have already been mentioned in the preceding one. The notions of 'forward-looking center' and 'backward-looking center', both of which refer to discourse elements, are used to facilitate formalizing this idea – each utterance U_n in a discourse has a ranked list of forward-looking centers $C_f(U_n) = \{C_f^1(U_n) \dots C_f^k(U_n)\}$; with the exception of the initial one, each utterance also has a unique backward-looking center $C_b(U_n)$, which must realize the highest-possible element of $C_f(U_{n-1})$. The definition of C_b implies that $C_b(U_n)$ is the highest-ranked element of $C_f(U_{n-1})$ that is realized in U_n (Brennan et al., 1987).

Grosz et al. identifies three types of transitions between utterances and orders them according to their relation to the perceived cohesiveness of the discourse:

[3] A more verbose version is available from Reinhart (1981): 'A given NP cannot be interpreted as coreferential with a distinct nonpronoun in its c-command domain'.

[4] The term 'utterance' is used in place of 'sentence' to stress that the theory operates in the domain of discourse, not sentences in isolation.

Center Continuation: $C_b(U_n) = C_b(U_{n-1}) = C_f^1(U_n)$[5]

Center Retaining: $C_b(U_n) = C_b(U_{n-1}) \neq C_f^1(U_n)$

Center Shifting: $C_b(U_n) \neq C_b(U_{n-1})$

In addition, Grosz et al. also identifies a constraint that prohibits the pronominalization of any member of $C_f(U_{n-1})$ in utterance U_n unless $C_b(U_n)$ is also realized with a pronoun[6]. In effect, this rule establishes the backward-looking center, which is often understood as the topic of the utterance, as the site that has the first priority for pronominalization.

One of the major issues with centering, as Poesio, Stevenson, et al. (2004a) suggested, is that some critical notions, including 'realization' and 'ranking', are only partially specified, resulting in a large number of competing 'instantiations' of the theory that attempt to solidify them. Poesio, Stevenson, et al.'s extensive investigation provides invaluable insights into the effects of these different parameterizations, and concludes that some of them allows the claims of centering to be statistically verified – provided that the claims are regarded as preferences, not constraints (Poesio, Stevenson, et al., 2004b).

Another issue particular to the theory's application to anaphora resolution is that it does not account for intrasentential anaphora. Since Walker (1989) raised the issue, most implementations included some kind of compensation for the problem. In addition, Kameyama (1998) proposed an intrasentential centering framework, which breaks each complex sentence into smaller units and uses clauses as utterances.

Selectional Restrictions

Selectional restriction refers to the semantic compatibility between a verb and an argument. In the context of coreference resolution[7], the semantic constraint that applies to the anaphor should remain valid after substituting it with the antecedent. The reverse should also apply if the anaphor is a definite description.

Recency

Recency measures the distance between an anaphor and a possible antecedent, expressed in number of words or sentences. A candidate that is closer to the anaphor usually receives higher preference.

Frequency of Mention

A frequently occurring word is often also central to the discourse and is in turn a more likely candidate of antecedent.

Grammatical Function

Keenan and Comrie (1977) proposed the accessability hierarchy (AH) of noun phrases on the basis of data from about fifty languages.

[5] C_f^1, the highest-ranked forward-looking center, is often called the preferred center, C_p.

[6] In an earlier paper (Grosz et al., 1983), a different version of the rule is proposed: if $C_b(U_n) = C_b(U_{n-1})$, $C_b(U_n)$ should be realized using a pronoun.

[7] Selectional restriction is not as useful for coreference to abstract entities and associative anaphora

Subject > Object > Indirect Object > Oblique > Genitives > Object of Comparison

Many anaphora resolution systems use similar lists, either directly as part of the salience metrics (e.g. a potential antecedent at subject position is more salient than one that is at object position) or as the ranking parameter in centering.

2.3 Personal Pronoun Coreference Resolution Systems

The pronominal coreference resolution systems surveyed in this section are organized based on their approach to antecedent selection. More specifically, the categorization depend on both a system's decision factors and its knowledge source, which can be obtained by answering the following two questions:

- What information does the system use to determine the antecedent?
- How does it acquire the needed information?

As the rest of the section illustrates, coreference resolution systems have evolved along both lines. On the one hand, the decision process becomes increasingly more complex as more factors are exploited; on the other hand, however, the effort needed to build the decision process has, to some extent, decreased, due to the introduction of machine-learning approaches. Another trend in the field is that while earlier systems usually focus on pronouns only, recent systems, especially the machine-learning based ones, tend to cover for both pronouns and definite descriptions.

2.3.1 Naive Approaches

The naive approaches are characterized by their relative 'simplemindedness' in that they do not exploit many knowledge sources. These approaches follow a major principle in identifying potential antecedents, and reject invalid ones based on syntactic constraints.

Hobbs' Naive Algorithm

Hobbs' (1976, 1978) naive algorithm remains one of the most influential pronoun resolution systems. On the one hand, its performance is comparable to many modern systems (Mitkov & Hallett, 2007); on the other hand, it only requires syntactic information and is relatively easy to implement, making it an ideal benchmark for new systems. The basic strategy of the algorithm is to perform left-to-right, breadth-first search on the parse tree, beginning from the clause or noun phrase that immediately covers the anaphor and gradually expand the search scope. In the case that no antecedent can be found, the algorithm searches the preceding sentences one by one using the top nodes as start points.

Implicitly, the naive algorithm employs both the recency and the grammatical function preferences – it starts from the vicinity of the anaphor, and the left-to-right search order coincides with the accessability hierarchy of English. The algorithm observes the binding requirements, albeit not to the fullest extent. Hobbs argues that incorporating all binding requirements would greatly complicate the program while gaining little improvement in real-world performance.

Hobbs reported an overall accuracy of 88.3% on a mixed corpus of technical writing, fiction, and news wire. When additional selectional restrictions are manually applied, the figure increases to 91.7%.

Despite its effectiveness, there is a particular issue about the naive algorithm that concerned many researchers (Ge et al., 1998; Mitkov & Hallett, 2007, inter alia) – the algorithm expects an $\bar{\mathrm{N}}$ node under NP, which is not compatible to the commonly used Penn Treebank style (Bies et al., 1995). The problem is not merely about style incompatibility since producing the $\bar{\mathrm{N}}$ requires classifying any attached prepositional phrase as either an argument or an adjunct to the head noun. However, a recent study by Merlo and Ferrer (2006) indicates that this distinction can be made quite reliably based on the preposition alone, therefore clearing up the barrier to the algorithm's application in Treebank-style corpora and parser output.

Centering Algorithm of Brennan, Friedman, and Pollard

Brennan et al. (1987) extended an earlier version of the centering theory and provided a solid algorithm for pronominal coreference resolution, which is often referred to as the BFP algorithm. The BFP algorithm applies the same criterion that is used in the original centering theory to distinguish center continuation and center retaining to the center shifting category, resulting in a four-way classification of center transition – Continuation, Retaining, Shifting-1, and Shifting (redefined). The new categories, Shifting-1 and Shifting, are defined as follows:

Shifting-1: $C_b(U_n) \neq C_b(U_{n-1}), C_b(U_n) = C_f^1(U_n)$
Shifting: $C_b(U_n) \neq C_b(U_{n-1}), C_b(U_n) \neq C_f^1(U_n)$

The algorithm ranks the forward-looking centers by their grammatical function. In addition, it also exploits syntactic constraints such as gender/number agreement and the binding requirements.

Walker (1989) performed a manual evaluation of the algorithm and compared it with Hobbs' naive algorithm. On the small corpus of 281 pronouns covering fiction, news wire, and dialogue, the BFP algorithm achieves an overall accuracy of 77.6%. The overall accuracy of the naive algorithm (81.1%) is slightly higher, but the difference is not statistically significant[8]. After a detailed error analysis, Walker discovered that the BFP algorithm's lower performance can be mainly attributed to its lack of intrasentential anaphora resolution mechanism. This issue is not specific to BFP but has its root in the centering theory. However, as Walker also noted, centering is a theory of discourse and may be held responsible for intrasentential anaphora either. In light of the problem, she proposed to enhance the BFP algorithm by adopting an additional rule from Carter (1987) that in general favors intrasentential antecedents over those from a previous utterance. Performance of the modified

[8] Unless otherwise noted, $\alpha = 0.05$.

approach is significantly higher than the original one, registering an increase of more than 8% in overall accuracy.

Left-Right Centering

Tetreault's (1999, 2001) Left-Right Centering (LRC) algorithm is a more recent centering-based approach. It uses the same left-to-right, breadth-first search strategy used by Hobbs' (1978) naive algorithm to produce the ranked list of forward-looking centers. The algorithm also searches the current utterance first before resorting to preceding ones, however, the strategy is not detailed. Evaluated on a subset of the WSJ corpus containing 1696 pronouns, the algorithm achieves an overall accuracy of 72.4%, slightly inferior to that of the Hobbs' naive algorithm (72.8%). The difference is not statistically important.

2.3.2 Factor-based Approaches

Factor-based approaches are characterized by their extensive use of heuristics for the antecedent determination task. Instead of following one major principle, these approaches resort to the emergent behavior resulted from the interaction of the weighted factors.

Lappin and Leass' RAP Algorithm

Like Hobbs' (1978) naive algorithm, Lappin and Leass' (1994) Resolution of Anaphora Procedure (RAP) is also an often-cited and well-studied algorithm. At the core of the RAP algorithm is a list of seven weighted pragmatic preferences that is used to compute the salience scores of potential antecedents. Among the most-highly weighted preferences are sentence recency (weighted 100) and grammatical functions (subject = 80, indirect object = oblique = 40); the rest of the preferences apply to specific syntactic constructs that the authors believe as more likely (or less likely) sites for antecedents, such as a head noun. Each of the potential antecedents is evaluated against all seven preferences and receives a weighted-sum of the matching preferences as its salience, which is subject to further modifications of additional local rules as well as degradation. The system recognizes and chains existing coreferential relationships into equivalence classes, which are treated as 'super' entities that unify the features of all chained elements. When one of the chained elements is presented to the weighting procedure, the unified features are evaluated instead. In addition, the system also employs a number of constraints, including number/gender agreement, binding requirements, and a built-in pleonastic pronoun detection module. Candidates masked by these constraints are not considered.

The RAP algorithm was evaluated on a corpus of computer manuals containing 360 third-person pronouns with an overall accuracy of 86%. A modified Hobbs' naive algorithm was also evaluated on the same corpus, with a slightly lower overall accuracy of 82%. It has to be noted, however, that the corpus was heavily filtered. In a more recent evaluation, Mitkov and Hallett (2007) reported that, without pleonastic pronoun detection, the RAP algorithm's performance is similar to that of the Hobbs' naive algorithm, and the difference is not statistically significant.

While the original RAP algorithm operates on full parse trees, Kennedy and Boguraev (1996) proposed an extension to the algorithm that releases it from this dependency. The extended algorithm operates on part-of-speech tags and grammatical function labels alone. In addition to the modifications need to accommodate the different form of input, the extended algorithm also introduces two new preferences and gives a reduced initial weight to oblique complements. Kennedy and Boguraev (1996) reported an overall accuracy of 75% on a corpus (306 third-person pronouns) containing a great variety of documents.

Mitkov's Knowledge-poor Algorithm

Mitkov's Knowledge-poor algorithm has its preferences expressed as a list of weighted antecedent indicators. Unlike the RAP algorithm, the preferences can be explicitly specified as boosting or impeding. In addition, the list also includes semantic information embedded in a list of indicating verbs. As a potential antecedent is evaluated, each indicator is matched against the antecedent and the matching ones add their weight to the candidate's score. The aggregated score of each candidate is then compared to decide the algorithm's pick. In addition to the antecedent indicators, the system also employs gender and number agreement as a constraint. What is particularly remarkable about the algorithm is its simpleness: The only external tools it needs is a part-of-speech tagger and an noun phrase extractor. The simple design also contributes significantly to its portability. The system has been adapted to Polish, Arabic, and French with good results, making it one of the few cross-language coreference resolution systems available to date.

The original algorithm is subsequently extended (Mitkov et al., 2002) so that it could operate in a fully automatic manner. Other major additions include a parser, three new indicators and a pleonastic pronoun classifier are also incorporated into the system. Evaluated on a set of technical manuals containing 2,263 anaphoric pronouns, the updated algorithm achieved an overall accuracy of 59.35% when the pleonastic pronoun classifier is in use. Turning it off improves the overall accuracy slightly (61.82%).

2.3.3 Machine-learning Approaches

Machine-learning approaches are characterized by their ability to automatically tune the parameters used for the decision process. As researchers begin to explore more heuristics, the number of parameters needing adjustment also increased. Machine-learning provides two models to solve the problem. The supervised model expects to be explicitly instructed what it should learn, which is usually provided in the form of annotated training data. The unsupervised model does not require training data but rather depends on cues embedded in its design.

Soon et al.'s Decision Tree Algorithm

Soon et al.'s (2001) decision tree algorithm uses 12 domain-independent features, including recency (measured in number of sentences), string matching of head nouns, alias resolution for named entities, semantic class (person/organization/location etc., obtained through WORDNET) compatibility, gender and number agreement, part-of-speech tags, and other syntactic features. A training set containing both positive and negative instances is drawn from manually annotated text. A positive instance contains an anaphor and its closest antecedent, the entities between the correct antecedent and the anaphor are used to form the negative instances. Once trained, the decision tree can estimate the likelihood of coreference given a pair of anaphor and potential antecedent. Soon et al.'s (2001) algorithm limits the search scope by setting up a threshold – the algorithm evaluates each potential antecedent, starting from the anaphor and scans towards the beginning of the text; the first candidate that have a likelihood above the threshold is chosen.

The system covers both pronominal and definite description coreference resolution. Evaluation performed on the MUC-6 and MUC-7 data sets shows that its performance (F-measures of 62.6% and 60.4% respectively) is on par with other manually-designed systems.

Many researchers proposed extensions to the original Soon et al. algorithm. For example, Ng and Cardie's (2002b) extension includes an additional 18 (mostly syntactical) features manually selected after evaluating a total of 41 feature additions. Ponzetto and Strube's (2006) extension focuses on expanding semantic features by incorporating additional knowledge from WORDNET and WIKIPEDIA and adding thematic role features. Ng's (2007) extension replaces the original semantic class marker with one that is based on machine-learning. These extensions typically achieve a less than 10% increase in F-measure.

Poon and Domingos' Markov Logic Network Algorithm

A new trend in coreference resolution is the rapid improvement of unsupervised machine-learning approaches. According to Poon and Domingos (2008), their system set up a new record for coreference resolution on the MUC-6 corpus, an F-measure of 79.2% for all nominal expressions. Poon and Domingos' system is based on the Markov Logic Network (MLN, Richardson & Domingos, 2006), which is a weighted first-order knowledge base serving as template to create Markov networks. The system adopts a cluster-based model[9], considering all mentions of the same entity as a cluster rather than comparing pairs of antecedents and anaphors. This design allows it to leverage the results obtained from 'easy' cases to help resolving the 'hard' ones. Another interesting feature of the system is that it does not need a separate process to determine whether a mention is anaphoric (which is necessary in order to propose it as an anaphor). The set of features incorporated in the system is rather simple, including gender and number agreement, head word determination, distance-based

[9] This is similar to Lappin and Leass' (1994) equivalence class. Cardie and Wagstaff (1999) have also used the same concept in an unsupervised approach that treats coreference resolution explicitly as a clustering task.

salience measure, as well as apposition and predicate nominal relationships.

2.3.4 Semantics-centric Systems

It has been recognized since the inception of computational treatment to anaphora resolution that semantics plays an important role. However, progress in this area has been relatively slow. To date, there are only a handful of fully automatic systems that emphasize the role of semantics in their implementation.

Dagan and Itai's Co-occurrence Statistics

Although many earlier systems (e.g. Carbonell & Brown, 1988; Rich & LuperFoy, 1988, inter alia) incorporate semantic knowledge in their decision process, Dagan and Itai's (1990) approach is one of the earliest to automatically acquire such knowledge and apply it in the domain of coreference resolution. The basic idea behind their approach is that co-occurrence patterns reflect regularized or canonical structures of a language (Grishman et al., 1986). The approach deviates from the common view that selectional restrictions should be based on semantic classes but rather apply them directly to words. Specifically, the approach collects statistics from a large corpus on tuples $\langle anchor, mention\rangle$, where $anchor = \langle lemma, function\rangle$ is the combination of a lemma functioning as either a verb or an adjective, and a grammatical function of `subject-verb`, `verb-object`, or `adjective-noun`. The collected information is used to approximate selectional restrictions. For example, the sentence 'The corrupt government collected the money.' instantiates three tuples:

- $\langle\langle collect, \texttt{subject-verb}\rangle, government\rangle$,
- $\langle\langle collect, \texttt{verb-object}\rangle, money\rangle$, and
- $\langle\langle corrupt, \texttt{adjective-noun}\rangle, government\rangle$

The system counts the number of occurrences of each usage tuple, and uses a threshold to determine its validity. The initial experiment uses statistics collected from 28 million words and a threshold of 5 occurrences. Out of the 59 instances[10] of *it* examined, 38 are covered by the system and 33 are correctly resolved.

A later extension (Dagan et al., 1995) introduces a 'normalized' statistic, $stat(C)$, allowing comparison to be made between different tuples. The statistic is modeled as the conditional probability of a tuple $\langle anchor, C\rangle$, given the mention C:

$$stat(C) = P(\langle anchor, C\rangle | C) = freq(\langle anchor, C\rangle)/freq(C)$$

The new statistic is incorporated into RAPSTAT, a direct extension to Lappin and Leass' (1994) RAP system. RAPSTAT examines the salience scores produced by RAP and overrides its decision if the lexical statistics strongly suggest otherwise. Results from a blind test show a moderate 2.5% increase

[10] The instances are randomly selected and filtered so that each instance has both its antecedent and at least one competing alternative in the same sentence

in overall accuracy with the addition of co-occurrence. However the difference is not statistically significant, partly due to the very small data set (cf. Section 2.3.2 for more details on the data set).

Kehler et al. (2004) dismiss the usefulness of co-occurrence statistics, describing them as ‘a poor substitute for world knowledge’ after experimenting with a supervised machine learning system and a slightly different set of patterns, `subject-verb`, `verb-object`, and `possessive-noun`. Their results show that incorporating statistics produced by such patterns provides no visible improvement to the system’s performance.

However, a subsequent study by Yang et al. (2005) seems to indicate otherwise. Their approach is similar to that of Kehler et al. (2004) – it uses the same set of patterns and also gathers statistics from the web – only more elaborate in the way the patterns are realized. Instead of translating a pattern instance literally to web queries, it is expanded both semantically and syntactically: Named entities are replaced by their semantic classes, nouns and verbs are expanded to include different inflected forms, and a combination of both definite and indefinite articles are used. Evaluations performed on two different supervised machine learning systems both show overall accuracy increase of around 3% when co-occurrence statistics are in use. Unfortunately, it is not clear from the paper whether this increase is statistically significant. Aside from the overall accuracy increase, there are also other interesting findings related to Yang et al.’s (2005) results. First, their results show an obvious contrast between the effectiveness of co-occurrence statistics on the gender-neutral pronouns and the ones marked with gender – the latter receive no visible benefits. Second, they show that co-occurrence statistics gathered from large corpus consisting 76 million words is consistently less helpful (by a small margin of around 2%) than those gathered from the web. Even in absence of statistical tests, this can at least indicate that the web is a source as good as a corpus when it comes to collecting co-occurrence patterns.

Bean and Riloff’s Contextual Role Knowledge

Bean and Riloff’s (2004) BABAR (see also Bean, 2004) is a supervised machine-learning based system. The system is domain-specific: it populates its knowledge bases from training materials of a certain domain and applies the knowledge to facilitate nominal coreference resolution for articles pertaining to the same domain. Similar to the systems developed by Dagan and Itai (1990) and Yang et al. (2005), BABAR focuses on co-occurrence patterns and their expansion using semantic classes. However it goes one step further to exploit the semantic correlation between a role and its activities under the context of a specific domain. In other words, the system not only identifies co-occurrence relationships between nominal expressions and verbs/adjectives but also identifies the co-occurrence of such relationships, in the sense that the nominal expressions in these relationships are coreferential.

In this approach, co-occurrence relationships are represented with ‘caseframe’ tuples, which are conceptually similar to Dagan and Itai's (1990) co-occurrence patterns but more detailed[11]. One of the major advantages of the caseframe patterns is the use of thematic roles 'agent' and 'patient' instead of the syntactic subject and direct object. This allows the cases to be grouped along the semantic axis, unaffected by passive voice clauses. Another important improvement offered by the approach is that in addition to the ordinary `<agent> verb`, it also introduces a special pattern, `<agent> verb dobj`[12], to capture the agent-verb relationship in the context of a specific patient (dobj). Instances of this pattern offer much more fine-grained selectional restrictions, and are especially useful for verbs with large variety of senses.

The learning process begins by scanning the training data to generate caseframes. Once the caseframe instances are produced, the likelihood ratio χ^2 statistic of the captured head noun (agent/patient) co-occurring with the caseframe terms (verb/adjective/direct object) is then calculated for each unique caseframe. The probability associated with the calculated value is used as the confidence of the caseframe. After discarding the low-confidence caseframes, the remaining ones are organized in a knowledge base of 'lexical expectations' (CFLEX). Another knowledge base, CFSEM, is constructed in a similar manner, but with semantic classes in place of the captured head nouns.

In order to establish the relatedness of the caseframes themselves, the system performs a high-precision coreference resolution on the training corpus, guided by only a handful of 'reliable' lexical and syntactic heuristics, such as identical proper names, reflexives, and simple appositions. If two caseframes both involve the same entity, they are considered potentially related. Again, the counts are subjected to statistical analysis to determine the strength of pair-wise correlations. The evaluated correlations are captured in the CFNET knowledge base.

BARBA combines the three knowledge bases with other general knowledge sources (e.g. lexical, syntactical etc) to make the final decision. Evaluation on the impact of different knowledge sources shows that the CFNET relations have the highest impact on pronoun resolution, increasing recall by around 15%. However, as Bean (2004, Section 7.2) noted, the performance gain may not apply to domain-general corpora, since the acquired contextual role relationships would be 'thinly spread'.

Bergsma and Lin's Path-based Approach

Bergsma and Lin's (2006) approach deviates from the traditional wisdom of selectional restriction modeling. Instead, it focuses on identifying 'coreferent paths' and 'non-coreferent paths', which are dependency paths that usually lead to coreferential/non-coreferential mentions on the two ends. A typical example of the coreferent paths is `noun` *`lost`* `pronoun's` *`job`*[13], which predicts that the two expressions occupying the `noun` and the `pronoun` slots refer to the same real-world entity. In the simplest version of the approach, the paths are learnt by scanning a large corpus for dependency

[11] Neither Bean and Riloff's paper nor Bean's thesis clearly specifies the full set of patterns being used. Therefore it is not possible to know the exact coverage of the patterns. However, from the examples provided and the more detailed descriptions by Riloff (1996), their patterns provide significantly more coverage than those offered by Dagan and Itai do.

[12] `dobj`: direct object. This pattern works for active voice clauses only.

[13] On the dependency structure, the `pronoun` is located below *job*. Also note that `noun` and `pronoun` are not parts of the path.

paths that have pronouns attached to both ends. A path instance is marked as likely coreferent if the two pronouns are the same (including inflected forms); otherwise it is marked as non-coreferent[14]. The final decision of a path's status is obtained by comparing the number of likely-coreferent instances with the total amount of instances – Paths with high ratios are classified as coreferent paths; those with low ratios are classified as non-coreferent paths.

Aside from the coreferent path, Bergsma and Lin's (2006) SVM-based system also features semantic compatibility and a probabilistic gender/number information[15], among others. In an evaluation performed on the MUC-7 test data set, the system achieves an accuracy of 71.6% over the third-person pronouns with nominal antecedents. It is also worth noting that in evaluations performed on two larger data sets, the performance gains contributed by both the coreferent path feature and the semantic compatibility feature are statistically significant.

2.4 Definite Description Anaphora Resolution Systems

Although it is gradually receiving more coverage, the number of systems that address definite description coreference is still relatively small. Even fewer systems tackle the issue of associative anaphora, which is a phenomenon primarily observed in definite descriptions. As mentioned in the earlier section, more recent, machine-learning based systems (e.g. those of Soon et al., 2001; Haghighi & Klein, 2007; Poon & Domingos, 2008; Ng, 2008, inter alia) tend to cover both personal pronouns and definite descriptions for coreference resolution. These systems usually consider definite descriptions as 'augmented' pronouns with additional features, thereby giving uniform treatments to the coreference resolution problem.

Vieira and Poesio's Empirically-based System

The study by Vieira and Poesio (2000) is one of the few studies that focus on definite descriptions. Their approach offers a reasoning-free and domain-general treatment of both coreferential and associative definite descriptions. Based on their earlier corpus-based investigations of definite description uses (cf. Vieira, 1998; Poesio & Vieira, 1998), Vieira and Poesio classify the definite descriptions into three broad categories: direct anaphora, bridging, and discourse-new. The first two categories differ in whether the anaphoric definite description shares the same head noun with its antecedent – when they do, the relationship is said to be direct, otherwise it is bridging. Compared to the categorization scheme established in Section A.3, the bridging category under this definition overlaps both coreferential and associative relationships. The discourse-new category, which contains non-anaphoric expressions, is established according the authors' observations that a large portion of the definite descriptions actually serve to introduce new entities to the discourse and are therefore not anaphoric.

[14] This method leverages the fact that same pronouns in a sentence refer to the same entity more often than not.

[15] Which is also boosted using coreferent paths, cf. their paper for more details.

At the core Vieira and Poesio's system is a manually-built decision tree with various heuristics as parameters. For anaphoricity determination, the identified heuristics include a list of 'special predicates', the presence of restrictive modification, appositive and copular constructions, the presence of proper noun as pre-modifier, and time-denoting expressions. The bridging relationships are captured by searching the manually constructed WORDNET knowledge base (see also Poesio et al., 1997). If the head nouns of two expressions are identified as synonyms, or constitute a meronym-holonym or hypernym-hyponym relationship, or share the same hypernym, the expressions are considered as having a bridging relationship. In addition, when one or both of the expressions are compound nouns, the system also looks for possible matchings between the head of one expression and the pre-modifier of the other, and between the two pre-modifiers (when both expressions are compound).

The system was developed and tested on two different subsets of the WSJ corpus. On the test data set, the system achieved an F-measure of 71% for direct anaphora, and an overall accuracy of 28% for the four types of bridging descriptions. The anaphoricity determination (discourse-new description identification) task was performed with an F-measure of 70%.

2.4.1 Anaphoricity Determination

As indicated by Poesio and Vieira (1998), around half of all definite descriptions in their study are non-anaphoric. This ratio highlights the need to include some form of anaphoricity determination in systems that process definite descriptions. Aside from Vieira and Poesio's (2000) rule-based approach and some recent unsupervised machine-learning approaches that implicitly handle the problem (e.g. Denis & Baldridge, 2007; Poon & Domingos, 2008), there are a few other studies that specifically target this area.

Bean and Riloff's Corpus-based Approach

Bean and Riloff (1999) describe a corpus-based approach that automatically gathers non-anaphoric definite descriptions that lack syntactic markers. The expressions are gathered from three sources – the expressions in the first sentence of each article (S1 extractions), the expressions that match the patterns generalized from the S1 extractions, and the expressions that are consistently definite (i.e. the head noun is always associated with the definite article *the*) throughout the corpus. Combining the list with with syntactic heuristics, the system achieves an F-measure of about 81% on a subset of the MUC-4 corpus. The methods are also incorporated into the authors' BARBAR (Bean & Riloff, 2004) anaphora resolution system

Uryupina's Web-based Approach

Instead of collecting statistics from a limited corpus, Uryupina (2003) uses the web to obtain the 'definite probability' of expressions. For each expression a maximum of six queries are submitted to search engine and the result page counts are used to obtain four ratios – (# "*the Y*")/(# *Y*),

(# "*the Y*")$/$(# "*a Y*"), (# "*the H*")$/$(# *H*), (# "*the H*")$/$(# "*a H*"), where *Y* and *H* denote respectively the original expression without determiner and the head noun of the expression. Aside from the web-based statistics, syntactic heuristics and recency (distance to the previous phrase sharing the same head) are also considered during the classification. Uryupina uses a four-way classification scheme, categorizing each noun phrase based on whether it is discourse-new ($\pm$*discourse_new*) and whether it is specific enough to uniquely identify an entity without referring to an antecedent ($\pm$*unique*). When evaluated on the MUC-7 corpus, the system achieves an F-measure of 83.5% in distinguishing $+$*discourse_new* and $-$*discourse_new* definite descriptions. An F-measure of 91.8% is recorded for the $\pm$*unique* subcategory. In both cases the 'definite probability' features contribute a modest but statistically significant increase to the results. Poesio et al. (2005) combine the 'definite probability' features with some additional lexical, positional, and syntactic features using a neural network. They observe an around 3% increase in F-measure (statistically significant at $\alpha = 0.1$) when the anaphoricity determination module is added to an existing anaphora resolution system.

Ng and Cardie's Decision Tree Algorithm

Ng and Cardie (2002a) use a set of 37 features, most of which are grammatical but the set also covers lexical information such as whether the head of the phrase appears in a preceding phrase, positional information such as whether the phrase is in the first sentence of the article, and semantic information such as whether there exists a preceding phrase that forms an ancestor-descendent relationship in the WORDNET with the phrase in question. The system achieves accuracies of 86% and 84% on the MUC-6 and MUC-7 evaluation data sets, using decision trees automatically induced from the respective dry-run data sets. However, further experiments integrating the anaphoricity module with a coreference resolution system show that the coreference resolution task benefits most from a high-precision anaphoricity determination module rather than one that has lower precision but higher recall. A modified version of the module, which is more accurate in determining non-anaphoric instances (around 90%), is shown to increase the overall system F-measure by 2-3%. In a later systematic evaluation on how should anaphoricity determination be integrated into the coreference resolution task, Ng (2004) re-affirms that the anaphoricity determination module should be relatively conservative in classifying an expression as non-anaphoric.

2.4.2 Associative Anaphora Resolution Systems

As general world knowledge plays an essential role in associative anaphora resolution, existing approaches usually focus on how to effectively exploit one or more knowledge sources to obtain the knowledge needed to interpret associative relationships. The three most commonly used knowledge sources are the WORDNET, manually compiled corpora, and the web.

Additional Studies by Poesio et al.

Aside from the WordNet-based approach (Poesio et al., 1997; Vieira & Poesio, 2000, see earlier discussion), other attempts of Poesio et al. include lexical clustering (Poesio et al., 1998), syntactic construction statistics (Poesio et al., 2002), and a multi-source approach realized using a neural network (Poesio, Mehta, et al., 2004).

In the lexical clustering approach, phrases are represented using vectors consisting the words surrounding the head of the phrase, with each word carrying a weight inverse-proportional to its distance towards the head. The degree of match between an potential anaphor and a candidate for antecedent is then calculated using one of the vector distance metrics. After comparing with the Manhattan distance and the Euclidean distance, the authors have chosen the cosine of the vectors' angle. The system does not perform as well as the WordNet-based approach for cases that latter is capable of processing (i.e. where the anaphor and the antecedent have a synonym/hypernym/meronym relationship). However, it provides limited coverage for the rest of the categories and therefore has a higher overall performance.

The syntactic construction statistics approach specifically targets the mereological cases, a class proven difficult for both the WordNet-based and the lexical clustering approaches. Poesio et al. (2002) identify four syntactic templates that are potentially relevant to the class: *the* NP *of* NP, NP *of* NP, NP*'s* NP, and NP N, the first three of which are eventually determined as useful. The system collects instance counts of the patterns for each noun from the training corpus and employs the mutual information statistic $I(x;y) = log\frac{P(x,y)}{P(x)P(y)}$ as the measure of likelihood for a given pair of anaphor and candidate for antecedent. The system successfully recognizes 8 out of 12 mereological cases, much higher than the WordNet-based approach (3/12) and the lexical clustering approach (2/12) do.

The neural network based approach, also targeting the mereological cases, combines lexical distance information obtained from either the WordNet or the web with two additional salience features, the utterance distance and whether the potential antecedent appears as the first-mentioned entity in its sentence. The WordNet distance is obtained by iterating through all word senses of the anaphor and the antecedent and finding the minimum distance between the two via a common hypernym node. The web distance is calculated using the page counts returned by the queries "*the* NP *of the* NP" or "*the* NP *of* NP". The inclusion of the first-mentioned feature reflects Poesio's (2003) earlier findings that anaphors in a bridging relationship is significantly more likely to find their antecedents in the first-mentioned position of the preceding sentence. The system is evaluated using 10-fold cross validation on two data sets, one balanced (1:1) and the other with three negative instances for one positive (1:3). Results obtained using the WordNet are similar to those obtained using the web, however the web-based method scales better as the ratio of negative instances increases. Switching from the balanced data set to the 1:3 data set causes an increase in precision figures obtained using both instance metrics, accompanied by a much larger decrease of the recall figures, and hence lower overall F-measure figures. In the absence of further experiments, it is hard to estimate the system's real-world performance.

Bunescu's Web-based Approach

The approach of Bunescu (2003) is purely web-based. It employs a simple query pattern that provides a broad coverage – "$\mathtt{n}_t$. *The* $\mathtt{n}_a$ `Verb`", where $\mathtt{n}_t$ is the head noun of the potential antecedent (the trigger term), $\mathtt{n}_a$ is the head noun of the anaphor, and `Verb` denotes the collection of inflected copula and modal verbs {*is/are, was/were, has/have, had, may, might, can, could, should, would*}. This case-sensitive pattern expects fragments from two adjoining sentences, the first of which ends with the trigger term (note the period following $\mathtt{n}_t$) and the ensuing one begins with the anaphor, followed by any of the verbs in the `Verb` list. The list of verbs are chosen because they have little semantics on their own and cannot serve to associate the two nouns. In addition to the main pattern, Bunescu (2003) also prepare queries for the individual components, "$\mathtt{n}_t$." and "*The* $\mathtt{n}_a$ `Verb`", and uses the mutual information metric to assess the strength of the association. Different preferences for precision and recall can be realized by varying the threshold of the mutual information values. Evaluation results of the system compare favorably to those obtained by Poesio et al. (1998): at the same level of recall, its precision is more than twice as much[16]. One of the drawbacks of Bunescu's approach is that major commercial search engines no longer support the kind of queries instantiated from the main pattern[17], which significantly reduces its practical value.

Fan et al.'s WordNet Semantic Path Search

Fan et al.'s (2005) approach aims at better utilizing the semantic information embedded in the WORDNET. The most significant contributors of the system's performance gain over its predecessors include its relaxed stopping criterion, the accessibility of the properties of a superclass (hypernym), and the greater maximum search depth. The system considers a search successful when a superclass or a subclass (hyponym) of the target expression is encountered. Evaluations reveal this relaxation makes the largest contribution, followed by property inheritance and search depth. Overall, the system achieves twice the recall as reported by Vieira and Poesio (2000) while maintaining the same precision. The main advantage of this approach over Bunescu's (2003) web-based one is that the obtained semantic path remains interpretable. On the other hand, it also offers significantly more coverage than previous WORDNET-based approaches.

[16] However, as noted by the author, the fact that the two systems use different evaluation data sets complicates the comparison.

[17] More specifically, the problematic areas are case-sensitive search and support for punctuation marks.

Chapter 3

Definiteness and Anaphora

Nominal anaphora have always been at the center of anaphora resolution research. However, the two major categories of anaphoric noun phrases, definite descriptions and pronouns, have received disproportionate attention: there is a large number of research on pronominal anaphora but relatively few dedicated to definite descriptions. Moreover, those researchers that consider definite descriptions often concentrate on MUC-style (Hirschman & Chinchor, 1997) coreference resolution tasks. Although there is considerable overlap between nominal anaphora and coreference, as noted by Deemter and Kibble (2000), the two notions are different in significant ways. Unlike pronouns, the majority of definite description uses are either 'discourse-new' in the sense that they do not have an antecedent in the discourse, or 'associative' in the sense that they are not coreferential with their antecedents. Pioneering empirical studies on definite description anaphora (e.g. Fraurud, 1990; Vieira, 1998) have revealed a number of difficulties that ultimately point to the essence of definiteness, which is still a subject of much debate among linguists and philosophers. On the other hand, the relatively vague notion of anaphora does not offer much help either, while researchers navigate through the myriad uses of definite descriptions.

Without appropriate understanding of the role definiteness plays in anaphora, it is very difficult to define the scope of definite description anaphora and to give adequate treatment for the various different uses. The primary focus of this chapter is to study the interaction between the two notions. Section 3.1 establishes the view of definite noun phrase anaphora as a device to satisfy the weak familiarity and informational uniqueness presuppositions of definiteness (Roberts, 2003), and investigates its relationship with some closely-related concepts. In Section 3.2, various uses of definite descriptions are examined from perspectives of familiarity and uniqueness to produce a categorization scheme of definite descriptions. Finally, the developed categorization scheme is applied to a part of the WSJ corpus and the results are presented in Section 3.3.

3.1 Basic Concepts Revisited

Given that anaphora covers such a broad and heterogenous set of linguistic phenomena[1], it is not surprising that there exists a number of different definitions for the notion of anaphora. Among the often-cited definitions are those of Halliday and Hasan (1976) and Hirst (1981), the latter of which will be discussed in details in Section 3.1.1. To some extent, both these accounts are developed from the utterer's perspective, i.e. anaphora as an instrument for discourse cohesion. There are also researchers (e.g. Carter, 1987; Deemter & Kibble, 2000; Denis, 2007) who opt for accounts centered around the need for 'interpretation', which is closer to the receiver's perspective (and hence that of an anaphora resolution system). Although this study prefers the latter, there is no fundamental conflict between the two perspectives. However, regardless which side they are on, the commonly-adopted notions of anaphora are too vague to provide sufficient guidance to the practice of anaphora resolution. The rest of the section will focus on clarifying the notion of anaphora for the subfield concerning nominal anaphors, which is the most active subfield of anaphora research and the topic of this study.

3.1.1 Definite Noun Phrase Anaphora

The aim of this section is to find a suitable view of anaphora for definite noun phrases. The first part of the section reviews Hirst's (1981) account of anaphora, which is one of the more elaborated among the often-cited definitions, and points out some of the inadequacies of its application to definite noun phrases as a whole. Following a brief overview of Roberts's (2003) analysis of definiteness, the third part of the section establishes the view of definite noun phrase anaphora as a device to satisfy the presuppositions of definiteness.

Hirst's Account of Anaphora

There are many different definitions of anaphora. For example, one may simply take the word literally (the act of referring back) or adopt the slightly more sophisticated view that an anaphor depends on its antecedent for its interpretation (Deemter & Kibble, 2000). In comparison, Hirst's (1981) definition[2] is more elaborate:

> ANAPHORA is the device of making in discourse an ABBREVIATED reference to some entity (or entities) in the expectation that the perceiver of the discourse will be able to disabbreviate the reference and thereby determine the identity of the entity. The reference is called an ANAPHOR, and the entity to which it refers is its REFERENT or ANTECEDENT. A reference and its referent are said to be COREFERENTIAL.

[1] Gathering from previous research such as those of Hirst (1981), Krahmer and Piwek (2000), and Mitkov (2002), it seems that although there lacks an absolute consensus as to which phenomena should be included in this set, it has a tendency of expansion.

[2] Hirst's (1981) definition may be seen as a more elaborate version of Halliday and Hasan's (1976), which describes anaphora (as paraphrased by Mitkov, 2002) as a cohesion which points back to some previous item.

> The process of determining the referent of an anaphor is called RESOLUTION. By ABBREVIATED, I mean containing fewer bits of disambiguating information (in the sense of Shannon & Weaver, 1949), rather than lexically or phonetically shorter.
>
> (Hirst, 1981, page 4)

Hirst's (1981) definition not only identifies the primary function of anaphora as a device of cohesion but also points out three characteristics of an anaphor: first, the anaphor points to a uniquely identifiable referent (or set of referents); second, the anaphor itself does not provide enough information about the identity of its referent; and third, it is expected that the receiver of the utterance is capable of filling this information gap.

There are a few obvious issues related to this definition. First, Hirst (1981) mainly concerns proforms in coreferential (i.e. the anaphor and its antecedent point to 'the same thing') relationships. In later discussion, I will nevertheless try to apply the definition to definite descriptions as well. Second, from the definition itself, it is not entirely clear what a 'referent' is – is it a discourse element, a real-world entity, or a mental representation of either? Reading some of the ensuing discussions, it seems that Hirst is referring to the last interpretation[3]. Similarly, the term 'abbreviated' is not clearly defined with regard to whether the anaphor should be compared to the discourse element that establishes the antecedent, or it simply means the information provided by the anaphor is not sufficient for locating the referent. I take it that comparison should be made between the anaphor and its intended antecedent in the discourse.

Regardless of the aforementioned issues, Hirst's (1981) definition can be used to explain a large number of anaphoric cases. To see how it works, consider example (1.1), repeated here as (3.1):

(3.1) Texas Instruments Japan Ltd.$_1$, a unit of Texas Instruments Inc., said *it*$_1$ opened a plant$_2$ in South Korea to manufacture control devices. *The new plant*$_2$, located in Chinchon about 60 miles from Seoul, will help meet increasing and diversifying demand for control products in South Korea, *the company*$_1$ said. *The plant*$_2$ will produce control devices used in motor vehicles and household appliances. WSJ 17:1-3

The example features two different chains of coreferential entities. In the first coreference chain, the proper name *Texas Instruments Japan Ltd.*$_1$ specifies a unique company, but the pronoun *it*$_1$ provides little semantics of its own, and the definite description *the company*$_1$ only contains type information. Interpretation of the second chain is not as straightforward since the antecedent *a plant*$_2$ and the anaphors *The new plant*$_2$ and *The plant*$_2$ share the same head noun, meaning that the antecedent itself does not provide more information about the identity of the referent than the anaphors do. This difficulty can be circumvented by taking the surrounding contexts into consideration. In this case, the antecedent is enriched by its context to the extent that it can be considered practically traceable to a unique real-world entity. In other words, it is not just 'a plant', but 'a plant belonging to Texas

[3] Hirst (1981, page 8) mentions that "for an anaphor to be resolvable, its antecedent must be in what we shall for the time being call the listener's 'CONSCIOUSNESS'."

Instruments Japan Ltd. that is (recently) opened in South Korea for the purpose of manufacturing control devices'. In contrast, the surrounding contexts of the two anaphors also provide additional information about the plant, but in neither case the information is sufficient to uniquely identify the referent.

While the addition of enrichment allows certain cases to fit into Hirst's (1981) framework, it has its limitations.

(3.2) In those old, old times, there lived two brothers who were not like other men, nor yet like those Mighty Ones who lived upon the mountain top. *They* were the sons of one of those Titans who had fought against Jupiter and been sent in chains to the strong prison-house of the Lower World.
The name of the elder of these brothers was Prometheus, or Forethought . . .
The younger was called Epimetheus, or Afterthought . . .
Baldwin (1895, The Story of Prometheus)

As illustrated by (3.2), the vague specifications of the two brothers provided by the first sentence does not preclude the use of the anaphoric *They* in the second, which also happens to contain more identifying information than the first one does. The true identity of the two brothers, Prometheus and Epimetheus, however, are not revealed until the following paragraph.

Examples like (3.2) pose real challenges to views of anaphora based on amount of disambiguating information, such as that of Hirst (1981). However, if it is not the need for disambiguation that links an anaphor and its antecedent, what could it be? Existing interpretation-based definitions do not provide an answer either. For example, Carter's (1987) account, which is also one of the more detailed definitions, simply states that the anaphor is "in isolation, somehow vague or incomplete, and can only be properly interpreted by considering the meanings of the other item(s)" (op. cit. Van Deemter, 1992). In order to obtain a satisfactory explanation, it is necessary to first examine the properties of definite descriptions and pronouns, which make up the majority of anaphoric cases.

Presuppositions of Definiteness

Definite descriptions are used both to refer to existing discourse entities and to introduce new entities into the discourse. The coreferential definite descriptions in example (3.1) are typical instances of the former use. The latter category is rather heterogeneous, consisting primarily of associative anaphors, deictic references, and the various 'unfamiliar' uses documented by Hawkins (1978), such as the definite descriptions in examples (3.3) through (3.5).

(3.3) I remember *the beginning of the war* very well . . . Hawkins (1978, ex. 3.83)

(3.4) Bill is amazed by *the fact that there is so much life on Earth.* Hawkins (1978, ex. 3.87)

(3.5) What's wrong with Bill?
Oh, *the woman he went out with last night* was nasty to him. Hawkins (1978, ex. 3.16)

The definite descriptions in examples (3.1) and (3.5) represent two distinct ends of definite description uses. In Prince's (1981, 1992) terms, the definite descriptions in (3.1) are both 'discourse-old' and 'hearer-old', while the one in (3.5) is both 'discourse-new' and 'hearer-new'. Most of the non-coreferential cases fell in between the two extremes. For example, *the beginning* in example (3.3) is arguably 'hearer-old' while being 'discourse-new', since it is generally understood that an event such as a war has a beginning. Similarly, many deictic references are also 'hearer-old' since the receiver is already aware of the referent, as shown in example (3.6).

(3.6) Pass me the water, please. Hawkins (1978, ex. 3.41)

The broad spectrum of definite description uses has triggered different interpretations of the essence of definiteness. As Poesio and Vieira (1998) noted, previous studies on the semantics of definiteness have yielded two competing views – some researchers (e.g. Russell, 1905) believe that uniqueness is the defining property of definiteness, while others (e.g. Heim, 1982) believe it presupposes familiarity. In addition, there are also researchers (e.g. Birner & Ward, 1994) who argue that neither theory provides adequate coverage. A recent study by Roberts (2003) provides a unified account of the two different views under the term 'informational uniqueness'. According to Roberts, definiteness presupposes both familiarity and uniqueness. The notion of familiarity, developed based on Heim's previous research, does not require that the referent is previously introduced into the discourse explicitly (i.e. a case of coreference) as many would assume, but rather indicates that the existence of the referent is entailed in the context. Similarly, the notion of uniqueness, developed upon Löbner's (1985) functional view of definite descriptions, is not Russellian but rather indicates that the referent is unique in all referents entailed by the context that satisfy the descriptive contents of the definite expression. Like Heim (1982), Roberts (2003, 2004) treats pronouns as a subclass of definite noun phrases – the weak familiarity presupposition also applies to pronouns. The difference between the two (e.g. pronouns carry little semantics and cannot be supplemented with additional descriptive contents) is accounted for with an additional salience requirement: referents of the pronouns must be maximally salient in the context of their utterance. The informal versions of Roberts's accounts for definite descriptions and pronouns are replicated below in Figure 3.1.

Roberts's (2003) account successfully explains the (strong) uniqueness effect Russell (1905) observed, and at the same time avoids predicting such strong uniqueness where it does not arise. However, it inherits the same difficulties from previous familiarity-based accounts in explaining cases such as (3.5) [4], where the context does not directly entail the existence of the referent. Roberts resorts to accommodation to resolve the contradiction, stating that in such cases the familiarity presupposition is satisfied when the receiver accommodates the existence of the referent. In the case of (3.5), the reader must accommodate that Bill did go out with a women the previous night. Roberts acknowledges that in some situations the familiarity presupposition can only be satisfied through accommodation. However, she argues that such cases are not really special considering that

[4] Birner and Ward (1994) used a similar example.

Given a context C, use of a definite NP_i presupposes that it has as antecedent a discourse referent x_i which is:

a) weakly familiar in C, and
b) unique among discourse referents in C in being contextually entailed to satisfy the descriptive content of NP_i.

Given a context C, use of a pronoun Pro_i presupposes that it has as antecedent a discourse referent x_i which is:

a) weakly familiar in C,
b) salient in C, and
c) unique in being the most salient discourse referent in C which is contextually entailed to satisfy the descriptive content suggested by the person, number and gender of Pro_i.

Figure 3.1: Roberts's account of the presuppositions of definiteness (informal)

other presuppositions are routinely satisfied by accommodation, and the role of accommodation in her model is not as significant as usually assumed in discussions about the familiarity presupposition.

At the face of it, having to resort to accommodation to satisfy a familiarity presupposition may sound strange. In this case, it helps to note that the weak notion of familiarity Roberts (2003) coined is essentially a notion of existence[5]. The familiarity effect arises when both the speaker and the addressee share the assumptions that the referent exists. If one adopts the (arguably more common) view that presuppositions apply to the common ground, the presupposition of existence automatically leads to Roberts's familiarity. The issue of (3.5), while seemingly specific to definiteness, is rather a part of the more complex problem related to the nature of presupposition and accommodation. For example, one may choose to reject the common ground theory of presupposition (e.g. Gauker, 1998) to avoid the problem all together at the cost of unexplained familiarity effect, or bring accommodation into the picture[6] to help maintain the consistency of the common ground.

Anaphora, Revisited

The presuppositions of definiteness explains why anaphoric definite noun phrases depend on their antecedents for interpretation. For most pronoun uses and a large portion of definite description uses, having a coreferential antecedent is the only way to satisfy the familiarity presupposition. Of course, it is also possible for a definite noun phrase to have a referent that is not strongly familiar – the referent may have entered the common ground of the interlocutors prior to the utterance, for example via visual perception or by virtue of world knowledge, or it may be introduced "on the spot" and requires the receiver to accommodate.

Most of the time, additional descriptive contents are needed in order to access a referent that is only weakly familiar. However, an utterer may strip the description contents when part of the information is salient in the context[7]. In such cases, the receiver must reconstruct the stripped

[5] Roberts (2003) also uses the term 'informational existence'.

[6] The mechanism of presupposition accommodation is an active field of research, cf. Fintel's (2008) recent discussion for relevant details.

[7] Note that unlike Hawkins (1978) and Prince (1981), this analysis treats associative anaphora as a derived form of

contents, sometimes via accommodation, in order to interpret the resulted associative anaphors. The reconstruction process gives the definite description a functional reading, which is necessary to satisfy both the familiarity presupposition and the uniqueness presupposition.

It is evident from the preceding analysis that the presupposition of familiarity is the definitive force behind anaphora. However, it does not explain the often-perceived property of anaphors as containing fewer bits of disambiguating information. Anaphors in associative cases are abbreviated references by nature, but anaphors with strongly familiar referents are not. For such anaphors, the perception is usually a side effect of the relative ease to satisfy the informational uniqueness presupposition. For example, in majority of situations there is only one strongly familiar referent of a given kind, which makes it possible for an author to access it using a 'simple' definite description devoid of additional descriptive contents. Although such an anaphor completely satisfies the uniqueness presupposition, it may have little disambiguation power when taken out of the context.

While the presupposition of familiarity necessitates the need for anaphoric interpretation, the uniqueness presupposition plays an important role in identifying the antecedent: it dictates that there exists one and only one weakly familiar discourse referent – which is the antecedent – that satisfies the descriptive contents of the anaphor. To make this point more clear, consider the following example:

(3.7) 01. The survival of spinoff Cray Computer Corp.$_1$ as a fledgling in the supercomputer business appears to depend heavily on the creativity – and longevity – of *its*$_1$ chairman and chief designer, Seymour Cray$_2$.

02. Not only is development of *the new company*$_1$'s initial machine tied directly to Mr. Cray$_2$[8], so is *its*$_1$ balance sheet.

03. Documents filed with the Securities and Exchange Commission on the pending spinoff disclosed that Cray Research Inc.$_3$ will withdraw the almost $100 million in financing *it*$_3$ is providing *the new firm*$_1$ if Mr. Cray$_2$ leaves or if the product-design project *he*$_2$ heads is scrapped. WSJ 18:1-3

There are three different coreference chains in (3.7) – two companies, *Cray Computer Corp.*$_1$ and *Cray Research Inc.*$_3$, and one person, *Seymour Cray*$_2$. The two subsequent mentions realized in definite descriptions, *the new company*$_1$ and *the new firm*$_1$, both point to *Cray Computer Corp.*$_1$, which in turn denotes a strongly familiar discourse referent in the context that is unique in satisfying both their type (i.e. *company / firm*) and the additional descriptive content (i.e. *new*). Note that

Prince's (1981) 'containing inferrable' uses. The reason for such an arrangement is mainly three fold. First, associative cases generally have corresponding 'containing' counterparts, but not vice versa. For example, consider '*I went to visit Mr. Doe today. The life* / His life was miserable.*' Second, associative anaphora is only possible when the corresponding trigger expression (anchor) is salient in the context. In addition, Poesio and Vieira (1998) have shown that associative anaphora is relatively rare but the discourse-new descriptions, most of which being the 'containing inferrable' type, are abundant. This can also be seen as an indicator that the former construction have more stringent conditions that guide its use. While none of these observations serve as direct evidence that associative anaphora is derived from 'containing inferrable', the combination makes it more plausible than the alternative.

[8] The term (together with another subsequent mention) is underlined instead of italicized in order to reflect the distinction between coreference and anaphora, which will be elaborated in Section 3.1.4.

the satisfaction of the additional descriptive content is not realized explicitly (e.g. using descriptions such as *new Cray Computer Corp.*) but rather hinted through the pre-modifier *spinoff*, the *as*-preposition *as a fledgling*, and to some extent by the use of *The survival of* – while survival is an essential property of any being, it is typically emphasized only when the underlying entity has some difficulty in surviving, which is a scenario often applicable to new beings. Stripping the additional descriptive contents (*new*) from both definite descriptions further verifies the guidance afforded by the uniqueness presupposition. As illustrated in (3.7′) below, replacing the first definite description with *the company* creates no negative effect on the felicity of the sentence. This can be readily explained by the fact that at the point the discourse referent denoted by *Cray Computer Corp.*$_1$ is also unique in satisfying the type *company*. However, replacing *the new firm*$_1$ with *the firm* will result in a much less acceptable sentence[9], since the context at the point entails two different discourse referents (*Cray Computer Corp.*$_1$ and *Cray Research Inc.*$_3$) satisfying the descriptive contents of the substitute.

(3.7′) The survival of spinoff Cray Computer Corp.$_1$ as a fledgling in the supercomputer business appears to depend heavily on the creativity – and longevity – of *its*$_1$ chairman and chief designer, Seymour Cray$_2$. Not only is development of *the company*$_1$'s initial machine tied directly to Mr. Cray$_2$, so is *its*$_1$ balance sheet. Documents filed with the Securities and Exchange Commission on the pending spinoff disclosed that Cray Research Inc.$_3$ will withdraw the almost $100 million in financing *it*$_3$ is providing *the firm*$_1$ if Mr. Cray$_2$ leaves or if the product-design project *he*$_2$ heads is scrapped.

Gathering from the above discussion, anaphora can be seen as a device to satisfy the weak familiarity and informational uniqueness presuppositions as coined in Roberts's (2003) theory of definiteness. The need of anaphoric interpretation arises when the anaphor cannot satisfy the presuppositions of definiteness by itself. This view does not contradict the fact that anaphora is a major device of text cohesion. However, in comparison to Hirst's (1981) definition, which is essentially a description of the utterers' motivations for choosing anaphoric expressions over repeating the antecedents, the view proposed in this study provides a more coherent account for the need for anaphoric interpretation from the receiver's perspective.

3.1.2 Definiteness and the Role of Salience

While Roberts (2003, 2004) treats pronouns as definites, she also recognizes some significant differences between pronouns and definite descriptions, which subsequently lead to a different strategy in interpreting pronouns (cf. Figure 3.1). However, this partial dichotomy may not be necessary; and

[9] The new sentence is not entirely ambiguous or unacceptable. A reader can still determine the real referent of *the firm*$_1$ following the line of reasoning that since *it*$_3$ points to Cray Research Inc.$_3$ and it is highly unlikely that one company provides financing to itself, *the firm*$_1$ must refer to the other company. However, this complexity is clearly unwarranted under the Gricean conversational maxim of manner.

its elimination will make Roberts's (2003) theory a completely unified, more attractive account of definites in general.

Perhaps the most significant difference between pronouns and definite descriptions is that pronouns convey little semantics by themselves. Co-occurring with their obvious lack of descriptive contents is the fact that their interpretation is usually highly dependent on salience. The following example, adapted from Roberts (2003), illustrates the salience-driven behavior of pronouns that is not typical to definite descriptions.

(3.8) A woman entered from stage left.
Another woman entered from stage right.
She / *#The woman* / *The SECOND woman* was carrying a basket of flowers.
Roberts (2003, ex. 40, adapted)

Roberts (2003) asserts that the use of pronoun *She* is felicitous while the definite description *The woman* is not[10]. In order to make the use of a definite description salient, it is necessary to add additional descriptive contents so that the presupposition of uniqueness can be satisfied, as in the case of *The SECOND woman*.

However, as Roberts (2003) also noted, it would be over-exaggerating to state that definite descriptions are free of influence from salience. In fact, as shown in the following excerpt[11], the role of salience in definite descriptions can be rather prominent, especially in the case where there are more than one discourse referents of the same type in the context.

(3.9) 26. Cray Computer$_1$ has applied to trade on Nasdaq.
27. Analysts calculate Cray Computer$_1$'s initial book value at about $4.75 a share.
28. … Cray Research$_2$ is transferring about $53 million in assets, primarily those related to the Cray-3 development, which has been a drain on Cray Research$_2$'s earnings.
29. Pro-forma balance sheets clearly show why Cray Research$_2$ favored the spinoff.
30. Without the Cray-3 research and development expenses, *the company*$_2$ would have been able to report a profit of $19.3 million …
31. On the other hand, had it existed then, Cray Computer$_1$ would have incurred a $20.5 million loss.
32. Mr. Cray … will work for *the new Colorado Springs, Colo., company*$_1$ as an independent contractor – the arrangement he had with Cray Research$_2$.
33. Regarded as the father of the supercomputer, Mr. Cray was paid $600,000 at Cray Research$_2$ last year.
34. At Cray Computer$_1$, he will be paid $240,000.

[10] I have some reservations regarding whether the use of *The woman* results in complete infelicity. However, it certainly leads to a less felicitous sentence than the case of *She*.

[11] Roberts (2003) offers a variant of (3.8) by inserting an adverbial, '*Later in the act,* ' in front of '*another woman*' in the second sentence (Roberts, 2003, ex. 43). The variant allows felicitous use of '*The woman*' in the third. While she seems to believe it is a special case, I see it as belonging to the same class of phenomena as illustrated in (3.9).

35. Besides Messrs. Cray and Barnum, other senior management at *the company*$_1$ includes ... WSJ 18:26-35

(3.9) is selected from the same news story as (3.7). The story is largely about a spinoff company, *Cray Computer Corp.*, and some details of the financial arrangement it has with its parent company, *Cray Research Inc.* While the beginning portion of the article (3.7) demonstrates that failure to satisfy the uniqueness presupposition leads to infelicitous results, the two occurrences of *the company* at sentences 30 and 35 in (3.9) seem to indicate otherwise. The most probable factor that licenses the felicitous use of *the company* in both sentences is salience[12].

One important thing to note is that accepting the role of salience in definite description anaphora does not necessarily allow interchangeable use of definite descriptions and pronouns – by now it is generally accepted that they are mostly not. The plethora of evidences cited by Roberts (2003) aside, a recent study by Preiss, Gasperin, and Briscoe (2004) also demonstrates that definite description anaphora cannot be readily resolved with an approach known to work reasonably well with pronouns. However, on the flip side, the fact that pronouns and definite descriptions are generally not interchangeable does not necessarily falsify the previous tentative conclusion either. This situation allows room for a further hypothesis that, although definite descriptions are less sensitive to salience than pronouns are in general, they may become more so when there is pressure to choose from multiple possible antecedents.

The main obstacle to this hypothesis is (3.8), which illustrates a scenario where definite descriptions become infelicitous due to the presence of multiple possible antecedents while pronouns can be used felicitously. I argue that this may be attributed to the difference in salience models applicable to definite descriptions and pronouns. The infelicity of simple definite description in (3.8) could be caused by the salience model's inability to disentangle the candidates. This phenomenon is not unique to definite descriptions. As (3.10) illustrates, pronouns can also become infelicitous at the presence of multiple equally-salient candidates for antecedent.

(3.10) Mary and Jane are good friends.
#She loves candy.

With (3.10) in mind, now we can examine Roberts's (2003) account for pronouns from a different perspective – instead of assuming that salience is an integral part of the uniqueness presupposition of pronouns, why not consider it as a pragmatic utility to enforce the uniqueness presupposition in general? This proposition would remove salience from the uniqueness presupposition of pronouns but at the same time explain the felicitous uses of *the company* in (3.9). Similarly, the salience requirement in clause b), which states that the antecedent of a pronoun must be salient in the context, may also be removed – if we consider salience as a pragmatic means of limiting the search scope

[12] Roberts (2003) discusses a similar case (footnote 3, page 292). She refers to domain restriction (or 'pragmatic enrichment' in her terms) as the licensing factor for the phenomenon observed in (3.9). My point here is that regardless of the name, what licenses (3.9) seems to be the same set of factors that is behind the common pronoun behaviors.

for antecedents, i.e. the accessibility of familiar discourse referents. Just as pronouns usually cannot access discourse referents outside a certain window[13], definite descriptions may also have difficulty accessing 'inactive' discourse referents. Example (3.11) illustrates an interesting phenomenon that arises when a non-salient discourse referent needs to be accessed.

(3.11) 01. Toni Johnson pulls a tape measure across the front of what was once a stately Victorian home.
02. A deep trench now runs along its north wall, exposed when the house lurched two feet off its foundation during last week's earthquake.
03. A side porch was ripped away.
04. The chimney is a pile of bricks on the front lawn.
05. The remainder of the house leans precariously against a sturdy oak tree.
06. The petite, 29-year-old Ms. Johnson, dressed in jeans and a sweatshirt as she slogs through the steady afternoon rain, is a claims adjuster with Aetna Life & Casualty.
... (topic shifted to insurance claims processing) ...
12. "That's my job – get policyholders what they're entitled to," says Bill Schaeffer, a claims supervisor who flew in from Aetna's Bridgeport, Conn., office.
13. *The Victorian house that Ms. Johnson is inspecting* has been deemed unsafe by town officials. WSJ 766:1-13

The discourse referent established in sentence 1 by *what was once a stately Victorian home* remained at the center of discussion until sentence 6-7, when the topic shifted to claims processing. After a lengthy discussion of the second topic, by sentence 13, it is clearly impossible to access the referent again through either '*The house*', '*The Victorian house*', or even '*The stately Victorian house*' for that matter, despite the fact that the house referent remains unique and familiar. The author resorts to associating the house with a proper name, *Ms. Johnson*, in order to facilitate access to the referent. The result of this strategy is a definite description post-modified by a restrictive relative clause, which, as discussed later in the chapter, happens to be a common form of 'unfamiliar' uses of definite descriptions.

To summarize, salience has similar effects – limiting access to only a subset of familiar referents and (partial) ranking of the accessible ones – on both pronouns and definite descriptions. Therefore a slight revision to Roberts's (2003) original analysis for definite descriptions will work for both definite descriptions and pronouns:

3.1.3 Anaphora and Indefinite Descriptions

Although one rarely feels the need to interpret a non-definite description as anaphoric, there are also obvious cases where the need does arise. One such example is the mention of *A side porch* in the

[13] There are many factors that come into play when determining the overall salience of a potential antecedent, distance-based window is only one (and probably the simplest) of them.

Given a context C, use of a definite description or pronoun DEF$_i$ presupposes that it has as antecedent a discourse referent x_i which, **subject to the influence of salience**, is:

a) weakly familiar in C, and
b) unique among discourse referents in C in being contextually entailed to satisfy the descriptive content of DEF$_i$.

Figure 3.2: A revision of Roberts's account of the presuppositions of definiteness (informal) that applies to both definite descriptions and pronouns

third sentence of (3.11) on page 35.

A number of observations can be made from the first three sentences of the excerpt. Firstly, it is impossible to determine whether an indefinite is anaphoric through syntactic means alone. For example, both the trench (sentence 2) and the side porch are introduced using indefinite descriptions, both function as subject of the sentence, and the non-anaphoric expression, *A deep trench*, is closer to the antecedent. Obviously, it is the semantic relationship between *porch* and *home* that licenses the anaphoric use of *A side porch*. In other words, *porch* is interpreted as relational (Löbner, 1985). The second observation is that in order to satisfy both the semantics of the indefinite article as denoting an unidentified instance[14] and the terms of anaphora as having a specific antecedent, an anaphoric indefinite expression implies the existence of multiple entities of the same type, all of which being associated to the antecedent. In other words, *A side porch* is an alternative expression of 'One of the porches'. This explains why the chimney, which is usually limited to one per household, is introduced using a definite phrase in sentence 3. The third point of interest is whether – and how – knowledge about the indefinite anaphora contributes to the understanding of the text. Failure to associate the side porch with the home does not significantly alter the meaning of the sentence on the surface level[15]: an unidentified side porch has been ripped away (from its original point of attachment). However, the implied message, that the previously mentioned house was damaged, is no longer available. The whole excerpt also becomes less coherent because the sentence no longer fits into the surrounding context.

The benefit of interpreting an indefinite description as anaphoric is less evident in many other cases, such as the expression *a higher offer* in (3.12):

(3.12) New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire, saying that the risks were too high and the potential payoff too far in the future to justify *a higher offer*. The move leaves United Illuminating Co. and Northeast Utilities as the remaining outside bidders for PS of New Hampshire … WSJ 13:1-2

[14] There are other uses of the indefinite articles, such as denoting a type. However they are not relevant to this particular situation.

[15] Note that the same cannot be said about associative anaphors realized using definite descriptions, e.g. *The chimney*, due to the presuppositions of definiteness as discussed earlier.

Unlike definite descriptions, the non-definite descriptions are not ‘marked’ for anaphora. Therefore the biggest issue is not locating the antecedents, but rather finding out whether to interpret them as anaphoric. Gathering from the limited number of examples encountered in this study, there are two main factors behind the need for anaphoric interpretation of indefinite descriptions. The first factor is discourse coherence – if a non-anaphoric reading renders the discourse incoherent, the description should be interpreted as anaphoric. For example, consider a centering-based analysis of sentence 3, (3.11): if associative anaphora is deemed as a form of realization[16], an anaphoric reading of *A side porch* would establish it as the backward-looking center of the sentence. In comparison, a non-anaphoric reading of the indefinite description would leave the sentence without a backward-looking center, violating one of the main claims of the centering theory. The modified sentences in (3.11$'$) both contain backward-looking centers, and the need to interpret *A side porch* as anaphoric seems to be significantly reduced.

(3.11$'$) Toni Johnson pulls a tape measure across the front of what was once a stately Victorian home. A deep trench now runs along its north wall, exposed when *the house* lurched two feet off its foundation during last week’s earthquake.

a. The earthquake also turned a side porch into pieces.

b. A side porch was ripped away by the earthquake.

The other factor driving the need for anaphoric interpretation is the nature of the head noun – if the head noun is relational (Löbner, 1985) and the argument(s) of the relation has not been associated to it in an obvious manner, the need for anaphoric interpretation arises. In the case of (3.12), the need arises because the essence of an offer is a price proposed to buy something. However, this factor is inherently fuzzy. As Löbner (1985) explained, many nouns are ambivalent with regard to sortal and relational uses. The strength of the perceived need may also vary depending on the degree of association provided by the text. For example, even though *friend* is a relational concept, the existence of an indirect albeit strong relationship (subject-verb-object) in (3.13.a) seems to weaken the perceived need to give *a friend* an anaphoric reading (compared to the case of 3.13.b).

(3.13) John did not go home directly after work.

a. He met a friend on the street and they went to a pub.

b. *A friend* called him and they went to a pub.

As (3.14) shows, one can also expect bare plurals to receive anaphoric readings under certain circumstances.

(3.14) 01. A form of asbestos once used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

[16] As Poesio, Stevenson, et al. (2004b) noted, allowing indirect realization have significant positive impact on the verifiability of Constraint 1 (each utterance has exactly one backward-looking center).

02. The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.

03. Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

04. Although *preliminary findings* were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem. WSJ 3:1-4

The bare plural *preliminary findings*[17] in sentence 4 satisfies both conditions presented in earlier discussions: the concept *finding* is inherently relational, and the coherence of the text is reduced[18] if it is not interpreted as anaphoric.

One might notice that the indefinite cases discussed so far are all associative. Other researchers (e.g. Ushie, 1986; Nishida, 2007; Fraurud, 1990) discussed scenarios where indefinite descriptions seem to refer to strongly familiar discourse referents. Cases presented by Ushie (1986) and Nishida (2007) are mainly related to predicative uses of indefinite descriptions, as in (3.15):

(3.15) S.-Y. Kuroda has illuminated a great many aspects of the study of language in his fascinating and wide-ranging contributions. This collection of essays . . . is a fitting tribute to the work of *an outstanding scholar*. Nishida (2007, ex. 2)

(3.16) "Funny Business"(Soho, 228 pages, $17.95) by Gary Katzenstein is anything but. It's the petulant complaint of *an impudent American whom Sony hosted for a year while he was on a Luce Fellowship in Tokyo – to the regret of both parties*. WSJ 37:27-28

A similar case, (3.16), is also found in the WSJ corpus. Under the 'interpretation'-based view of anaphora as adopted in this study, these cases are not considered anaphoric[19]. Fraurud (1990) is mainly concerned with the generic uses of non-definite descriptions. Following G. N. Carlson (1977b), bare plurals bearing strictly generic readings are 'proper names of kinds of things', and are therefore not anaphoric. Obviously, other generic cases, including but not limited to those expressed in bare singular forms, indefinite descriptions, or even definite descriptions, fall into the same basket and should receive similar treatment[20].

[17] The expression *preliminary findings* used under the context of sentence 4 clearly bears an indefinite reading, not a generic one. If one follows G. N. Carlson's (1977a, 1977b) analysis on the semantics of bare plurals, the indefinite reading can be seen as created by the context (i.e. *reported more than a year ago*).

[18] In this case, measuring coherence is not as straight-forward as it is for (3.11). Centering cannot be applied directly unless *preliminary findings* is considered a realization of *crocidolite* in sentence 3, which is unfortunately not an optimal choice for antecedent. The best choice of antecedent is probably the fact that crocidolite is proven hazardous, or *researchers*, both appearing in sentence 1.

[19] In fact, if one follows the discussion in Section 3.1.4, they are not even coreferential. However, these cases do represent a significant category of indefinite uses, and should probably be treated under a separate task definition (as suggested by Deemter & Kibble, 2000).

[20] Note that this does not solve all the issues raised by Fraurud. As Fraurud (1990) puts it, "One might of course choose to regard these syntactically indefinite NPs as semantically definite, due to their genericity. But this does not provide an immediate solution to the problems of (i) how to recognize that a particular indefNP is generic and thus potentially co-referent with a preceding one, and (ii) how to model the interpretation of such instances of indefNPs." Adopting G. N. Carlson's (1977b) analysis (see also G. Carlson & Pelletier, 1995) seems to solve the second issue. However, there does not seem to

It is still too early to make a conclusion on the nature of indefinite anaphora. However, discourse coherence seems to be the strongest driving force when there is a perceived need to interpret an indefinite as anaphoric. Once it is decided that an indefinite should be interpreted in association with an antecedent, it can generally be translated into a form that involves definite description[21] and therefore treated under the same framework of definite description associative anaphora.

3.1.4 Anaphora and Coreference

According to Deemter and Kibble (2000), two discourse entities are coreferential if and only if they have the same referent. Deemter and Kibble also point out that while anaphora and coreference can coincide, they are essentially different, noting the differences between the relationships in some key properties such as symmetricalness and context-sensitivity of interpretation. Deemter and Kibble's analysis provides valuable insights into the notion of coreference and issues related to existing annotation practice. This section follows their line of reasoning and discusses the relationship between coreference and anaphora.

Deemter and Kibble (2000) raise three issues related to the MUC coreference annotation practice: problems with non-referring noun phrases and bound anaphora, problems with intentionality and predication, and difficulties determining what is markable. The problem with bound anaphora, as illustrated in example (3.17), is that there does not seem to be a real-world entity that correspond to the bound variable. The remaining two problems are both related to the scope of coreference. Examples (3.18) and (3.19) are used by Deemter and Kibble to demonstrate the complications that can arise when intensional descriptions (e.g. *Henry Higgins*, Hirschman & Chinchor, 1997) are marked as coreferential with extensional descriptions (e.g. *sales director of Sudsy Soaps* and *president of Dreamy Detergents*).

(3.17) Every TV network reported *its* profits. Deemter and Kibble (2000, ex. 1c)

(3.18) Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents. Deemter and Kibble (2000, ex. 3)

(3.19) The stock price fell from \$4.02 to \$3.85; Later that day, it fell to an even lower value, at \$3.82. Deemter and Kibble (2000, ex. 4)

(3.18) and (3.19) are of less interest to this study, as the discussed entities certainly do not involve in any anaphoric relationship. However, the fundamental reasons why they are not coreferential are closely related to many anaphoric phenomena. For example, consider the definite description *The adjuster* in (3.20):

be a satisfactory answer to the first one yet (Dahl, 1995 points out some minimal grammatical markers, but they are probably too weak for a practical system).

[21] For bare plurals, the translation can be either '*all of the NPs*' or '*some of the NPs*', but the difference is not essential to the purpose of this study.

(3.20) 06. The petite, 29-year-old Ms. Johnson, dressed in jeans and a sweatshirt as she slogs through the steady afternoon rain, is a claims adjuster with Aetna Life & Casualty.

...

75. *The adjuster* hadn't completed all the calculations, but says:"We're talking policy limits." WSJ 766:6-75

Should the antecedent be *Ms. Johnson*, or rather *a claims adjuster with Aetna Life & Casualty*? If one makes the distinction between individuals and functions (which is suggested by Deemter and Kibble (2000) as one of the possible remedies[22] and is also the view adopted by this study), it follows that *a claims adjuster* does not introduce an independent discourse referent that is compatible with *The adjuster*[23]. On the other hand, allowing *a claims adjuster* to serve as the antecedent implies that it is also coreferential with *Ms. Johnson*.

Bound anaphora also happens to be one of the central issues to anaphora. Among other things, it has to do with what exactly does a discourse entity refer to. If one follows the analyses of Heim (1982) and Roberts (2003, 2004) that they point to 'discourse referents' instead of real-world objects, the expression *Every TV network* does introduce a corresponding discourse referent (under the scope of quantification) to the context, which is available to serve as the antecedent of *its*. In other words, if referents are defined on the common ground of the interlocutors instead of the real world, a bound anaphor can be seen as coreferential to its antecedent.

As mentioned earlier, Deemter and Kibble (2000) also noted that coreferential expressions can be context-insensitive, i.e. the interpretation of one need not depend on the other. They used the name *President W. I. Clinton* and the description *Hillary Rodham's husband*[24] as an example to illustrate that coreferential entities are not necessarily anaphoric. This case represents a very challenging problem that is quite different from the ones tackled by most existing intra-document coreference resolution systems. On one hand, a significantly larger search space needs to be explored. On the other hand, it also demands much more detailed world knowledge. If no additional information is given about the relationship between Mr. Clinton and Hillary in the same document (assuming an intra-document coreference situation), resolving the coreferential link would require world knowledge about the relationship between two specific individuals. Even if such information is provided in the document, it still takes non-trivial capability of inference and truth-condition tracking to solve the problem. Reintroducing the same entity without providing explicit links to the previous mention is not a very common practice due to the confusion it may cause. However, an author may take the liberty to do it when the entity is not considered central to the message the article conveys. For example, in (3.21) a paper factory is introduced in sentence 17 as *the West Groton, Mass., paper*

[22] For relevant details on the subject, see discussions by Dowty, Wall, and Peters (1981, Appendix iii) and Partee (1987), among many others.

[23] As noted by Deemter and Kibble (2000), it is still possible to serve as the antecedent for expressions such as '*the position*'.

[24] This is probably not the best example, since the relationship between them is essentially the same as what is illustrated by (3.18), especially if they are mentioned in the same document and an explicit link is provided through a copula or apposition.

factory and later on reintroduced using an indefinite description *a factory* in sentence 24.

(3.21) 17. The percentage of lung cancer deaths among the workers at the West Groton, Mass., paper factory appears to be the highest for any asbestos workers studied in Western industrialized countries, he said.

18. *The plant*, which is owned by Hollingsworth & Vose Co., was under contract with Lorillard to make the cigarette filters.

...

24. About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s. 25. Areas of *the factory* were particularly dusty where the crocidolite was used.

...

28. "There's no question that some of those workers and managers contracted asbestos-related diseases," said Darrell Phillips, vice president of human resources for Hollingsworth & Vose. WSJ 3:17-28

The only clue that the two discourse entities may be coreferential is the quotation of a Hollingsworth & Vose employee towards at the end of the excerpt, which is by far too vague to serve as evidence. In fact, it takes considerable effort even for a human being to find a proof[25] that they are indeed coreferential.

Fortunately, many (if not most) of the coreferential-only relationships are not so difficult to discover, and they are often closely related to the process of anaphora resolution. Taking proper names for example: it is comparatively easy to establish the coreference relationship between multiple proper name expressions using some form of string matching. Once the relationship is established, the expressions all points to the same individual whose identity remain unchanged across the discourse. The various properties that are attributed to the individual at different locations could then be uniformly applied[26] to the same individual. Needless to say, this can be very helpful for anaphora resolution.

3.2 Classification of Definite Descriptions

As discussed in Section 3.1, this study adopts Roberts's (2003) view that the use of definite noun phrases presupposes both weak familiarity and informational uniqueness. Her theory makes it possible to explore the issue of definite description classification from two distinct angles – familiarity and uniqueness – while still maintaining a coherent picture. This section presents an attempt towards this direction.

[25] For example, by reading between the lines of the story by Levin (1987).

[26] Ideally, some kind of truth-condition tracking should be involved here to handle cases such as (3.18).

3.2.1 Related Research

A number of researchers have looked into the classification of definite descriptions. Some of the classification schemes, such as those of Hawkins (1978), Prince (1981, 1992), and Löbner (1985), are essentially by-products of the researchers' inquiries into the essence of definiteness. These taxonomies provide invaluable insights into how definite descriptions behave with regard to familiarity and uniqueness. A notable exception is the study by Poesio and Vieira (1998), whose empirical examination sheds some new light on the issue with corpus statistics and inter-annotator agreement data.

Hawkins' Descriptive Analysis

Extending on previous studies of Christophersen (1939) and Jespersen (1949), Hawkins (1978) identifies eight distinct uses of definite descriptions:

1. Anaphoric use
   These are definite descriptions whose referents are introduced in the discourse.

   (3.22) Bill was working at a lathe the other day. All of a sudden *the machine* stopped turning. Hawkins (1978, ex. 3.30)

2. Visible situation use
   These definite descriptions are used to refer to entities visible to both parties of the conversation.

   (3.6) Pass me *the water*, please.

3. Immediate situation use
   This type is similar to the visible situation use, however, the existence of the referent is inferred from the local situation.

   (3.23) Harry, mind *the table*! Hawkins (1978, ex. 3.53)

4. Larger situation use relying on specific knowledge about the referent
   The referents of these definite descriptions are not located in the immediate situation. However, the interlocutors have shared knowledge about the referents. For example, there is a gibbet in the English town of Halifax, West Yorkshire, and the inhabitants are aware of its existence at Gibbet Street. If the gibbet is taken down, the local press can simply report:

   (3.24) *The Gibbet* no longer stands. Hawkins (1978, p. 119)

5. Larger situation use relying on general knowledge
   The referents of these definite descriptions are also located outside of the immediate situation. However, their existence is inferred from general knowledge rather than personal experience. Consider the following conversation between two guests upon arrival of a wedding:

   (3.25) Have you seen *the bridesmaids*? Poesio and Vieira (1998, ex. 5)

6. Associative anaphoric use

   Interpretation of associative definite descriptions also depend on the interlocutors' shared knowledge of generic relationships between objects. However, unlike the larger situation cases, at least one participant of the relationship is already mentioned in the discourse.

   (3.26) The man drove past our house in a car. *The exhaust fumes* were terrible.

   Hawkins (1978, ex. 3.61)

7. 'Unfamiliar' use (with explanatory modifiers)

   Definite descriptions that are not familiar in Christophersen's (1939) sense (i.e. by virtue of being anaphoric, being related to the situation of the utterance, or being associated to another discourse entity) are put under this category. Hawkins (1978) further identifies four subtypes according to the syntactic construct of the definite descriptions, namely the associative clauses (3.3), the NP-complements (3.4), the establishing relative clauses (3.5), and the nominal modifiers (3.27).

   (3.3) I remember *the beginning of the war* very well ...

   (3.4) Bill is amazed by *the fact that there is so much life on Earth*.

   (3.5) What's wrong with Bill?
   Oh, *the woman he went out with last night* was nasty to him.

   (3.27) I don't like *the colour red*. Hawkins (1978, ex. 3.115)

8. 'Unexplanatory' modifiers use

   Certain modifiers, such as *same*, *first*, and the superlatives, require the use of the definite article. Definite descriptions in this category generally fit well into Donnellan's (1966) 'attributive use', i.e. they are usually used without requiring both (or even either) parties of the conversation to have knowledge about the specific referent[27].

   (3.28) *The fastest person to sail to America* was an Icelander.

   Hawkins (1978, ex. 3.133)

Overall, Hawkins' (1978) list provides fine-grained and comprehensive coverage[28] for definite descriptions. Critiques of Hawkins' analysis are mainly concentrated in three areas: the fundamental assumptions about definiteness, appropriateness of the term 'unfamiliar', and issues related to its practical application to annotation. As discussed earlier, researchers reached different conclusions on the nature of definiteness. The theory proposed by Hawkins, according to Fraurud (1990), is essentially familiarity-based. There are other researchers (e.g. Löbner, 1985) who clearly favor non-ambiguity/uniqueness. The remaining two issues are also raised by Fraurud. As Fraurud puts it, "it should also be pointed out that 'unfamiliarity' of the referent is neither a necessary property

[27] As Poesio and Vieira (1998) pointed out, it is possible to use an 'unexplanatory' modifier definite description in a situation where all interlocutors have specific knowledge about the referent.

[28] With the notable exception of definite descriptions representing generic concepts, as pointed out by Poesio and Vieira (1998).

of definite NPs with 'explanatory modifiers', nor is it a property that is confined to this structural type". One of the naturally occurring examples supporting this view is excerpt (3.11) on page 35, in which an 'unfamiliar' description (*The Victorian house that Ms. Johnson is inspecting*) is used to help the reader access a discourse referent that is no longer salient. However, 'Unfamiliar' uses are not limited to non-salient referent either, as evidenced by (3.29):

(3.29) Frank told Sheriff Smith that Ringo would arrive on Thursday. *The news that Ringo would be in town* filled the Sheriff with worry. Vieira (1998, ex. 2.11)

Although one could arguably remove the complement clause to further increase the cohesion of the text, (3.29) seems felicitous in its original form. Fraurud (1990) also points out that some of the categories in Hawkins' (1978) list overlap each other, causing ambiguities in many cases. One of the examples is (3.30), where the interpretation of *the next train* depends on both the local situation and the destination, *Gothenburg*, which is given in the discourse:

(3.30) (at a ticket office of the central station in Stockholm)
I am going to Gothenburg. When does *the next train* leave? Fraurud (1990, ex. 9)

Prince's Assumed Familiarity

Dissatisfied with the traditional binary distinction between 'given' and 'new', Prince (1981, 1992) proposes a more fine-grained division along the axis of familiarity from the perspective of what an author assumes about the receiver(s)' knowledge. Under the general categories of 'New', 'Inferrable', and 'Evoked', she makes further distinctions based on the source of familiarity – whether the information is provided by the discourse (Discourse-new/old) or it is assumed that the hearer is already aware of it (Hear-new/old). Prince (1981) presents this three-dimensional taxonomy using a tree-like structure containing seven leaf nodes: Brand-new Unanchored, Brand-new Anchored, (new but) Unused, Containing Inferrable, Non-containing Inferrable, Textually Evoked, and Situationally Invoked.

As Poesio and Vieira (1998) discussed, Prince's (1981) classification groups some of the distinct uses identified by Hawkins (1978) on semantic basis. However, even more important to the purpose of this study is the notion of Hear-new/old. What I consider essential to this notion is whether the author assumes that some of the recipients may not have have sufficient knowledge about a discourse entity, rather than whether or not a particular reader is familiar with it. For example, Prince (1992, 1981) discusses that a containing-inferrable to one reader may be hearer-old and discourse-new for another, which makes it suitable for multi-receiver discourses.

Poesio and Vieira's Empirical Study

Poesio and Vieira's (1998) empirical study (also see Vieira, 1998) is an important supplement to Hawkins' (1978) list. The study not only reveals, in a measurable way, the gap between Hawkins' theoretically-oriented offerings and the realities faced by the task of large-scale corpus annotation, but also provides additional insights into the properties of definite descriptions in general.

Poesio and Vieira (1998) discuss two annotation experiments. The first one is performed on on a total of 1,040 definite descriptions from 20 randomly selected WSJ articles, and the second one is performed on a total of 464 definite descriptions from 14 (one of which is also present in the first set) articles in the same corpus. The classification schemes of both experiments are motivated by Hawkins' (1978) list. The categories used for the first experiment are: Anaphoric same head, Associative, Larger Situation/Unfamiliar, Idiom, and Doubt. The first category includes only anaphoric definite descriptions that share the same head nouns with their antecedents. The rest of the anaphoric cases are moved to the 'Associative' category, which includes the cases that fall under Hawkins' (1978) 'Associative anaphoric use' as well. The 'Larger Situation/Unfamiliar' category combines the four categories of 'Larger situation' uses (4 and 5) and 'Unfamiliar' uses (7 and 8) from Hawkins' (1978) list. Considering the corpus consists of only news stories, this scheme effectively covers all uses described by Hawkins (1978)[29]. The second experiment uses a revised scheme that also contains five categories: Coreferential, Bridging, Larger Situation, Unfamiliar, and Doubt. Under this scheme, the 'Bridging' category is largely in alignment with Hawkins' (1978) 'Associative anaphoric use'. Instead of repeating the details of the experiments, only part of the important findings are discussed in this section.

The first interesting observation from Poesio and Vieira's (1998) study is the distribution of categories in the corpus:

- Only about half of the definite descriptions are anaphoric (either directly or associatively). Around 48% of the definite descriptions in the first experiment are anaphoric according to the authors. A slightly higher percentage (51%-54%) of the definite descriptions in the second experiment are found to be either coreferential or associative anaphoric according to the annotators.
- Majority (over 60%) of the anaphoric instances share the same head noun with their antecedents in the first experiment.
- Associative anaphora represent a relatively small but non-ignorable category (6%-11% of all instances in the second experiment, according to the annotators).

Equally important are the inter-annotator agreement data and sources of annotator disagreement:

- Annotators can generally reach reasonable agreement on which definite descriptions are coreferential and which are discourse-new, but not on a more fine-grained categorization.
- A large portion of the inter-annotator disagreements, according to Poesio and Vieira (1998), can be attributed to the overlapping between the categories proposed by Hawkins (1978).
- Associative anaphora causes a lot of difficulties in annotation. Discourse entities that have

[29] However, as Poesio and Vieira (1998, footnote 10) noted, there are a few cases of 'Immediate situation' uses.

associative anaphoric readings according to one annotator are often interpreted as discourse-new or coreferential by another.

Löbner's Functional View

Löbner (1985) proposes that definite descriptions denote functional concepts that unambiguously assign objects to given situations or other objects in a given situation. Löbner starts by illustrating that nouns divide themselves into roughly two large groups – sortal and relational. Although many nouns are ambivalent, the distinction between a sortal concept (e.g. woman) and a relational one (e.g. wife) is fundamental. The subtype of relational noun whose defining relationship is a one-to-one mapping (e.g. father) is identified as functional. The idea is then further generalized to cover a large group of definites, which Löbner calls semantic definites:

> An NP is a *semantic definite* iff it represents a functional concept, independently of the particular situation referred to. (Löbner, 1985, page 299)

The rest of the definites are identified as pragmatic definites. Additional sub-categories are also identified for each type: semantic definites are grouped by the number of arguments attached to the functions and whether they are provided explicitly; and the pragmatic definites are grouped by their distinct uses, namely endophoric, anaphoric, and deictic. The three pragmatic uses largely correspond to Hawkins' (1978) 'Unfamiliar' use with establishing relative clauses[30], Anaphoric use, and Visible situation use[31], respectively [32].

It is important to note that Löbner's (1985) functional concepts always take situation as one of their arguments. Thus in a given situation, an FC1 functional concept (e.g. a proper name) unambiguously maps to an object. Similarly, an FC2 functional concept takes one additional argument (e.g. '*the President of the U.S.*'), although sometimes implicitly, as in '*The Prime Minister has resigned.*'[33]

Aside from those corresponding to the pragmatic uses, all other items on Hawkins' (1978) descriptive list fall under the semantic definite category. For some of the items, such as the Larger situation use relying on general knowledge and the Associative anaphoric use, this provides a natural grouping from the semantic point of view. Löbner's (1985) account also provides satisfactory explanation to many other items, such as the 'Unexplanatory' modifiers use and the Larger situation use relying on specific knowledge about the referent. Löbner explains that in the former case the adjectival attributes are functions, and the latter is considered as a special form of proper names.

[30] Definite descriptions post-modified by prepositions other than *of* also belongs to this category.

[31] Löbner (1985) regards the Immediate situation use as semantic, not deictic. See his discussion for details.

[32] Vieira (1998, Tables 2.1-2.4) provides a detailed comparison of the different terms used by Hawkins (1978), Prince (1981), and Löbner (1985), among others.

[33] As noted by Löbner (1985), this case is somewhat ambiguous between FC1 and FC2. However, for the purpose of this study, further distinction beyond semantic definite is not necessary.

3.2.2 Viewing Definite Descriptions from Both Perspectives

Because definiteness presupposes not only familiarity but also uniqueness, it helps to examine definite descriptions from both perspectives. Since Hawkins' (1978) list is the most fine-grained, a slightly modified version will be used as a starting point. Below is a recapitulation of the categories, followed by the shorter names or acronyms that will be used in the rest of the section:

- Anaphoric use (Anaphoric)
- Visible situation use (Deictic)
- Immediate situation use (Immediate)
- Larger situation use relying on specific knowledge about the referent (LSU-specific)
- Larger situation use relying on general knowledge (LSU-general)
- Associative anaphoric use (Associative)
- 'Unfamiliar' use with explanatory modifiers (Unfamiliar)
  - NP-complement (Complement)
  - associative clause (*of*-Prep)
  - nominal modifier (Nominal)
  - establishing relative clause (Pragmatic)
- 'Unexplanatory' modifiers use (Adjectival)

The Visible situation use in Hawkins' (1978) list is renamed to 'Deictic' in order to reflect the observation by Löbner (1985) that perceptions other than visual can give rise to the same effect. Similarly, the original 'Unfamiliar' use with establishing relative clause category is replaced with Löbner's (1985) pragmatic endophoric category, because it covers both the establishing relative clause use and post-modifications by prepositions other than *of*. In addition, the 'Unexplanatory' modifiers use category is replaced with Löbner's (1985) 'Complex FC1' class, which contains definite descriptions modified by adjectives such as superlatives and ordinals as well as *next*, *last*, *only*, *same*, and *other*, etc.

Familiarity

For the purpose of this study, three mutually exclusive levels of familiarity are identified. Following Roberts (2003), definite descriptions with strong familiarity have referents that are already introduced in the discourse; those fail to meet the requirement of strong familiarity but nevertheless have referents whose existence are entailed in the common ground of the interlocutors are categorized as being weakly familiar[34]; and the instances that do not have to meet even the weak familiarity in order to be felicitous are marked as potentially requiring accommodation.

[34] Note that Roberts's (2003) weak notion of familiarity subsumes strong familiarity.

Use	Strong	Weak	Accommodation
Anaphoric	X		
Deictic		X	
Immediate			X
LSU			
-specific		X	
-general		X	
Associative			X
Unfamiliar			
-Complement		X	
-Nominal		X	
-*of*-Prep			X
-Pragmatic			X
Adjectival		X	

Table 3.1: Minimum familiarity requirements for definite descriptions

Table 3.1 gives a summary of the minimum requirement for familiarity of each category. The familiarity requirements of the first two categories can be straight-forwardly derived from their definitions. Similarly, by definition, definite descriptions belonging to the Immediate category requires the receiver to accommodate the existence of the referent. Roberts (2003) outlines the necessary conditions of presupposition accommodation as follows:

> Necessary Conditions on Presupposition Accommodation:
>
> (a) Retrievability: what the hearer is to accommodate is easily inferable, by virtue of its salience and relevance to the immediate context, and
>
> (b) Plausibility: the accommodated material is unobjectionable.
>
> (Roberts, 2003, page 303)

Therefore, in the same scenario (reminding a blind friend in my house) where (3.23) is felicitous, one cannot shout (3.23′).

(3.23′) Harry, mind *the elephant*!

The two LSU categories both require weak familiarity, because there is no immediate context to support the process of accommodation. However, certain cases in the Associative category and some subcategories of the Unfamiliar class may require the receivers to accommodate the existence of the referent. Consider the following examples:

(3.31) John was murdered yesterday. *The knife* lay nearby. H. H. Clark (1975, ex. 20)

(3.32) (Bill, when asked where he was going:)
A friend asked me to fix *the sunroof of his car*.

(3.5) What's wrong with Bill?
Oh, *the woman he went out with last night* was nasty to him.

These cases can be used felicitously without requiring previous knowledge that John was murdered with a knife, that the car has a sunroof, or that Bill did go out with a woman.

Uniqueness

As noted by Roberts (2003), the (Russellian, or semantic) uniqueness effect consistently arises when the referent is only weakly familiar. For example, consider (3.33):

(3.33) This car has a statue on *the dashboard.* Roberts (2003, ex. 5)

In (3.33), the dashboard, by virtue of being connected to a particular car, is understood as unique in the world. As Roberts explains it, the semantic uniqueness is a conversational implicature[35] arising from the need to satisfy informational uniqueness, since semantic uniqueness and weak familiarity together entails informational uniqueness. In this particular case, the weak familiarity presupposition is satisfied by interpreting *the dashboard* in connection with *this car*. The common knowledge that a particular car can have at most one dashboard also guarantees that there cannot be an additional dashboard on the common ground that also belongs to the car, therefore satisfying the informational uniqueness presupposition. In more complicated cases like (3.34)[36] where accommodation is required to satisfy the familiarity presupposition, a receiver will interpret the definite description in question as 'meant to be semantically unique'.

(3.34) I found a box in my attic the other day. I opened the lid and pushed *the button I found inside*. You won't believe what happened. Roberts (2003, ex. 4)

Roberts's (2003) analysis clears the way for analyzing the majority of Hawkins' (1978) categories under Löbner's (1985) functional framework while maintaining familiarity as one of the fundamental presuppositions of definite descriptions. If one adopts the functional view, the inevitable question is then what exactly those functions are. Löbner answers the question in a descriptive manner. Table 3.2 provides a brief overview of the weakly-familiar definite descriptions with regard to the sources of the functions and their parameters.

There are a number of ways for a definite description to acquire a functional reading. Proper names and the likes (including instances of the categories Unfamiliar-Complement[37] and Unfamiliar-Nominal) are usually understood as unique in any given situation[38]. The Immediate category is also analyzed as independent, since from the receiver's point of view, the discourse referents have to

[35] Roberts (2003) gives a different analysis for titles (e.g. '*the Ohio State University*') and regards them as having an epistemic version of semantic uniqueness.

[36] There is no real difference between this example and (3.5). It is selected because it clearly requires accommodation to be felicitous.

[37] This is Löbner's (1985) point of view. Another way to see this category is that the head nouns are used predicatively. Most, if not all of the instances in this category can be paraphrased into *it*-extrapositions and copula constructions. For example, *the fact that* ... can be paraphrased as *it is a fact that* ..., or equivalently *that* ... *is a fact*.

[38] This does not mean that '*John*' refers to a unique person in the world, but rather that it is generally understood as a unique individual. Determining the identity of John is another issue, which may be addressed through 'pragmatic enrichment' (cf. Roberts, 2003). For example, one may regard the individual as 'the person named John that appeared in the article / mentioned by Bill' etc.

	Argument(s)				Function Specification	
Use	Situation	Discourse	Perception	Self	Gen. Knowledge	Self
Deictic	X		X			
Immediate						
LSU						
-Specific	X					
-General	X				X	
Associative		X			X	
Unfamiliar						
-Complement						
-Nominal						
-*of*-Prep				X		X
-Pragmatic				X		X
Adjectival	?[a]	?[a]	?[a]			X

Table 3.2: Functions and arguments of weakly-familiar definite descriptions. Situation and General World Knowledge are only marked when their role is prominent.

[a] Different subtypes of the Adjectival category have different sources of arguments. See discussion on the Adjectival class (page 53) for details.

be setup via accommodation, which gives rise to the uniqueness effect. Some times the signal for functional reading[39] is provided directly with a modifier, as in the cases of Unfamiliar-*of*-Prep, Unfamiliar-Pragmatic, and Adjectival. In addition, there are also times when the decision for functional reading has to be derived from general knowledge, as in the case of the Associative category. Taking (3.33) for example, the functional reading for *the dashboard* comes from the fact that each car has at most one dashboard. Instances of the LSU-General category are somewhat ambiguous[40] as to whether they are more like the LSU-Specific cases or rather similar to the Associative cases. The category is marked as relying on general knowledge considering cases like (3.25).

One important thing to note is that not all categories are homogenous from the semantic point of view, and Table 3.2 only reflects the typical use of the respective syntactic constructs. For example, both Hawkins (1978) and Löbner (1985) note that instances of the Unfamiliar-Pragmatic category may be used anaphorically when there is a strongly familiar discourse referent that meets the prescribed descriptive contents (we shall return to this point later). Similarly, an instance of the Adjectival category may be used to single out one of the previously-established discourse referents. Even the seemingly homogenous category of Unfamiliar-*of*-Prep contains instances such as '*the threat of U.S. retaliation*', which are rather similar to proper names.

[39] As noted by Löbner (1985), many nouns have both sortal and relational readings. It is often the case that the head nouns in Unfamiliar-*of*-Prep are inherently functional and do not really need a 'trigger' for functional reading. However, there are also cases like '*the car of my uncle*', where functional reading is necessitated by the preposition.

[40] Löbner (1985) faced a similar dilemma.

3.2.3 A New Classification Scheme

Setting aside the difficulties mentioned earlier, there does seem to be a common theme behind the definite descriptions that are capable of discourse-new use – they do not rely on any other discourse entities to satisfy the weak familiarity and informational uniqueness presuppositions. In other words, they are not anaphoric. Gathering from Tables 3.1 and 3.2, I propose the following classification scheme for definite descriptions:

De Facto Proper Names (PN)

This category contains the definite descriptions that behave like proper names, i.e. they are weakly familiar and are understood as semantically unique without heavy reliance on any particular situation. This class covers titles (3.35), the definite descriptions denoting kinds (3.36), instances of the Unfamiliar-Complement category (3.37), and the Unfamiliar-Nominal instances (3.38).

(3.35) The company has $1 billion in debt filed with *the Securities and Exchange Commission*. WSJ 351:29

(3.36) It's probably true that many salarymen put in unproductive overtime just for the sake of solidarity, that the system is so hierarchical that only *the assistant manager* can talk to *the manager* and *the manager* to *the general manager*, and that Sony was chary of letting a young, short-term American employee take on any responsibility. WSJ 37:31

(3.37) Others grab books, records, photo albums, sofas and chairs, working frantically in *the fear that an aftershock will jolt the house again*. WSJ 766:19

(3.38) The dealership dutifully recorded the sale under *the name "Judge O'Kicki."* WSJ 267:68

A common issue among generic definite descriptions is that the types are often further restricted by other discourse elements, creating an 'associative' kind of effect. For example, *the assistant manager* in (3.36) is obviously referring to the assistant managers in *the (Japanese management) system*. However, this kind of association does not warrant the generic definite descriptions to be interpreted as associative anaphoric, since it only makes them more specific, in the same way *assistant* affects *manager*, but does not change the fact that they are generic and behave like proper names.

The category also covers cases like (3.39), where the head is accompanied by a proper name denoting an entity of the same kind. In (3.39), the head noun *machine* serves to clarify what *Cray-3* is upon its first introduction.

(3.39) The documents also said that although the 64-year-old Mr. Cray has been working on the project for more than six years, *the Cray-3 machine* is at least another year away from a fully operational prototype. WSJ 18:4

In addition, cases such as (3.40) and (3.41) as well as a part of the Adjectival class[41] also belong to this category. From the semantic point view, (3.40) and (3.41) closely resembles (3.37) and (3.39), respectively, despite their overt syntactic structure[42].

(3.40) Gary Hoffman ... said *the threat of U.S. retaliation*, combined with a growing recognition that .... WSJ 20:15

(3.41) The Soviet orders were compressed into *the month of October* because of delays. WSJ 192:35

Semantically Unique Definite Descriptions with Restrictive Modifications (RM)

This syntax-based category contains the definite descriptions that are interpreted as semantically unique due to restrictive nominal or adjectival pre-modification, and/or post-modification by preposition or restrictive relative clause. Most of the instances from the previous-mentioned Unfamiliar-*of*-Prep and Unfamiliar-Pragmatic classes belong to this category. Typical examples of the category include:

(3.42) Regarded as *the father of the supercomputer*, Mr. Cray was paid $600,000 at Cray Research last year. WSJ 18:33

(3.43) Mr. Nixon is traveling in China as a private citizen, but he has made clear that he is an unofficial envoy for *the Bush administration*. WSJ 93:19

(3.44) *The Polish government* increased home electricity charges by 150% ... WSJ 37:59

(3.45) "*The secret to being a good adjuster* is counting," says Gerardo Rodriguez, an Aetna adjuster from Santa Ana. WSJ 766:42

(3.46) Alan F. Shugart, currently chairman of Seagate Technology, led *the team that developed the disk drives for PCs*. WSJ 22:14

Note that the expression *the Bush administration* in (3.43) has a functional reading that is essentially the same as paraphrases like '*(President George H. W.) Bush's administration*', or '*the administration of Bush*'. In comparison, the similarly-constructed '*the Vichy government*' (WSJ 39:31) lacks an internal information structure and is considered as a de facto proper name.

Also included in the category are members of the Adjectival class that have accompanying restrictive post-modifications, such as (3.47):

(3.47) The declaration by Economy Minister Nestor Rapanelli is believed to be *the first time such an action has been called for by an Argentine official of such stature*. WSJ 21:2

[41] More specifically, these are the type II and type V (page 55) instances. See discussion on the Adjectival class (page 53) for details.

[42] In fact, examples (3.37) through (3.41) are all appositions of one kind or another (cf. Quirk, Greenbaum, Leech, & Svartvik, 1985, sections 17.65-17.93).

Despite that Löbner's (1985) treats Unfamiliar-*of*-Prep and Unfamiliar-Pragmatic classes differently, they are put in the same category due to the similarity between them with regard to familiarity requirements and sources of functions and their arguments. In fact, although in theory restrictive post-modifications can be used to distinguish between a number of strongly-familiar discourse referents in the same way that '*the new company*' identifies a fledgling firm, data from corpus analysis seem to indicate that such constructs would not be the preferred method. It is difficult to think of an example where a disambiguating attribute used to pick up a particular individual from a limited set cannot be expressed in 'simpler' forms such as an adjectival or a nominal pre-modifier. On the other hand, there are examples like (3.34), in which the lack of even weak familiarity leads to uniqueness effect. Finally, there are also cases in the Unfamiliar-Pragmatic class, such as (3.48), that are guaranteed to be semantically unique because the heads denote the how, why, where, or when of events (see also Quirk et al., 1985, section 18.30).

(3.48) I mention the picture only because many bad movies have a bright spot, and this one has Gregory Peck, in a marvelously loose and energetic portrayal of an old man who wants to die *the way he wants to die*. WSJ 39:42

The uniqueness effect in definite descriptions post-modified by prepositions other than *of* also seem reasonably robust, with the notable exception of *by*. The exception is quite understandable: in general, it is unlikely that an individual or organization only creates one thing of a particular sort. There are only two instances of definite descriptions with *by*-prepositions in the portion of WSJ corpus annotated in this study, both of which are used to supply additional information, as shown in (3.49). An extended search in the corpus text reveals that this seems to be the typical use of definite descriptions with *by*-prepositions[43].

(3.49) 45. <u>A marketing study</u> indicates that Hong Kong consumers are the most materialistic in the 14 major markets where the survey was carried out.

46. *The study by the Backer Spielvogel Bates ad agency* also found that the colony's consumers feel more pressured than those in any of the other surveyed markets, which include the U.S. and Japan. WSJ 37:45-46

Digression: The Problematic Adjectival Class

As much as I would like to subsume the whole Adjectival class under the previously discussed category of restrictive modifications, there are practical difficulties that disallow this: the Adjectival class is simply not coherent. Consider the following excerpt:

(3.50) 68. The next day, as she ... she jumps at *the slightest noise*.

[43] I have reviewed all 19 instances of the entire WSJ corpus matching the pattern `the noun by the`. Only 2 cases are semantically unique; one of which is due to the presence of additional post-modification, the other one (WSJ 1852:4) has the head noun *cooperation*. One (WSJ 2102:34) of the remaining 17 cases is best interpreted as associative; the others are anaphoric (some of them have events as antecedents). Of the 16 anaphoric cases, 12 supply new information in one way or another.

69. On further reflection, she admits that venturing inside the Hammacks' house the previous day wasn't "such a great idea."

70. During her second meeting with the Hammacks, Ms. Johnson reviews exactly what their policy covers.

71. They would like to retrieve some appliances on *the second floor*, but wonder if it's safe to venture inside. WSJ 766:69-71

(3.50) contains two instances of the Adjectival category, *the slightest noise* in sentence 68, and *the second floor* in sentence 71, with the former being discourse-new and the latter best interpreted as associative. Obviously, there is a second x in any set X with a cardinality greater than one, and *second* provides the function to single out x_2. Then, why do we feel compelled to interpret *the second floor* in relation to *the Hammacks' house*? There are a number of possibilities here: (a) *floor* has a relational reading; (b) *the second floor* needs to be relevant; and (c) a weakly-familiar set of ordered floors needs to be present in order for the definite description to be felicitous. It is even possible that all three are in effect here. However, the first two are a bit vague, because *floor* is not inherently relational (as in *wife*), and to my best knowledge there does not seem to be a rigorous and computationally executable standard for relevance. This leaves (c) the only viable path for now.

The biggest obstacle on the path is the contrast between *the slightest noise* and *the second floor*. I believe the difference has its root in that *noise* (at least when used with superlatives) is uncountable, while *floor* is countable. Uncountable concepts are already weakly familiar and understood as continuous functions whose values can be compared and may have minimums and maximums in a given situation. On the other hand, a definite description with countable head noun demands a weakly familiar set of the same sort, regardless of whether the head is combined with a superlative (e.g. '*the tallest girl*') or an ordinal.

Unfortunately, there are also examples like (3.51), in which the set consists of all chickens ever existed.

(3.51) *The first chicken* was hatched from an egg that was not a chicken's egg.

Hurford, Heasley, and Smith (2007, page 85)

By itself, (3.51) does not pose a threat to the previously discussed familiarity hypothesis: a set of all chickens ever existed on earth is readily conceivable and can be said as weakly familiar. The problem is why *the second floor* does not receive the same treatment. I admit that I have no definite answer at this moment, although it is possible that unlike floors in a building, chickens do not have a default attribute that can be used for ordering, leaving time as the only option available.

Finally, (3.52) provides a slightly twisted version of another common Adjectival use – singling out elements from a set that is already strongly familiar.

(3.52) 01. Japanese investors nearly single-handedly bought up two new mortgage securities-based mutual funds totaling $701 million, the U.S. Federal National Mortgage Association said.

02. The purchases show the strong interest of Japanese investors in . . . .

03. He said more than 90% of *the funds* were placed with Japanese institutional investors.

04. The rest went to investors from France and Hong Kong.

05. Earlier this year, Japanese investors snapped up a similar, $570 million mortgage-backed securities mutual fund.

06. That fund was put together by Blackstone Group, a New York investment bank.

07. *The latest two funds* were assembled jointly by Goldman, Sachs & Co. of the U.S. and Japan's Daiwa Securities Co. WSJ 29:1-7

The initial sentence of (3.52) mentions two funds, which is referred to in sentence 3 as *the funds*. Sentence 5 mentions yet another fund, which serves as the antecedent of *That fund* in the immediately-following sentence. This brings the total number of strongly-familiar funds in the common ground up to three. In sentence 7, the author uses *The latest two funds* to refer to the two funds mentioned initially. Had the two funds not been introduced explicitly, (3.52) would have been a typical example of the 'element-in-a-strongly-familiar-set' kind. But now there are two different ways to satisfy the presuppositions of definiteness: a strongly-familiar set coupled with the function provided by the superlative, and a strongly-familiar discourse referent that has matching descriptive contents. While the former seems aesthetically more appealing because it follows the same line of reasoning applicable to the more typical cases of the kind, there are also researchers who advocate the latter strategy (e.g. the 'If Possible Use Identity' proposed by Asher & Lascarides, 1998)[44]. For the purpose of classification, however, I believe it is reasonable to adopt the first strategy; the latter strategy can be implemented as a preference in the anaphora resolution procedure.

To summarize, thus far five different subtypes of the Adjectival class have been observed:

(I) `the + adj + noun with restrictive modification`, as in '*the fastest person to sail to America*';

(II) `the + adj + uncountable noun`, as in '*the slightest noise*';

(III) `the + adj + countable noun with strongly familiar set`, as in '*chickens . . . the first chicken*'; and

(IV) `the + adj + countable noun with set obtained through association`, as in '*the house . . . the second floor*'.

(V) `the + adj + countable noun with universal set`, as in '*the first chicken was hatched from an egg that was not a chicken's egg.*'

[44] Psycholinguistic studies on the online processing of definite descriptions (cf. Schumacher, 2008, and the cited research therein) suggest that associative anaphora generally has higher processing costs in comparison with 'direct' anaphora, both in terms of the cost associated with information retrieval and inferencing and the cost of establishing new discourse referent. This can be seen as collateral evidence that supports the later strategy. However, it must be noted that the first strategy does not necessarily lead to establishing new discourse referent, as further reasoning may identify that *The latest two funds* has the same referent as the initially-mentioned two funds – but this kind of reasoning is beyond the scope of anaphora resolution.

Type I has already been subsumed under the RM category (page 52). If we ignore the internal structures of the type II and type V instances, they can be considered as proper names as well. Following H. H. Clark (1975), type III instances are associative (set-membership bridging). By definition, type IV instances are also associative. However, it is worth noting that when the anchor itself is originated in the situation rather than the discourse, as in (3.53), the instance should be labeled LSU.

(3.53) Norton Co. said net income for *the third quarter* fell 6% to $20.6 million, or 98 cents a share, from $22 million, or $1.03 a share.
Poesio and Vieira (1998, ex. 18b, WSJ 760:1)

Semantically Unique Definite Descriptions Bound by Situation (BS)

This category covers instances of the Deictic class and the two LSU classes. Regardless of whether the receiver has specific knowledge about the intended referent, the common theme behind these uses is that there is a relevant domain, derived from factors not directly related to the discourse, in which the unique existence of the referent is known by the interlocutors. The size of the domain can vary greatly – some objects, such as the world or the sun, have such a large domain that they can almost be deemed as proper names, while others, especially those referred to deictically, have much smaller domains. However, there is always a possibility that uniqueness can no longer be guaranteed in a domain larger than the relevant one.

Treating the three uses uniformly allows us to move focus away from some of the minor but often distracting details. For example, Löbner's (1985) 'bedroom bottle' (3.54) no longer needs different treatments depending on whether the bottle is always there or merely present in the bedroom by accident.

(3.54) John and Mary (believe it or not) always have a bottle of mineral water beside their common bed. One night, John is already sleeping. Mary wakes up and feels thirsty. She fumbles for the bottle in the darkness, but cannot find it and wakes up poor John:
John, would you pass me *the water*, please? Löbner (1985, ex. 42)

The clause 'not directly related to the discourse' is needed in order to prevent some (perhaps undesirable) generalizations of the LSU-general category. For example, *the government* in (3.55) can no longer be interpreted as LSU-general, since the relevant domain for *the government* only covers the U.S. government.

(3.55) 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, ex. 18c, WSJ 761:6-7)

The Immediate use is similar to the Deictic use in the sense that it is also bound by situation. The main difference between them is that the former requires the addressee to accommodate the existence of the referent. Once the accommodation is successful, however, the distinction between the two becomes less obvious. Since this study is mainly concerned about written text, I will simply ignore the differences between Immediate and Deictic uses.

Direct Anaphoric (DA)

The Direct Anaphoric category contains the definite descriptions that a) are not members of the PN (page 51) or RM category, and b) have strongly-familiar discourse referents matching their descriptive contents. An anaphor of this category may have the same head noun with its antecedent, or one of them may be a synonym, hypernym, or epithet of the other. In addition, an anaphor with abstract head noun denoting actions can also have a clause as its antecedent.

Definite descriptions of this category are coreferential to their antecedents. However, as discussed earlier, not all coreferential relationships involve anaphora. Definite descriptions that can have their weak familiarity and informational uniqueness presuppositions satisfied independent of a previous mention or the situation do not belong to this category[45]. It follows that 'discourse-new' is no longer an appropriate term in the context of this study. Instead, distinction is made between definite descriptions that are anaphoric and the 'discourse-new-capable' ones, or simply put, between anaphoric and non-anaphoric definite descriptions.

Associative Anaphoric (AA)

This category contains the definite descriptions that a) are not members of the PN or RM category, and b) have merely weakly-familiar discourse referents matching their descriptive contents.

3.2.4 The Gray Areas

In the previous section I have presented an attempt to analyze the multifarious linguistic phenomena pertaining to definite descriptions from perspectives of both uniqueness and familiarity and produced a categorization scheme for definite descriptions. While special care has been taken to make the boundaries among the categories more clear, I must admit that there are still some gray areas left. On the other hand, the effort of demarcating the boundaries also creates certain undesirable 'artifacts'. For example, '*the sun*' would be interpreted differently according to whether the head word is capitalized. This section discusses some of the known issues related to the proposed scheme.

Methodological Issues

The first, and perhaps the most obvious gray area in the current categorization scheme is that the last three categories are not mutually exclusive. As attested by both Fraurud (1990) and Poesio and Vieira (1998), the same issue is also present in Hawkins's (1978) scheme. Annotators often have difficulties agreeing on whether a specific discourse referent should be interpreted as originated from the situation or from the discourse, or whether it should be interpreted as associative or

[45] Similarly, subsequent mentions of a BS instance should not be interpreted as anaphoric.

direct anaphoric. These are real problems, but the answers cannot, and should not be provided by a categorization scheme, because the ultimate choices have to be made based on a number of complex factors, one of which being personal preferences.

The situation-vs-discourse problem is well-illustrated by (3.56), in which *the country* is categorized as discourse-new by one annotator and bridging[46] by another in Poesio and Vieira's (1998) first annotation experiment.

(3.56) The missing watch is emblematic of the problems Mr. Wathen encountered in building his closely held California Plant Protection Security Service into the largest detective and security agency in the U.S. through acquisitions.
. . . (other 5 sentences) . . .
Over the next 20 years, California Plant Protection opened 125 offices around *the country*.
Vieira (1998, ex. 3.8, WSJ 305:5-10)

Obviously, both readings are acceptable because ultimately the definite description points to the same discourse referent. However, rather than blaming on the non-overlapping categorization scheme, one must realize that the problem has its root in the artificial distinction we are drawing between discourses and the situations they are bound to. As suggested by Fraurud (1990), the problem is of methodological nature and cannot be solved categorically. One could, however, circumvent the problem procedurally, for example by giving priority to either interpretation or by marking both as acceptable.

Similarly, most of the associative-vs-direct anaphoric issues are also of methodological nature: once both presuppositions are satisfied, an associative anaphor often becomes coreferential with any strongly-familiar discourse referent that also matches its descriptive contents. For the purpose of corpus annotation and anaphora resolution system development, it is probably desirable to have a specific standard that prefers direct anaphora over associative anaphora due to the relative ease of resolving the former.

Having said that, some of the associative-vs-direct anaphoric issues are more complicated and require special attention. For example, consider *the debt problem* in (3.57):

(3.57) Argentina said it will ask creditor banks to halve its foreign debt of $64 billion – the third-highest in the developing world.
. . .
Mr. Rapanelli recently has said the government of President Carlos Menem, who took office July 8, feels a significant reduction of principal and interest is the only way *the debt problem* may be solved.
WSJ 21:1-7

At least four different readings for *the debt problem* are possible: associative with *the government of President Carlos Menem*, *Argentina*, or *its foreign debt of $64 billion* as anchor, or consider

[46] Poesio and Vieira's (1998) definition of bridging also covers certain coreferential relationships.

the non-head component *debt* as direct anaphoric with *its foreign debt of $64 billion* and at the same time mark *the debt problem* as non-anaphoric because it is semantically unique. Vieira (1998, section 4.5.3) follows the third interpretation. Personally, I prefer the last one because it reflects the internal structures of the definite description in question. But regardless of the choice, the solution of the problem boils down to a detailed specification of the category boundaries, preferably at the implementation level.

Descriptions with Restrictive Modifications

Although restrictive modifications often lead to independent satisfaction of both weak familiarity and informational uniqueness presuppositions, there are always exceptions[47]. In both (3.58) and (3.59), the head nouns are clearly functional, but the post-modifications do not provide adequate support. In fact, the relative clause in (3.59) is arguably non-restrictive.

(3.58) No fewer than 24 country funds have been launched or registered with regulators this year, triple *the level of all of 1988*, according to Charles E. Simon & Co., a Washington-based research firm. WSJ 34:3

(3.59) In Robert Whiting's "You Gotta Have Wa" (Macmillan, 339 pages, $17.95), the Beatles give way to baseball, in *the Nipponese version we would be hard put to call a "game."* WSJ 37:19

However, things can become tricky when information needed to achieve semantic uniqueness is split between the modification and the context. For example, nouns denoting events are often only relational but not functional against one of its arguments. Consequently, definite descriptions with such head nouns may need to be interpreted as anaphoric even though they are restrictively modified, as in the case of *the withdrawal of New England Electric* in (3.60).

(3.60) it was just another one of the risk factors" that led to the company's decision to withdraw from the bidding, he added.
Wilbur Ross Jr. of Rothschild Inc. . . . said *the withdrawal of New England Electric* might speed up the reorganization process. WSJ 13:12-13

Another interesting example (3.61) appears later in the same article, but with a sortal head noun modified by a restrictive relative clause.

(3.61) Northeast said it would refile its request and still hopes for an expedited review by the FERC so that it could complete the purchase by next summer if its bid is *the one approved by the bankruptcy court*. WSJ 13:17

Setting aside the issue of *one*-anaphora in (3.61), both *the withdrawal of New England Electric* and *the one (bid) approved by the bankruptcy court* can be interpreted as associative anaphoric using

[47] The RM category does not cover the exceptions.

the previously mentioned *the bidding* as anchor. Each of the two definite descriptions also has an alternative reading. The former can be interpreted as direct anaphoric with *to withdraw from the bidding* as antecedent[48], and the latter may be deemed non-anaphoric – although it is not likely that a court only proves one bid, the fact that the definite description is used as predicate nominal relaxes the requirement for uniqueness[49]. As mentioned in the earlier section, the ambiguity between direct and associative reading can only be solved procedurally. In comparison, the dual-reading issue of (3.61) is relatively easy to solve. Since the RM category and the AA category are mutually exclusive by definition, the problem goes away as soon as one decides on the acceptability of the extension.

On the flip side, a definite description interpreted as non-anaphoric may be coreferential with another discourse entity. From an anaphora resolution system's point of view, it is entirely accidental. However, as mentioned earlier, recognizing some of the coreferential relationships also helps anaphora resolution. Gathering from the corpus, there are three situations where non-anaphoric definite descriptions often become coreferential. The first, and most reliable situation is the subsequent mentions of certain de facto proper names, such as '*the SEC*' against the previously-mentioned '*the Securities and Exchange Commission*'. Sometimes a definite description with restrictive premodification is used in sentences immediately following a previous mention to give the referent a 'name', as in (3.62):

(3.62) In a victory for environmentalists, Hungary's parliament terminated a multibillion-dollar River Danube dam being built by Austrian firms. *The Nagymaros dam* was designed to be twinned with another dam, now nearly complete, 100 miles upstream in Czechoslovakia.
WSJ 37:62-63

Note how (3.62) closely mirrors (3.63), in which a named entity is used for the same purpose:

(3.63) A painting by August Strindberg set a Scandinavian price record when it sold at auction in Stockholm for $2.44 million. *"Lighthouse II"* was painted in oils by the playwright in 1901... WSJ 37:68-69

Identifying the driving force behind the coreferential interpretation of such cases is beyond the scope of this study. However, from (3.63), it is clear that the force must be different from what is behind the anaphoric cases. Finally, there are also cases like (3.11) on page 35, in which the effort to bring a 'backgrounded' discourse referent back to focus leads to the non-anaphoric expression *The Victorian house that Ms. Johnson is inspecting*. Another interesting example is (3.64), in which the non-anaphoric *Dinkins campaign* (sentence 87) is most likely a side effect of disambiguation from the two presidential campaigns mentioned in the preceding sentence.

[48] There is another probably more appropriate candidate for antecedent in the first sentence of the article, '*New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire ...*'

[49] See ensuing discussion of quasi-nonanaphoric definite descriptions for more details about the licensing condition for non-unique predicate nominal.

(3.64) 70. On some occasions when Mr. Dinkins has discussed the issues during the campaign, he has run into a familiar kind of trouble.

…(continues to discuss Dinkins' policies, with no mention of campaign) …

84. Mr. Dinkins's inner circle of advisers appears to include both ideologues and pragmatists, leaving voters with little clue as to who will be more influential.

85. The key man seems to be the campaign manager, Mr. Lynch.

86. …Mr. Lynch is a veteran union organizer who worked on the presidential campaigns of Sen. Edward Kennedy and Mr. Jackson.

87. But as the *Dinkins campaign* hit tough times this month, Andrew Cuomo, the politically seasoned son of the New York governor, is also … WSJ 765:70-87

Situations like (3.11) and (3.64) do not automatically lead to non-anaphoric definite descriptions, although they are often 'accidentally' used in these situations. From this perspective, the narrower definition of anaphora developed in this study may seem suboptimal. Nevertheless, I consider it more important to separate the definite descriptions whose interpretation depend on other discourse entities and the ones that merely share the same referents with other discourse entities. Recognizing this difference allows different strategies to be applied[50] to the different kinds of phenomena that were uniformly labeled as coreferential.

The Problematic Adjectival Class, Continued

In the previous discussion on the definite descriptions pre-modified by functional adjectives, I have shunned away from *next*, *other*, and *same*[51]. Applying weak familiarity requirement on instances containing these adjectives sometimes leads to unconventional interpretations. For example, consider *the other two outside bidders* in (3.65):

(3.65) 12. it was just another one of the risk factors" that led to the company's decision to withdraw from the bidding, he added.

…

14. The fact that New England proposed lower rate increases – 4.8% over seven years against around 5.5% boosts proposed by *the other two outside bidders* – complicated negotiations with state officials, Mr. Ross asserted. WSJ 13:12-14

Traditional wisdom tells us to associate the expression with *New England (Electric System)* in the same sentence. This interpretation fills the argument required by the function of *other* – "those N(s) other than xy" (Löbner, 1985). However, it tells little about where the discourse referents for *two outside bidders* come from. More interestingly, it seems to work reversely and project the sort '*bidder*' onto *New England*. The weak familiarity presupposition can only be satisfied by association

[50] Consider cases like (3.62), for example, since the coreferential relationship is essentially implied by the context, the antecedent would most likely be found in an immediately preceding sentence.

[51] *Only* can be analyzed under the same framework as previously discussed.

with *the bidding* in sentence 12[52]. Once the familiarity requirement is met, however, no further action needs to be taken, since the adjective already offers a functional reading (i.e. informational uniqueness is satisfied).

Cases containing *next* and *same* are pestered by exactly the same issue – whether to follow intuition and try to identify the function arguments or to satisfy the presuppositions of definiteness. In my view, the issue is not isolated but rather a manifestation of the difference between definiteness and determinedness (Löbner, 1985, p. 303). Neither Löbner's (1985) theory nor that of Roberts (2003) is concerned about overall determinedness of the referent. I believe the same applies to anaphora, at least with regard to the notion developed in this study.

Quasi-Nonanaphoric Definite Descriptions

Aside from the definite descriptions that are semantically unique, there are a few special conditions under which anaphoric readings do not seem to be necessary. However, whether to accept these cases as non-anaphoric is essentially a personal preference, and hence the term 'quasi-nonanaphoric'. Below is a list of some quasi-nonanaphoric conditions identified in this study:

(I) Definite descriptions functioning as predicates

Löbner (1985) describes a special class of definite descriptions, the configurational uses, that do not satisfy sortal uniqueness. (3.66) is an often-cited example of the class:

(3.66) He was *the son of a poor farmer*. Löbner (1985, ex. 17)

Löbner further specifies that these definite descriptions are not referential (which is obvious for predicate nominal, as in (3.66), but less so in some other configurational uses) and that the head nouns need to be relational. I believe the second requirement can be relaxed to include sortal head nouns that acquire relational readings through modifications as well, as shown in (3.67).

(3.67) She is *the girl I met last week*.

As Abbott (2001) points out, in cases like (3.66) and (3.67), there is no need to use the definite descriptions to identify an individual in order to interpret the propositions expressed, since the individuals have already been identified. Therefore it seems reasonable to overlook the semantic uniqueness requirement of restrictively-modified[53] predicate nominals and simply regard them as non-anaphoric.

Aside from the previously discussed example (3.61), there is also another interesting case in the corpus that is closely related to this type, as shown below:

(3.68) Much of Mr. Lane's film takes a highly romanticized view of life on the streets (though probably no more romanticized than Mr. Chaplin's notion of the Tramp as

[52] There is also a strongly-familiar discourse referent established earlier in the text. But it is not important for the purpose of this discussion.

[53] It seems that restrictive modification is generally required in order to fulfill the familiarity requirement.

the good-hearted free spirit). WSJ 39:10

It seems that the author assumes that the readers are already familiar with the Tramp's characteristics (presumably through personal experience by viewing one of Chaplin's films), otherwise it would be difficult to justify the use of the definite article.

(II) Definite descriptions in cleft-like sentences

Definite descriptions such as *the problem* in (3.69) have been previously identified by Vieira (1998, section 4.4.4) and was subsumed under the 'copula constructions' heuristics for identifying discourse-new definite descriptions. While I am suspicious that Vieira's (1998) copula heuristics is too broad, definite descriptions in (3.69) and their likes can be genuinely ambiguous between anaphoric and non-anaphoric. Following Doherty (2001), whose analysis provides much insight into the construct, a sentence like (3.69) is called a 'cleft-like sentence', which is characterized by a pre-copula abstract noun and a post-copula structure of a *that*-clause, a *wh*-clause, or a full clause without complementizer.

(3.69) If the answers to these questions are affirmative, then institutional investors are likely to be favorably disposed toward a specific poison pill.
However, *the problem* is that once most poison pills are adopted, they survive forever.
WSJ 275:13-14

In (3.69), *the problem* does not have a referential reading. One could certainly associate it with the previous sentence, but doing so does not give rise to semantic uniqueness (there are always more than one problems associated with anything) and consequently cannot introduce a weakly-familiar discourse referent for *the problem*. However, if the sentence is treated as a pseudo-cleft, which carries existential and exhaustive presuppositions (cf. Hedberg, 2000; J. K. Gundel, 1977), semantic uniqueness of *the problem* can be obtained for free, as illustrated in (3.69′)[54].

(3.69′) However, *what is problematic* is that once most poison pills are adopted, they survive forever.

Essentially, (3.69′) means that there is something and only one thing that is problematic, and that problematic thing (the problem) is what is provided by the *that*-clause.

While *the problem* in (3.69) does not have a referential reading, certain definite descriptions in this category can be interpreted as associative anaphoric or even direct anaphoric. For example, *The bottom line* in (3.70) may be interpreted as associative anaphoric using both *Malcolm Hoenlein* and *Mr. Dinkins* as anchors.

(3.70) Mr. Giuliani is finding that Mr. Dinkins, in his many years in public life, has built up considerable good will that so far has led many voters to overlook certain failings.

[54] Only some cases in this category can be nicely paraphrased into a pseudo-cleft. Nevertheless, I believe the same principle holds for all of them.

"*The bottom line* is that he is a very genuine and decent guy," says Malcolm Hoenlein, a Jewish community leader. WSJ 765:107-108

However, in light of cases like (3.69), it is best to consider the category non-anaphoric if it is to be treated as a whole.

(III) Definite descriptions with local anchors

Sometimes an associative anaphoric definite description has its anchor in the same clause. Theoretically speaking, this does not make them different from their non-local counterparts. However, having the presuppositions satisfied locally does give such cases the additional benefit of being discourse-new capable. One of Hawkins' (1978) original 'unexplanatory modifiers' examples finds its place here:

(3.71) My wife and I share *the same secrets.* Hawkins (1978, ex. 3.21)

As discussed earlier, definite descriptions with 'unexplanatory modifiers' often need to have their weak familiarity presuppositions satisfied through association as long as they are not restrictively modified. In the case of (3.71), the anchors are right there in the same sentence – almost everybody has secrets, so do '*my wife*' and '*I*'. There is also a similar case (3.72) in the corpus. Interestingly, annotators in Vieira's (1998) experiment could not reach an agreement on the case. One annotator identifies *the same neighborhood* as coreferential to *Oakland*, and another identifies it as associative on *the collapsed section of double-decker highway Interstate 880.* However, if *the same neighborhood* is considered as associative on *a new home*, which also implies the existence of an 'original' home that can serve as the other argument for *same*, the phrase '*build a new home in the same neighborhood*' becomes discourse-new capable as well.

(3.72) When Aetna adjuster Bill Schaeffer visited a retired couple in Oakland last Thursday, he found them living in a mobile home parked in front of their yard.
The house itself, located about 50 yards from the collapsed section of double-decker highway Interstate 880, was pushed about four feet off its foundation and then collapsed into its basement.
The next day, Mr. Schaeffer presented the couple with a check for $151,000 to help them build a new home in *the same neighborhood.* WSJ 766:49-51

3.3 Annotation Results

For the purpose of this study, a small part of the WSJ corpus, which has been originally used by Poesio and Vieira (1998) in their second experiment[55], is annotated according to the categories developed in Section 3.2. The following additional rules are enforced during the annotation:

[55] The data set includes articles 3, 13, 15, 18, 20, 21, 22, 24, 26, 29, 34, 37, 39, and 766 of the WSJ corpus.

- Type I and type II quasi-nonanaphoric definite descriptions are considered non-anaphoric. The former is subsumed under the RM category, and instances of the latter are treated as proper names.
- Idioms are considered proper names.
- The three potentially overlapping categories are processed in the order of BS $>$ DA $>$ AA, i.e. the definite descriptions bound by situations category has the highest priority, followed by the direct anaphoric category, and finally the associative anaphoric category.
- Due to practical restrictions of the annotation system, cases like (3.57) on page 58 are treated as associative anaphoric. There are only three such instances in the annotated corpus.

Table 3.3 summarizes the annotation results. Out of the 511[56] definite descriptions, 303 (59.3%) are non-anaphoric and only 208 (40.7%) are anaphoric. Many of the non-anaphoric definite descriptions have coreferential antecedents; their distribution is outlined in Table 3.4. Removing these 57 instances from the non-anaphoric group results in a ‘discourse-new’ ratio of 48.1%, which is largely in accordance to the figures reported by Vieira (1998) (47% for both data sets according to the standard annotation). In addition, personal pronouns are also annotated and the result is presented in Table 3.5.

[56] Vieira (1998) reports a total of 464 definite descriptions in the same data set, the source of the discrepancy is not clear.

Category	**Example**		**Items**
De facto proper names			147
Title	*the National Cancer Institute*	87	
Generic	life on *the streets*	24	
Idiom	on *the other hand*	10	
Apposition	*the shame of defeat*	6	
	the Backer Spielvogel Bates ad agency	6	
	the fact that . . .	3	
Adjectival	from *the most sympathetic angle*	3	
Quasi-nonanaphoric	*The reason*: Share prices of many . . .	2	
Others	*the 1920s*	6	
Unique definite descriptions with restrictive modifications			140
Of-preposition	*the end of 1990*	53	
Pre-modification	*the San Francisco area*	31	
Relative clause	*the damages caused by the earthquake*	29	
Other prepositions	*the interest rate on the refund*	14	
Mixed modifications	*the class of asbestos including crocidolite*	4	
Quasi-nonanaphoric	He isn't *the inheritor of Charlie Chaplin's spirit*	9	
Unique definite descriptions bound by situation			16
Time-related	*the past century*	9	
Location-related	*the country*, *the world*	6	
Immediate	*the steady afternoon rain*[a]	1	
Direct Anaphoric			157
Nominal antecedent	the . . . factory . . . *the plant*	147	
Event antecedent	Argentina said . . . *the declaration*	10	
Associative Anaphoric			51
Part-whole	the house . . . *the chimney*	8	
Topic antecedent[b]		4	
Embedded	its foreign debt . . . *the debt problem*	4	
Others	demolishing the house and clearing away *the debris*	35	
Non-anaphoric			303
Anaphoric			208
Total			**511**

Table 3.3: Profile of the definite descriptions in the annotated corpus

[a] There is no previous mention of weather condition in the article.

[b] Vieira (1998, section 3.4) uses the name 'Discourse Topic'. Although the corresponding discourse referent is only weakly familiar and is clearly licensed by the surrounding context, there is no overt anchor in the text.

Category		**Items**
De facto proper names		41
Title	38	
Generic	2	
Apposition	1	
Unique definite descriptions with restrictive modifications		15
Pre-modification	8	
Relative clause	6	
Other prepositions	1	
Unique definite descriptions bound by situation		1
Location-related	1	
Total		**57**

Table 3.4: Distribution of non-anaphoric definite descriptions with coreferential antecedents

Category			**Items**
First-person			28
Anaphoric	"*We* can lose money on this," he says.	23	
Generic	. . . the Japanese are more like *us* than most of *us* think.	3	
Deictic	*I* say "contained dialogue" because . . .	2	
Second-person			11
Generic	"It really brings *you* down to a human level," she says.	10	
Anaphoric	"And *you* . . . " says Mrs. Hammack, . . . tapping his hand.	1	
Third-person: *it*			108
Anaphoric	the company . . . *it*	91	
Extraposition[a]	*It*'s a shame their meeting never took place.	14	
Cleft[a]	But *it* is Mr. Lane . . . who has been obsessed . . .	1	
Local situation[a]	The ground shakes . . . *It* is an aftershock	1	
Idiom[a]	You either believe Seymour can do *it* again or you don't.	1	
Third-person: others			175
Single antecedent	Mr. Ross . . . *he*	172	
Split antecedent	she . . . her war-damaged husband . . . *their* home	3	
Anaphoric			290
Non-anaphoric			32
Total			**322**

Table 3.5: Distribution of personal pronouns

[a] See Chapter 5 for detailed discussion.

Chapter 4

Web-assisted Anaphora Resolution

The primary goal of this chapter is two fold: to design an anaphora resolution system that uses the web as its main source of information, and to apply the insight gained from the previous chapter to develop a definite description anaphoricity detector that is accurate enough to be used as a filter for anaphors.

Since many components of the system introduced in this chapter heavily exploits the web, we begin the chapter with a brief discussion on the characteristics of the web, how it differs from manually-compiled corpora, as well as some of the pitfalls that should be avoided. Sections 4.2 and 4.3 discuss the preprocessing component and the methods used to acquire the various information needed for anaphora resolution. The implementation details of the pronominal anaphora resolution subsystem, definite description anaphoricity detector, and the definite description anaphora resolution subsystem are discussed in Sections 4.4, 4.5, and 4.6, respectively.

4.1 Using the Web as a Corpus

The first question regarding using the web as a corpus is whether it can be regarded as a corpus at all. As Kilgarriff and Grefenstette (2003) pointed out, following the definition of corpus-hood that 'a corpus is a collection of texts when considered as an object of language or literary study', the answer is yes. With the fundamental problem resolved, what remains is to find out whether the web can be an effective tool for NLP tasks.

As a corpus, the web is far from being well-balanced or error-free. However, it has one feature in which no other corpus can be even remotely comparable – its size. No one knows exactly how big it is, but each of the major search engines already indexes billions of pages. Indeed, the web is so large that sometimes a misspelled word can yield tens of thousands of results (try the word *neglectible*). This sends out a mixed signal about using the web as a corpus: on the good side, even relatively infrequent terms yield sizable results; on the bad side, the web introduces much more noise than manually-compiled corpora do. In Markert and Nissim's (2005) recent study evaluating different knowledge sources for anaphora resolution, the web-based method achieves far higher recall ratio than those that are BNC- and WordNet-based, while at the same time yielding slightly lower precision. Similar things can be said about the web's diverse and unbalanced composition, which means that it can be used as a universal knowledge source – only if one can manage not to get overwhelmed by non-domain-specific information.

That being said, it is still very hard to overstate the benefits that the web offers. As the largest collection of electronic texts in natural language, it not only hosts a good portion of general world knowledge, but also stores this information using the very syntax that defines our language. In addition, it is devoid of the systematic noise introduced into manually-constructed knowledge sources during the compilation process (e.g. failure to include less frequent items or inflexible ways of information organization). Overall, the web is a statistically reliable instrument for analyzing various semantic relationships stored in natural languages by means of examples.

As also suggested by Kilgarriff (2007) and many others, it is technically more difficult to exploit the web than to use a local corpus and it can often be dangerous to rely solely on statistics provided by commercial search engines. This is mainly due to the fact that commercial search engines are not designed for corpus research. Worse, some of their design goals even impede such uses. For example, search engines skew the order of results using a number of different factors in order to provide users with the 'best' results. Combined with this is the fact that they only return results up to certain thresholds, making it essentially impossible to get unbiased results. Other annoyances include unreliable result counts, lack of advanced search features[1], and unwillingness to provide unrestricted access to their APIs. A recent, exciting development in the field is the release of the Google N-grams (Brants & Franz, 2006) data. The Google N-grams corpus is essentially a snapshot of the live web, and therefore affords the major benefit of the web as containing large amount of information. On the other hand, since search engines are out of the picture, their problematic behaviors are no longer an issue. Unfortunately, some inherent problems of the N-grams corpus cast limits on its overall usefulness. The first and most prominent issue is that the current release is limited to a maximum *N* of 5. In other words, it only provides counts for phrases containing 5 or less consecutive words. This limit probably also reflects what can be managed by present-day personal computers. As *N* grows, the number of unique expressions will quickly increase to the point that it is no longer possible to efficiently handle them with a personal computer. Another major problem is that the N-grams are limited within the sentence boundary, and no document-level information is provided. Therefore it is impossible to perform queries such as the ones proposed by Bunescu (2003) or to obtain information about phrase co-occurrence in the same document. In light of these difficulties, the Google N-grams can only be used as a supplement, not as a replacement, to the search engines. Before a new search engine specifically designed for corpus research is available, it seems we will have to work around some of those restrictions and live with the rest.

[1] For example, the wildcard ($*$) feature on Google, which could be immensely useful for query construction, no longer restricts its results to single words since 2003; Yahoo's ability to support alternate words within quoted texts is limited, while MSN does not offer that feature at all.

4.2 Preprocessing

The preprocessing component transforms the syntactic information embedded in natural language texts into machine-understandable structures. During the preprocessing stage, each word is assigned a part-of-speech tag, and the whole sentence is parsed using a dependency grammar (DG) parser. For simplicity's sake, the current system is designed to use the WSJ corpus, which is already tagged and parsed with context-free grammar (CFG). A head percolation table similar to that proposed by Collins (1999) is used to obtain the head component of each phrase. The rest of the phrase constituents are then rearranged under the head component to form the dependency tree using a procedure detailed by Xia and Palmer (2001). Figure 4.1 illustrates the syntactic structure of a sentence in the WSJ corpus. Both the original CFG parse tree and the derived dependency structure are shown side-by-side, with head entities underlined in the CFG diagram.

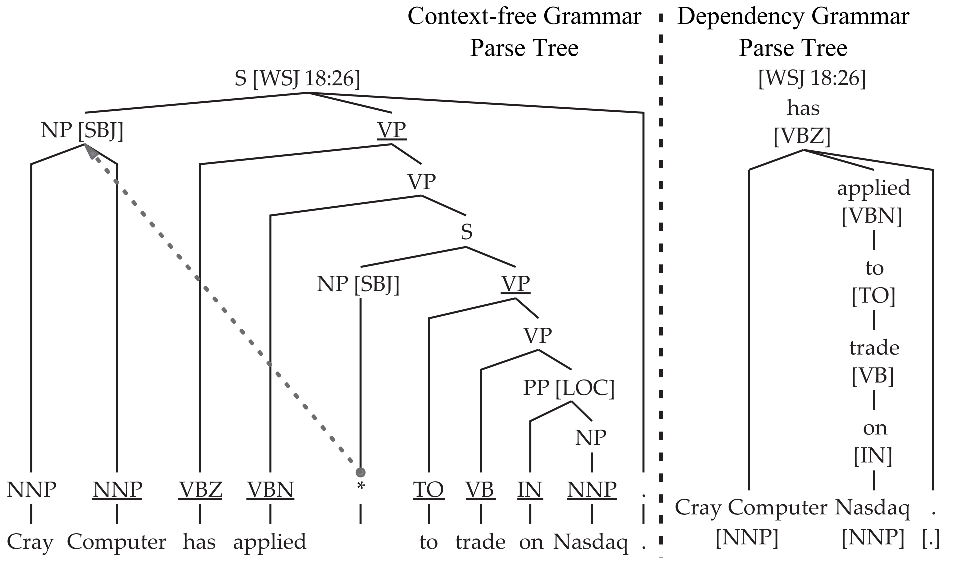

Figure 4.1: Illustration of a sentence's syntactic structure, both as annotated in the WSJ corpus (left) and after head percolation (right).

As shown in Figure 4.1, the function tags (e.g. SBJ and LOC), null elements, and tracing information present in the context-free parse tree are not ported to the dependency tree. This is because real-world parsers usually do not produce such tags. Except this deliberate omission, both parse trees contain essentially the same information, only presented in different manners. In this study, dependency structure is preferred over the more popular phrase structure mainly because of its explicit marking of both the head components and the complementing/modifying relationships among various components. This feature is very helpful for instantiating the web-based query patterns proposed in the rest of this study.

4.3 Acquiring Information

As briefly introduced in Section 2.2, number and gender agreement is a commonly used constraint among anaphora resolution systems. It plays a key role in pronominal anaphora resolution, since number and gender are the only descriptive contents available from pronouns. Consider the following examples, adapted from Mitkov (2002):

(4.1) John Bradley spoke to Jane McCarthy about the forthcoming project. *He* said this enterprise would cost millions. Mitkov (2002, ex. 2.2, adapted)

(4.1′) John Bradley spoke to Jane McCarthy about the forthcoming project. *She* said this enterprise would cost millions.

In these two examples, the interpretation of the pronouns are solely dependent on their gender. The two cases are unambiguous only because the name *John Bradley* is generally assumed to refer to a male person, and *Jane McCarthy* is assumed to refer to a female. With the additional knowledge that *the forthcoming project* cannot be referred to using either *he* or *she*, there remains only one plausible antecedent in each example. To a less extent, the tasks of name alias matching and definite description resolution also benefits from gender and number information. For example, although the distance between '*John Smith*' and '*Mr. Smith*' is the same as that between '*John Smith*' and '*Jane Smith*', only the former are valid aliases based on gender information.

Type information of proper names is equally important for definite description anaphora resolution. For example, in (4.2), knowing that *New England Electric* denotes a company is essential to the correct resolution of *the company*.

(4.2) John Rowe, president and chief executive officer of New England Electric, said *the company*'s return on equity could suffer if ... WSJ 13:6

Determining the gender, number, and type information of a noun phrase is not always a straightforward task. Often there are various heuristics from the text, which needs to be carefully balanced and combined in order to make the final decision. Depending on the source, clues can be categorized as either intrinsic or external. Intrinsic clues, such as that '*John*' is likely the name of a male, are directly available from the nominal expression in question. External clues are information chained from other expressions in the ambience. For example, in (4.3), *J.P. Bolduc* can be identified as the name of a person who is most likely male, because of the appositive *vice chairman*.

(4.3) J.P. Bolduc, vice chairman of W.R. Grace & Co., which holds a 83.4% interest in this energy-services company, was elected a director. WSJ 5:1

4.3.1 Word-level Information Acquirement

The most explicit, and perhaps the most reliable sources of gender and type information are the various pre- and post-modifiers in proper names, such as '*Mr.*', '*Jr.*', and '*Corp.*' Proper names that are not clearly marked for gender often nevertheless give clues about what their referents are (or are not). One would assume that '*Securities and Exchange Commission*' refers to an organization, and exclude the possibility of '*Cray Computer*' being a person. Common nouns provide similar clues. For example, '*director*' is very likely a title of a person, while '*project*' almost certainly is not. Like names of persons, common nouns also varies in their distributions across genders. For example, '*professor*', '*nurse*', and '*farmer*' can all serve as antecedents of both masculine and feminine pronouns, but the likelihood varies significantly.

Gazetteer and thesaurus are the two commonly-used external sources of gender and type information. Both of them require laborious manual efforts. Gazetteers provide quick and relatively accurate information for proper names. A thesaurus such as WORDNET, on the other hand, is more useful for identifying information about common nouns. Aside from the difficulties involved in compiling and subsequently extending these manual knowledge bases, which cast a limit on their size, the other major issue that limit their usefulness is that they usually do not contain information on the prior distribution of an entity across multiple possible categories. Many proper names and a large portion of common nouns have different meanings, and the different senses often map to different genders. For example, '*Washington*' may refer to either a city, a government, or a person; and '*John*' can serve as the first name of either a male or female. Gender information of a common noun can be obtained by tracing its senses up the hyponymy hierarchy: if a particular sense falls under `person`, the word is not likely to serve as referent of the neutral pronoun *it*. Obviously, this approach is not capable of capturing more fine-grained information such as a '*workman*' is more likely a male. The most prominent issue however, as noted by Pantel and Ravichandran (2004) and others, is that while WORDNET fails to cover certain domain-specific senses of words, it includes many infrequent senses of words, which are distractive to the purpose of gender identification. For example, the nouns '*dog*', '*computer*', and '*company*' each has a sense that points to `person`. The relative order of the sense is not a reliable indicator either, one example is that the word '*calculator*' has its first sense in WORDNET as a hyponym of `person`. A possible solution to this problem is to perform sense disambiguation prior to gender identification. However, word sense disambiguation is by itself a difficult problem, perhaps even more so than gender identification.

Acknowledging the difficulties associated with manual sources, some researchers resorted to corpus-mining to automatically induce gender information associated with nominal entities. Ge et al. (1998) presented a bootstrap-based approach to automatically harvest gender information in large corpora. The algorithms is rather simple in concept – it runs the simple pronoun resolution system designed by Hobbs (1978), without applying the gender constraint, against a large un-annotated corpus. The antecedent of each coreferential pair is assigned the gender information of its pronominal anaphor. The resulting set is then used to assess the gender information of the individual nouns. The result set is tested on a list of proper names with gender-marked designators and achieved an accuracy of 70.3%.

A more recent advance in this field is Bergsma's (2005a) web-based approach. It identifies five relatively robust syntactic paths that associate pronominal anaphors and their antecedents as well as the respective search engine queries designed to find such instances:

1. Reflexives, as in '*Mary explained herself . . .*':
   himself, *herself*, *itself*, and *themselves* in ‘‘noun * reflexive’’
2. Possessives, as in '*IBM bought its supplies . . .*':
   his, *her*, *its*, and *their* in ‘‘noun * possessive’’
3. Nominatives, as in '*Alice thought she should . . .*':
   he, *she*, *it*, and *they* in ‘‘noun * nominative’’
4. Predicates, as in '*He is a father.*':
   he, *she*, *it*, and *they* in ‘‘nominative is/are [a] noun’’
5. Designators, as in '*Mr . Brown*':
   Mr. and *Mrs.* in ‘‘designator noun’’

An SVM using the web-mined features achieved an F-score of 90.4 in executing a four-way classification of masculine/feminine/neutral/plural, slightly higher than average human performance.

The system developed in this study performs gender and type identification by leveraging multiple sources of information: fixed patterns in proper names, a list of personal names, the WORDNET, word frequency, and the web. In general, the system tries to process proper names with explicit markers and those identifiable by WORDNET first. If gender and type information can be completely determined from these sources, no further processing is required. For the rest of the proper names, the name list, the word frequency list, and the web are consulted.

Pattern Matching

As discussed earlier, many proper names are explicitly 'marked' with gender and type information. The system recognizes a small group of common markers, as detailed in Table 4.1, and assigns corresponding gender and type information to the marked names. A small number of common nouns, such as '*chairman*' and '*spokeswoman*', are also marked for gender. However, the reliability of such markers vary, making them unsuitable targets for simple pattern matching.

Rule	**Type**	**Gender**
Mr. Mister Sir	PERSON	Masculine
Mrs. Ms. Miss	PERSON	Feminine
Dr Prof Rep	PERSON	Masculine/Feminine
Jr.	PERSON	Masculine
Corp Inc PLC LLC GMBH Co Ltd Group	company	Neutral
Name ␣ , ␣ State Acronym[a]	city	Neutral

Table 4.1: Patterns and their corresponding gender and type information

[a] As in '*Gettysburg, Pa.*'

In addition to the rules identified in Table 4.1, the system also recognizes '*Messrs.*' and '*Mr. and Mrs.*'. Entities marked with these modifiers receive special treatments so that they can be matched with mentions of the individuals. A similar rule applies to names of couples, such as '*William and Margie Hammack*' and '*the Hammacks*', which are interpreted as '*William Hammack and Margie Hammack*'[2] and '*Mr. Hammack and Mrs. Hammack*'[3] respectively.

Name List

The personal name list used by the system contains frequently-occurring last names and first names (categorized by gender) compiled from a subset[4] of the 1990 U.S. census data by the U.S. Census Bureau (1995)[5]. Unlike other gazetteers, the list also provides the frequencies associated with each name, a feature especially useful when a first name is be used by both male and female. For example, according to the list, the first name '*John*' is used by 3.27% of the males surveyed and 0.01% of the females. Since the number of males and females are roughly equal in the subset, it can be estimated that the probability of '*John*' referring to a male is roughly 99.7%, provided it is used to denote a person.

The WordNet

The WORDNET is used to identify if a common noun or the last word of a proper name denotes an organization, location, unit of measure, or time period, or if the name is associated with a 'well-known' location such as a continent, a country, a big city, or a large body of water.

Tests for well-known locations are performed first. In these tests, the WORDNET is essentially used as a gazetteer of geographic names. The names are tested for hypernyms against the WORDNET senses `city`, `state`, `country`, `location`, `land (4)`, and `body of water`. If one the tests is successful, the name is marked as 'city', 'state', 'country', or 'LOCATION' accordingly.

If a proper name does not belong to one of the well-known locations, the head component is further tested for hypernyms against WORDNET senses of `organization`, `location`, `measure`, and `time period`. Upon successful completion of the tests, the names are marked as being an instance of either the head word, or one of 'MEASURE' or 'TIME'.

In addition, the WORDNET is also used for identifying job titles embedded in names of persons (e.g. '*U.S. Trade Representative Carla Hills*'). Recognizing the titles not only gives additional type information to the names but also help gender determination by isolating the names from additional descriptive contents.

[2] The system identifies that both '*William*' and '*Margie*' are valid first names, while '*Hammack*' is a valid last name.

[3] The system checks for previous mentions of individuals having the last name before this kind of interpretation is made.

[4] According to the accompanying document, the subset contains roughly six million census records, or about one-fortieth of the U.S. population at the time.

[5] Barbu (2003) uses the same list for named entity recognition.

Proper Name Ratio

The proper name ratio is the relative frequency of a word appearing in a proper name. This statistic is helpful in identifying names of organizations and job titles. Names of organizations often contain portions indicating their characteristics, such as the main business or type of organization. These words, usually located at the end of the names, generally have low proper name ratios since they are more often used as common nouns. Similarly, job titles also have low proper name ratios in general. In contrast, names of persons are usually only used as proper names and therefore have high proper name ratios. In the current implementation, the ratio is determined using a word frequency table obtained from the BNC corpus and is mainly used to provide 'guesses' for type information. For example, the system assigns the tentative type of 'hotel' for the name '*Amfac Hotel*' based on the information that the word *hotel* only appears as proper name in about 16% of its uses.

Web-based Gender/Number Determination

Finally, the system directly queries the web to obtain the probability distribution of a word's gender. The four query patterns, including two high-precision sets and another two backups that offer higher coverage but lower precision, are similar in spirit to those proposed by Bergsma (2005a) but provide more accurate results.

$$\texttt{Reflexives (high-precision)} \qquad \textit{The} \textvisiblespace \texttt{noun} \textvisiblespace \texttt{reflexive} \tag{4.1}$$

$$\texttt{Possessives (high-precision)} \qquad \textit{The} \textvisiblespace \texttt{noun} \textvisiblespace \textit{and} \textvisiblespace \texttt{possessive} \tag{4.2}$$

$$\texttt{reflexives (backup)} \qquad \textit{the} \textvisiblespace \texttt{noun} \textvisiblespace \texttt{reflexive} \tag{4.3}$$

$$\texttt{relative pronouns (backup)} \qquad \textit{a} \mid \textit{an} \textvisiblespace \texttt{noun} \textvisiblespace \texttt{relative pronoun} \tag{4.4}$$

The choice of the pronouns used to instantiate the patterns are as follows:

$$\begin{aligned} \texttt{reflexive} \in & \quad \{\textit{himself}, \textit{herself}, \textit{itself}, \textit{themselves}\} \\ \texttt{possessive} \in & \quad \{\textit{his}, \textit{her}, \textit{it}, \textit{their}\} \\ \texttt{relative pronoun} \in & \quad \{\textit{who}, \textit{which}\} \end{aligned}$$

With the capitalized *The* at the front (also note the capitalization in names of the patterns), the two high-precision patterns are intended to only match expressions appearing at the beginning of sentences. This requirement eliminates false-positives such as '*Although she has not met the man herself, she* ...'. The backup patterns are used when the high-precision queries do not produce enough results. The first one simply removes the capitalization from the high-precision reflexives pattern. The other backup pattern only attempts to identify the likelihood of a given noun referring to a human being – a high number of hits obtained from the *who* query is interpreted as that the noun is more likely to serve as referent for either *he* or *she* than to *it*, and vice versa.

With some minor modifications, the patterns can be adapted for proper names:

`reflexives (proper noun)`	`proper noun ␣ reflexive`	(4.5)
`possessives (proper noun)`	`proper noun ␣` *`and`* `␣ possessive`	(4.6)
`relative pronouns (proper noun)`	`proper noun ␣ , ␣ relative pronoun`	(4.7)

To illustrate the use of the patterns, consider the nouns *man* and *house*. The instantiated queries and their respective result counts are listed in Table 4.2. In the implemented system, the result counts returned from queries are further normalized by the frequencies of the respective pronouns; this process is not reflected in Table 4.2.

Pattern	Masculine		Feminine		Neutral		Plural	
	Query	Cnt	Query	Cnt	Query	Cnt	Query	Cnt
`Ref`	The man himself	8.8E3	. . . herself	0	. . . itself	0	. . . themselves	0
`Pos`	The man and his	7.9E3	. . . her	0	. . . its	0	. . . their	0
`ref`	the man himself	1.4E5	. . . herself	1.6E2	. . . itself	2.3E2	. . . themselves	9.3E1
`rel`	a|an man who			2.3E6	. . . which	1.0E4	-	-
`Ref`	The house himself	0	. . . herself	0	. . . itself	1.4E4	. . . themselves	0
`Pos`	The house and his	0	. . . her	0	. . . its	1.5E3	. . . their	0
`ref`	the house himself	1.8E3	. . . herself	8.2E2	. . . itself	2.9E4	. . . themselves	1.0E3
`rel`	a|an house who			1.1E3	. . . which	2.6E4	-	-

Table 4.2: Results of gender/number queries instantiated with the words *man* and *house*, obtained from the Google N-gram corpus. The abbreviated pattern names `Ref`, `Pos`, `ref`, and `rel` refer to `Reflexives`, `Possessives`, `reflexives`, and `relative pronouns`, respectively.

As shown in Table 4.2, instances of the two backup patterns yields considerably higher result counts, but the results are also significantly more noisy. Some of the original Bergsma queries are also affected by the noise problem. For example, the `predicates` pattern is affected by noise originated from the broad spectrum of the pronoun *it*'s usage – the query "he is a man" only yields about five times the result count produced by "it is a man"[6]. The `nominatives` pattern is also probably[7] not very reliable, since a pronoun serving as the subject of a finite sub-clause does not always refer to the matrix subject, with the degree of likelihood partly associated with the matrix verb.

The patterns are designed to be executed not by a search engine but against the Google N-grams corpus. For the purpose of this study, the corpus is pre-processed and subsequently indexed with hash values. Two different hash algorithms – a 48-bit version of the SDBM algorithm by Seltzer and Yigit (1991) and Pearson's (1990) 8-bit hashing algorithm, are used to create indices from the

[6] A valid example that matches this query is '*At least I recognize that it is a man and not an apple.*'(Feynman & Gilbert, 1960).

[7] Since Google no longer limits the content matched by the wildcard (*) to a single word, the original queries cannot be performed as intended. However, as a collateral evidence, the query "She|she thought she should" against the Google N-grams corpus generates around four times the result count obtained from "She|she thought he should". Running the same queries against the Google search engine, the first one only produces around twice as many results as the second.

unique N-grams. For each N, a total of 256 index groups are established with the value of the 8-bit hashes serving as group identifiers. Besides the SDBM hash values, the index files also contain the occurrence count of each N-gram, represented in a custom, 16-bit unsigned floating point structure[8] composed of 5 exponent bits and 11 bits for the significand. The structure is capable of exactly representing integers up to 2^{12}, or 4096, with the largest-possible value being 4095×2^{30}, or roughly 4×10^{12}. Both the choice of the hashing algorithms and the adoption of the custom floating point structure are motivated by practical considerations. The two hashing algorithms are both known for their speed, and the SDBM algorithm also exhibits relatively low collision rate[9]. The probability of collision is further reduced by grouping indices according to the values produced by the Pearson's (1990) algorithm, which operates based on a different concept from that of the SDBM. Of the one billion 5-grams, a total of 6 collisions are detected in 5 (out of the 256) index groups. The custom floating point format is used to alleviate the pressure inflicted on computer memory by the huge amount of N-grams. The decrease in data size also has a positive impact on the speed of query execution. The inaccuracy caused by the format in representing larger numbers ($n > 4096$) is roughly 0.02%, which can be safely ignored considering the nature of this study.

4.3.2 Text-level Information Acquirement

In addition to the information available from the expressions per se, the system also captures various relationships available in the text. Consider examples (4.3), repeated here, and (4.4):

(4.3) J.P. Bolduc, vice chairman of W.R. Grace & Co., which holds a 83.4% interest in this energy-services company, was elected a director.

(4.4) Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group. WSJ 1:1-2

The system captures external evidences from six difference sources:

Number agreement with verbs as in the relationship between *which* and *holds* in (4.3). While *which* can refer to both singular and plural entities, the singular present *holds* in this case rules out the latter possibility. The subject-verb agreement is not a particularly strong source of gender/number information: it only makes the distinction between plural and singular uses, and is only available to a relatively small number of entities. However, in cases where gender/number information cannot be reliably determined from a word itself, such as the word *fish*[10] in (4.5), it can be the crucial piece of the puzzle.

(4.5) The fish often are plentiful around the pilings of the old gas wells that dot the flat surface like the remains of sunken ships. WSJ 1323:11

[8] This all-integer format belongs to a family of such formats detailed by Munafo (2006).

[9] The choice of SDBM is based on some limited empirical tests. A more thorough evaluation of different hashing algorithms is available from Henke, Schmoll, and Zseby (2008).

[10] The word is annotated as singular (`NN`) in the corpus. Aside from *fish*, there are a few other nouns in English, such as *sheep* and *deer*, that are not inflected for plural uses. The system does not offer any special treatment for these nouns.

Non-restrictive apposition as in the relationship between *vice chairman of W.R. Grace & Co.* and *J.P. Bolduc* in (4.3). Apposition is the most reliable source of external information, not only because the relationship is easy to identify, but also due to that apposition is generally used to provide objective information about an entity.

Predication as in the relationship between *Mr. Vinken* and *chairman of Elsevier N.V.* in (4.4). From anaphora resolution's point of view, information provided by the predication relationship is not always reliable (the same can be said about apposition, but to a less extent). For example, in (4.6), *publishing sensation* is used to give Mr. Murakami the status of being a famous writer. However, the word *sensation* is much less often associated with a human than non-human.

(4.6) The 40-year-old Mr. Murakami is a publishing sensation in Japan. WSJ 37:15

***As*-preposition** as in the relationship between *a nonexecutive director* and *Pierre Vinken* in (4.4). Although there are some subtle differences (cf. Jäger, 2003) between the two, *as* preposition can be seen as roughly equivalent to predication for the purpose of this study. In other words, (4.4) can be processed in the same manner as '*Pierre Vinken is a nounexecutive director.*'

Relative pronoun as in the relationship between *which* and *W.R. Grace & Co.* in (4.3). With some notable difficulties (Cardie, 1992), relative pronouns are usually resolved with high confidence. The choice of the *wh*-word, especially between *who*/*whom* and *which*, is a good indicator for the gender of the relative pronoun's antecedent.

Alias as in the relationship between *Mr. Vinken* and *Pierre Vinken* in (4.4). Aliases and repeated occurrences of the same name can be detected with relatively high confidence, and information gathered from one of the aliases can be safely applied to the others.

Interpreting *As*-prepositions

Unlike appositions and predications, which are usually directly recoverable from surface structures, correctly interpreting an *as*-preposition involves some additional work. The main difficulty lies in determining the subject to which the preposition is attached. In the case of (4.4), the distracter *the board* is the object of the verb *join* and is also semantically incompatible with the predicate nominal *a nonexecutive director*. However, neither needs to be true for many other cases. Consider (4.7):

(4.7) The SEC documents describe those chips, which are made of gallium arsenide, as being so fragile and minute they will require special robotic handling equipment. WSJ 18:18

Here the distracter *The SEC documents* is the subject of *describe*, and also happens to be compatible with *fragile* and *minute*. In fact, *fragile* is more often used to describe documents than it is for *chips*[11]. From (4.7), it is obvious that the most significant factor is the verb itself rather than its

[11] In the Google N-grams corpus, the term "documents" is only four times more frequent than "chips", but querying "fragile documents" yields a hit count of more than 1,500 while "fragile chips" returns 0 (the Google N-grams corpus has a cut-off count of 40).

arguments. In a relationship '*x* describes *y* as *z*', the predicate *z* most-likely applies to *y*, while in '*x* joins *y* as *z*', *z* is more often applicable to *x*.

This study uses a web-based approach to determine the subject of the *as*-predication. For each relationship '`subject verb object` *as*', a set of eight queries are constructed according to the following patterns:

$$\mathit{As}\text{-}\mathtt{prep}_{\text{Subject,NumberMatch}} \qquad \mathtt{singular\ pronoun} \textvisiblespace \mathtt{verb} \textvisiblespace \mathit{as} \textvisiblespace \mathit{a}|\mathit{an} \tag{4.8}$$

$$\overline{\mathit{As}\text{-}\mathtt{prep}}_{\text{Subject,Singular}} \qquad \mathtt{singular\ pronoun} \textvisiblespace \mathtt{verb} \textvisiblespace \mathit{as} \tag{4.9}$$

$$\mathit{As}\text{-}\mathtt{prep}_{\text{Subject,NumberMismatch}} \qquad \mathtt{plural\ pronoun} \textvisiblespace \mathtt{verb} \textvisiblespace \mathit{as} \textvisiblespace \mathit{a}|\mathit{an} \tag{4.10}$$

$$\overline{\mathit{As}\text{-}\mathtt{prep}}_{\text{Subject,Plural}} \qquad \mathtt{plural\ pronoun} \textvisiblespace \mathtt{verb} \textvisiblespace \mathit{as} \tag{4.11}$$

$$\mathit{As}\text{-}\mathtt{prep}_{\text{Object,NumberMatch}} \qquad \mathtt{verb} \textvisiblespace \mathtt{singular\ pronoun} \textvisiblespace \mathit{as} \textvisiblespace \mathit{a}|\mathit{an} \tag{4.12}$$

$$\overline{\mathit{As}\text{-}\mathtt{prep}}_{\text{Object,Singular}} \qquad \mathtt{verb} \textvisiblespace \mathtt{singular\ pronoun} \textvisiblespace \mathit{as} \tag{4.13}$$

$$\mathit{As}\text{-}\mathtt{prep}_{\text{Object,NumberMismatch}} \qquad \mathtt{verb} \textvisiblespace \mathtt{plural\ pronoun} \textvisiblespace \mathit{as} \textvisiblespace \mathit{a}|\mathit{an} \tag{4.14}$$

$$\overline{\mathit{As}\text{-}\mathtt{prep}}_{\text{Object,Plural}} \qquad \mathtt{verb} \textvisiblespace \mathtt{plural\ pronoun} \textvisiblespace \mathit{as} \tag{4.15}$$

The choice of the pronouns are as follows:

$$\mathtt{singular\ pronoun} \in \qquad \{\mathit{He}|\mathit{he}|\mathit{She}|\mathit{she}|\mathit{It}|\mathit{it}, \mathit{him}|\mathit{her}|\mathit{it}\}$$

$$\mathtt{plural\ pronoun} \in \qquad \{\mathit{They}|\mathit{they}, \mathit{them}\}$$

In order to further expand the coverage of the queries, the verbs are also expanded to include various inflected forms. For example, instantiating $\mathit{As}\text{-}\mathtt{prep}_{\text{Subject,NumberMatch}}$ pattern with (4.7) yields "He | he | She | she | It | it describes | described as a". Similarly, the patterns $\mathit{As}\text{-}\mathtt{prep}_{\text{Subject,NumberMismatch}}$ and $\mathit{As}\text{-}\mathtt{prep}_{\text{Object,NumberMismatch}}$ translate into "They | they describe | described as a" and "describe | describes | described them as a", respectively.

The patterns exploit the fact that the predicate nominal and its subject generally agree in number. When a mismatched pronoun is used, the position that holds the real subject of the *as*-predication should be affected more adversely than the one that does not. The patterns marked with overlines are auxiliary patterns used to produce normalizing factors. Let $n_{Sbj,Match}$, $\bar{n}_{Sbj,Match}$, $n_{Sbj,Mismatch}$, $\bar{n}_{Sbj,Mismatch}$, $n_{Obj,Match}$, $\bar{n}_{Obj,Match}$, $n_{Obj,Mismatch}$, and $\bar{n}_{Obj,Mismatch}$ denote the respective result counts obtained from query instances of the eight patterns, two ratios can be calculated:

$$r_{Sbj} = \frac{n_{Sbj,Mismatch} / \bar{n}_{Sbj,Mismatch}}{n_{Sbj,Match} / \bar{n}_{Sbj,Match}} \tag{4.16}$$

$$r_{Obj} = \frac{n_{Obj,Mismatch} / \bar{n}_{Obj,Mismatch}}{n_{Obj,Match} / \bar{n}_{Obj,Match}} \tag{4.17}$$

r_{Sbj} and r_{Obj} indicate respectively the degree to which the subject and the object positions of the verb are affected when a number-mismatched pronoun is used: the smaller the respective r value, the more affected the position is. After comparing the two r values, the position that is affected more is chosen as the subject of the *as*-predication.

Table 4.3 illustrates the results of the queries instantiated with (4.4) and (4.7). As expected, in both cases the positions that hold the subjects of the *as*-predications are more heavily influenced by the 'wrong' pronouns. For (4.4), *Pierre Vinken* is chosen as the subject of *as a nonexecutive director*, since $r_{Sbj} < r_{Obj}$. Similarly, *those chips* is chosen in (4.7) instead of *the SEC documents* because r_{Obj} is smaller.

Verb	Position	n_{Match}	$\bar{n}_{Match}$	$n_{Mismatch}$	$\bar{n}_{Mismatch}$	r
join	subject	2.94E3	6.27E3	2.21E2	1.02E3	**0.46**
	object	3.27E3	1.56E4	2.90E3	1.61E4	0.86
describe	subject	5.58E4	2.68E5	1.08E4	5.01E4	1.03
	object	1.33E5	4.87E5	5.43E3	7.98E4	**0.25**

Table 4.3: Query results for examples (4.4) and (4.7)

Occasionally, *as*-prepositions appear as components of noun phrases, as in (4.8) and (4.9):

(4.8) But it is Mr. Lane, as movie director, producer and writer, who has been obsessed with refitting Chaplin's Little Tramp in a contemporary way. WSJ 39:3

(4.9) The survival of spinoff Cray Computer Corp. as a fledgling in the supercomputer business appears to depend heavily on ... WSJ 18:1

These cases are processed based on surface structure only. If the noun phrase involves an *of*-preposition, as in the case of (4.9), its object (i.e. *Cray Computer Corp*) is assigned as the subject of the *as*-predication; otherwise the head noun is used.

Alias Matching

In modern English, proper names, i.e. names of persons, companies, places, and other specific things, are usually capitalized and therefore easy to identify. However, name variants (or aliases) are often used to denote the same entity throughout a document. For example, the person first introduced under the title '*Seymour Cray*' may later be referred to as either '*Mr. Cray*' or simply '*Seymour*', or mentioned in '*Messrs. Cray and Barnum*', an expression combining both '*Mr. Cray*' and '*Mr. Barnum*'. Names of organizations are often shortened or substituted with acronyms in subsequent mentions, e.g. '*Cray Computer*' used in place of '*Cray Computer Corp.*' and '*SEC*' for '*Securities and Exchange Commission*'.

Since proper names are names of specific entities, it is not surprising that the different aliases usually bear certain literal similarity. Taking the name '*Seymour Cray*' as an example, both alternative forms '*Mr. Cray*' and '*Seymour*' in the subsequent mention share one word with the more complete form. Therefore, a distance measure that quantifies the similarity between two aliases is essential to the resolution of proper names. This study employs a form of the Levenshtein distance (or minimum edit distance, Levenshtein, 1966), a measure of the minimum cost of transforming one string literal to another through a set of editing operations, to represent the degree of similarity between aliases. Strube, Rapp, and Müller (2002) have also used normalized minimum edit distance between the anaphor and the potential antecedent as part of their feature set and obtained good results on German articles. The Levenshtein distance is probably the most flexible algorithm among the commonly used distance functions for string comparison[12]. The algorithm supports three editing operations – insertion, deletion, and substitution, each of which can be assigned a distinct cost. Using dynamic programming, Levenshtein distance can be obtained by incrementally filling an $(m+1)\times(n+1)$ cost matrix, D, where m and n are the respective length of the string literals being compared (denoted x and y). The matrix is first partially initialized with values $D_{i,0} = C_{ins}$ and $D_{0,j} = C_{del}$, where C_{ins} and C_{del} denote the costs of insertion and deletion, respectively. The rest of the matrix is then gradually computed according to the following formula:

$$D_{i,j} = \begin{cases} D_{i-1,j-1} & ,\text{if } x_i = y_j \\ min(D_{i,j-1} + C_{ins},\ D_{i-1,j} + C_{del},\ D_{i-1,j-1} + C_{sub}) & ,\text{if } x_i \neq y_j \end{cases} \tag{4.18}$$

, where C_{sub} denotes the cost of substitution. Once the computation is done, the final distance d can be easily retrieved as $d = D_{m,n}$. Table 4.4 illustrates the computation of the Levenshtein distance using equal costs. The entry located at the bottom-right corner of the table is the final result.

		m	**a**	**x**	**i**	**m**	**a**
	0	1	2	3	4	5	6
m	1	**0**	1	2	3	4	5
i	2	1	**1**	2	2	3	4
n	3	2	2	**2**	3	3	4
i	4	3	3	3	**2**	3	4
m	5	4	4	4	3	**2**	3
u	6	5	5	5	4	**3**	3
m	7	6	6	6	5	4	***4***

Table 4.4: The Levenshtein distance (with equal costs of operations $C_{ins} = C_{del} = C_{sub} = 1$) between the literals '*minimum*' and '*maxima*', obtained using dynamic programming. The path to the final result (***4***) is marked as **bold**.

As Strube et al. (2002) demonstrated, the Levenshtein distance can be a useful tool for coreference resolution in its original form. The implementation used in this study features some adjustments to make it more effective for English proper names. The first adjustment is to use words instead of letters as the minimum granularity. This adjustment reflects the fact that proper names are generally not subject to morphological transformations and are therefore better handled at the granularity of words. To illustrate the inappropriateness of letter-based metrics, the letter-based distance between '*John Smith*' and '*John*', a pair that is likely to be coreferential, is 6, while the distance between '*John Smith*' and '*Jane Smith*', a very unlikely pair of aliases, is only 3. The second adjustment is to set higher costs for insertion and substitution operations (set to 2 and 3 respectively in the cur-

[12] Navarro (2001) identifies the four most commonly used distance functions – the Levenshtein distance, the Hamming distance, the episode distance, and the longest common subsequence distance.

rent implementation). Empirical evidence suggests that when multiple names are used to describe the same entity, the first mention is usually more 'complete'. For example, if '*John*' is used at the beginning of the article to refer to John Smith, it would be unusual (albeit not impossible) for him to be mentioned later as '*John Smith*'. With the two adjustments, the distance between '*John Smith*' and '*John*' is merely 1, and the distance between '*John Smith*' and '*Jane Smith*' becomes 3. Another difference between the approach adopted in this study and that of Strube et al. (2002) is that the system recognizes the pre- and post-modifiers, such as *Mr.* and *Corp.*, inside proper names. Many other systems (e.g. that of Soon et al., 2001) process proper names differently according to the modifiers. However, in this study, these modifiers are considered non-essential for the purpose of alias matching and are ignored during the process.

Auxiliary Relationships

Aside from the aforementioned relationships, the system also captures possessions, prepositions, and verb-argument structures from the parsed text. These relationships are referred to as 'auxiliary relationships' in the rest of the chapter.

Possessions and prepositions are generally easy to identify – the former is marked by either a possessive case or an *of*-preposition; the other prepositions (except *as*) belong to the latter. The biggest issue is the versatility of the *of*-preposition (cf. Quirk et al., 1985, section 9.54). Aside from the genitive use, *of*-prepositions can be used in place of apposition, as in '*the month of October*', or to express complex relationships, as in '*a boat of fiberglass*', and quite often used partitively, as in '*a number of controversial issues*'. The partitive cases are especially troublesome for relationship identification, since from the semantic point of view, the head nouns in these cases are only auxiliary while the larger part of the meaning is contributed by the partitives. For example, consider (4.10), identifying the relationship '*Shea & Gould held discussions*' would be much more helpful than merely recognizing '*Shea & Gould held a number*'.

(4.10) He said Shea & Gould held a number of discussions with the five partners during the past few weeks to get them to stay but . . . WSJ 1446:39

The system identifies partitive constructions with syntactic cues and the WordNet. When the head word is an adjective or determiner[13], or when the construction is non-definite and the head noun is a hyponym/synonym of the WordNet sense `kind`, `quantity`, or `collection`, or is a synonym of `group`.

Gathering verb-argument structures is not always a straight-forward task either. Consider example (4.11), the syntactic structures of which are shown in Figure 4.1:

(4.11) Cray Computer has applied [*] to trade on Nasdaq. WSJ 18:26

[13] The system considers words marked with the part-of-speech tags `JJ`, `JJR`, `JJS`, `RB`, `RBR`, `RBS`, or `DT`.

There are two verb-argument relations in (4.11) involving *Cray Computer*, one covered by the verb *apply*, and the other, *trade*. The original WSJ annotation marks the second relationship with an NP-* pointing to *Cray Computer*. However, since the null elements originally present in the annotated WSJ corpus have been deliberately removed during pre-processing, the relationship has to be reconstructed. Identifying and resolving the empty categories is a complex problem on its own (see e.g. Campbell, 2004; Gabbard, Marcus, & Kulick, 2006). Instead of trying to fully recover the removed null elements, the system follows some simple heuristics to process a subset of the 'subjectless' nonfinite clauses.

Given a gerund or infinitive, the system first searches for subject at local level. If a subject cannot be found and the non-finite clause does not function as a subject or predicate itself, the search continues by navigating up the dependency tree and checking for presence of subject at each node[14]. With a few exceptions, the iteration is stopped whenever a non-verb node, such as a noun, adjective, or preposition[15], is encountered on the path. This simple method generally works well. However, there are two special situations that require additional attention. The first is when the matrix verb also has an object or when the matrix clause is a passive construct, as illustrated by examples (4.12) through (4.14):

(4.12) . . . federal thrift regulators ordered it [*] to suspend dividend payments . . . WSJ 2360:1

(4.13) Commonwealth Edison Co. was ordered [*] to refund about $250 million . . . WSJ 15:1

(4.14) The figures in both reports were adjusted [*] to remove the effects of . . . WSJ 36:10

In (4.12), the object *it* is the antecedent to the subject NP-* of *to suspend*. Similarly, *Commonwealth Edison Co.*, the surface subject of the passive construct *was ordered*, is the antecedent to the subject NP-* of *to refund* in (4.13). In (4.14), the subject NP-* of *to remove* has no antecedent. A pair of web-based query patterns are used to help identify cases like (4.12), (4.13), and (4.14):

$$\texttt{nominal clause} \qquad \texttt{verb}_{VBD} \textvisiblespace\ \textit{that} \textvisiblespace\ \textit{the} \tag{4.19}$$

$$\texttt{passive} \qquad \texttt{is} \textvisiblespace\ \texttt{verb}_{VBN} \textvisiblespace\ . \tag{4.20}$$

$\texttt{verb}_{VBD}$ and $\texttt{verb}_{VBN}$ are the inflected forms of the verb corresponding to past tense and past participle, respectively. The `NominalClause` pattern tests the matrix verb's ability to take nominal clauses. If the verb is regularly used with nominal clauses, the matrix object is identified as the antecedent to the subject NP-* of the infinitive. The `Passive` pattern is used to estimate how often the verb takes regular noun phrases as objects. The decision is made based on the ratio of the hit

[14] More specifically, the system captures 'verb chains' (cf. Section 4.4.1) on the dependency structure. Lower-level chains with 'missing' subjects simply peg the null elements to either the subject or object positions of their parents. The system resolves the empty categories in an iterative manner, as one null element maybe attached to another.

[15] For gerund clauses serving as object of preposition, the search starts from above the preposition. The system only processes gerunds following the prepositions *by*, *without*, *before*, *after*, *since*, *upon*, *while*, *from*, *for*, and *of*. The last three link to the logical objects of the parent clauses (e.g. '*I thanked him for being so nice.*' and '*He was punished for breaking the glass.*'), and the rest link to the surface subjects.

counts of queries generated by these two patterns, $r = n_{NominalClause}/n_{Passive}$. When a small r value is observed (the current implementation uses a threshold value of 0.1), the system identifies the subject of the matrix clause as that of the infinitive; if the matrix clause is a passive construct, no subject is assigned.

The other issue is that many gerunds are dangling participles, as in (4.15):

(4.15) South Korea registered a trade deficit of $101 million in October, [*] reflecting the country's economic sluggishness ... WSJ 11:1

Out of the 1,267 cases in the WSJ corpus having syntactic structures similar to that of (4.15), only 828, or about 65%, have antecedents assigned to the corresponding subject NP-* elements. The system uses another pair of web-based patterns to handle these cases:

`PronominalSubject`	`Pronoun` ␣ $\texttt{verb}_{VBD}$ ␣ `stub`	(4.21)
`DemonstrativeSubject`	`Demonstrative` ␣ $\texttt{verb}_{VBD}$ ␣ `stub`	(4.22)

In the current implementation, `Pronoun` = *He* | *She* | *They* | *It*, and `Demonstrative` = *This*. If the gerund verb is immediately followed by a preposition, the preposition is used as `stub`; otherwise `stub` is simply set to *the*. The design of this set of patterns is motivated by the fact that a large portion of the antecedentless cases actually have actions or events as the antecedents of their NP-* elements, as in the case of (4.15). Since demonstratives are often used to refer to actions or events, it is expected that cases like (4.15) would yield more results ($r = n_{DemonstrativeSubject}/n_{PronominalSubject} > 0.1$) for the `DemonstrativeSubject` queries. When evaluated on the 1,267 cases matching the syntactic pattern `(VP ...(, ,) (S (NP-SBJ (-NONE- *)) (VP (VBG) ...)))`, the system achieves an accuracy of 78.2% (precision= 84.0%, recall= 82.4%) in identifying subject NP-* elements with antecedents.

The system only collects auxiliary relationships that involve at least one noun, pronoun, or named entity, and indexes them by these nominal entities. When needed, the collected relationships can be easily converted to web queries using Patterns 4.23 through 4.32.

`possession`	`possessor` ␣ *'s* ␣ `possessee`	(4.23)
$\overline{\texttt{possession}}$	{*'s* ␣ `possessee`, `possessor` ␣ *'s*}	(4.24)
of-`prep`	`possessee` ␣ *of* ␣ *the* `possessor`	(4.25)
$\overline{\textit{of}\texttt{-prep}}$	{`possessee` ␣ *of* ␣ *the*, *of* ␣ *the* `possessor`}	(4.26)
`prep`	`head` ␣ `preposition` ␣ *the* ␣ `object`	(4.27)
$\overline{\texttt{prep}}$	{`head` ␣ `preposition` ␣ *the*, `preposition` ␣ *the* ␣ `object`	(4.28)
$\texttt{verb}_{\text{Active}}$	`subject` ␣ `verb` ␣ `stub`	(4.29)
$\overline{\texttt{verb}}_{\text{Active}}$	`subject`	(4.30)
$\texttt{verb}_{\text{Passive}}$	`object` ␣ *was* | *were* ␣ $\texttt{verb}_{VBN}$	(4.31)

$$\overline{\texttt{verb}}_{\text{Passive}} \qquad\qquad \texttt{object} \qquad (4.32)$$

A possession relationship such as '*their house*', '*the company's products*', or '*products of the company*' has two 'ends', a possessee and its possessor. Given a possession, one of the end positions, and a replacement entity, the system generates a possession query and the accompanying $\overline{\texttt{possession}}$ query for normalization purposes. In the generated queries, both the possessor and the possessee are expanded to include both singular and plural forms. For example, replacing the possessor in '*the company's products*' with '*firm*' results in the possession query "firm | firms 's product | products" and the $\overline{\texttt{possession}}$ query "firm | firms 's". The system also generates *of*-prep queries for possession relationships, which are intended to be run in parallel[16] with the possession queries. Similar to the possession pattern, the possessor and the possessee are also expanded. The same applies to the prep pattern.

The treatment of verb-argument relations are slightly more complex. If the replaced end is the surface subject of the relationship and the relationship in active voice, the pattern $\texttt{verb}_{\text{Active}}$ is used; otherwise the system chooses the $\texttt{verb}_{\text{Passive}}$ pattern. There are two variables in the $\texttt{verb}_{\text{Active}}$, verb and stub. If the verb in the original relationship is in past tense or past participle tense, verb is rendered as $\texttt{verb}_{VBD}$ (past tense). Otherwise it is rendered as $\texttt{verb}_{VBP}|\texttt{verb}_{VBZ}$, the combination of the non-singular and singular forms of the present tense. The stub includes any particle and/or preposition immediately following the verb, or *a*|*an*|*the* if the verb takes one or more objects (particle is still included if there is one). The system also generates possession and *of*-prep queries for verb-argument relationships having noun phrases at both ends. For example, given the relationship '*the Cray-3 will contain 16 processors*' and a replacement '*machine*' for the surface subject, three queries are generated: "machine | machines contain | contains a | an | the", "machine | machines 's processor | processors", and "processor | processors of the machine | machines". Obviously, possession relationship does not hold for the subjects and objects of all verbs. This method is only used as a work-around because the $\texttt{verb}_{\text{Active}}$ pattern cannot include the original object due to the problem of data sparseness.

Assigning Types to Proper Names

As discussed earlier, the system leverages a number of information sources to determine the types of proper names or to produce 'educated guesses' when types cannot be reliably determined. After the alias matching process, only types obtained from the most 'complete' alias are preserved. For example, in the article WSJ 13, the company '*New England Electric System*' is referred to later as both '*New England Electric*' and '*New England*'. At word level, the system is not able to determine the type of the entity, but provides the guessed types 'system' and 'electric' to the first two aliases based on the low proper name ratios. The last alias is recognized as a location based on information

[16] A pair of possession and *of*-prep queries only yield one result – the higher value of the normalized hit counts produced by the two queries.

provided by the WORDNET. After alias matching, only the first guess, 'system', is retained for the whole alias group, the guess 'electric' and the type 'LOCATION' are discarded as they are obtained from inferior sources. Unfortunately, 'system' is not the correct type, and there is no appositive or predication in the text that provide type information for the entity.

The system tries to overcome problems like this by further expanding the search scope for type information to the adjacency of the proper name aliases if an alias group has no definitive type assigned to it. Within a predefined window size (three sentences in the current implementation), the system extracts the head nouns of each 'simple' definite description that appears after any of the aliases in the group. The term 'simple' definite description refers to the definite descriptions that are either devoid of additional descriptive contents or have at most one adjective / past participle modifier. The extracted head nouns are further tested for gender/number compatibility against the alias group; the incompatible ones are removed from the candidate list. If a guessed type is offered by word-level analysis, it is also proposed as a candidate.

The auxiliary relationships, i.e. possessions, prepositions, and verb-argument relationships, are used to rank the candidates. The system identifies the relationships involving each of the aliases in the group, generates queries for each relationship and type candidate, execute them, and collects normalized query results. The type candidate that fits best in the original contexts of the aliases is assigned to the group. Figure 4.2 illustrates the details of the algorithm. In the current implementation, the threshold for minimum percentage of successful queries is set at 0.5, and an alias group needs to participate in at least 3 different relations in order to qualify.

4.3.3 Combination of Evidences

The general principle of this study is to make use of as much information as possible. In the previous sections various ways to obtain gender/number information are discussed. At word level, heuristics for gender/number may come from many distinct sources such as the WORDNET or the web. The number of evidences grows further when coreferential entities, such repeated mentions or aliases of proper names, are grouped together as clusters. It is clear that a method to weight and combine the various heuristics is essential for this study.

The Dempster-Shafer Theory of Evidence

The Dempster-Shafer (DS) Theory of Evidence (Shafer, 1976) provides a framework for representing uncertain evidence and for combining evidences from different sources. The DS theory can be considered a generalization of the Bayesian theory of subjective probability, with the most prominent characteristic being that under DS it is no longer necessary to assign weights to states where evidence is not available. As Dempster (2008) describes it,

> DS theory is founded on appending a third category "don't know" to the familiar dichotomy "it's true" or "it's false". More precisely, a DS model provides three non-

```
input  : a list of type candidates $Types$,
         a list of proper name aliases $Aliases$,
         a threshold for minimum percentage of successful queries $Threshold$.
output: the type to be assigned, or $NULL$ if none is applicable.

totalnumpatterns $\leftarrow 0$;
scoreset $\leftarrow \emptyset$;
hitset $\leftarrow \emptyset$;
foreach $a \in Aliases$ do
    rels $\leftarrow$ FindRelationships($a$);
    foreach $r \in$ rels do
        numpatterns $\leftarrow$ GetPatternCount($r$,$a$);
        totalnumpatterns $\leftarrow$ totalnumpatterns $+$ numpatterns;
        patternset $\leftarrow$ GetPatterns($r$,$a$);
        normpatternset $\leftarrow$ GetNormalizerPatterns($r$,$a$);
        for $p \leftarrow 1$ to numpatterns do
            patternresultset $\leftarrow \emptyset$;
            foreach $t \in Types$ do
                hc $\leftarrow$ ExecuteQuery(patternset $[p]$,$a$);
                nhc $\leftarrow$ ExecuteQuery(normpatternset $[p]$,$a$);
                patternresultset $[t] \leftarrow$ hc / nhc;
                if hc $> 0$ then hitset $[t] =$hitset $[t] + 1$;
            end
            normalize patternresultset so that $\sum$patternresultset $= 1$;
            foreach $t \in Types$ do scoreset $[t] =$scoreset $[t]+$patternresultset $[t]$;
        end
    end
end
besttype $\leftarrow \arg\max_t$ scoreset;
if hitset [besttype ] / totalnumpatterns $> Threshold$ then return besttype;
else return $NULL$
```

Figure 4.2: Algorithm for assigning types to proper name alias groups

> negative probabilities (p,q,r) with $p+q+r=1$ to the three categories of the modal triad "known to be true", "known to be false", and "don't know" associated with each assertion specified in the model. It remains true that every statement defined within the model is in fact either true or false, but "you", the DS analyst, is no longer restricted to p and q with $p+q=1$ as in Bayesian theory. Since probabilities to which "you" commit tentative belief are presumed to be evidence-based, a probability p is construed to represent "your" evidence "for" the truth of an assertion, while probability q measures evidence "against", and probability $r=1-p-q$ quantifies residual ambiguity. (Dempster, 2008)

The "don't know" category is very helpful since it allows withholding judgement on what is not supported by evidence, instead of having to overcommit under the Principle of Insufficient Reason. It follows naturally that the DS theory has the ability to model the constriction of a hypothesis set as evidence accumulates.

Under the DS theory, a decision problem is modeled as a finite set of basic hypotheses that are exhaustive and mutually exclusive. This set, denoted $\Theta = \{\theta_1, \ldots, \theta_K\}$, is referred to as the *frame of discernment*. The powerset 2^Θ is the set of all subsets of Θ, including the empty set and Θ itself. For instance, $\Theta_g = \{m(asculine), f(eminine), n(eutral), p(lural)\}$ represents the alternatives for gender/number assignments, which has a corresponding powerset $2^{\Theta_g} = \{\emptyset, m, f, n, p, m \vee f, m \vee n, \ldots, n \vee p, m \vee f \vee n, \ldots, f \vee n \vee p, m \vee f \vee n \vee p\}$.

Following are the basic concepts of the theory:

Basic probability assignment

Given a frame of discernment Θ, the function $m : 2^\Theta \mapsto [0;1]$ is called a *basic probability assignment* if it satisfies the following:

$$m(\emptyset) = 0 \tag{4.33}$$

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{4.34}$$

It is worth noting that $m(A)$ represents the *belief* committed exactly to A and does not support any of its strict subsets. In other words, a positive value of $m(masculine \vee feminine)$ indicates that there is evidence that the corresponding entity refers to a human being, but tells nothing about the belief whether the person is male or female – such information is provided by $m(masculine)$ and $m(feminine)$, respectively. In comparison, a Bayesian system would have to distribute the probability to the subsets, therefore specifying more than what the evidence supports.

Belief function

Given the basic probability assignment m, the *belief function* $\text{Bel} : 2^\Theta \mapsto [0;1]$ can be defined as a function that assigns a value in $[0,1]$ to every nonempty subset B of Θ, such that for any

$B \subset \Theta$:

$$\mathrm{Bel}(B) = \sum_{A \subseteq B} m(A) \tag{4.35}$$

A subset A for which $m(A) > 0$ is called a *focal element* of the belief function. Some additional properties of Bel include:

$$\mathrm{Bel}(\emptyset) = 0 \tag{4.36}$$

$$\mathrm{Bel}(\Theta) = 1 \tag{4.37}$$

and, for any collection $A_1, A_2, \ldots, A_n (n \geq 1)$ of the subsets of Θ:

$$\mathrm{Bel}(A_1 \cup A_2 \cup \ldots \cup A_n) \geq \sum_{I \subseteq \{1,2,\ldots,n\}, I \neq \emptyset} (-1)^{|I|+1} \mathrm{Bel}(\bigcap_{i \in I} A_i) \tag{4.38}$$

Doubt, Plausibility, and Belief Interval

In addition to the belief function, Shafer (1976) offers two more measures, *doubt* and *plausibility*, to provide further details on level of certainty entailed by the evidence. For a nonempty subset B of Θ, the doubt function $\mathrm{Dou}(B) = \mathrm{Bel}(\overline{B})$ reflects the amount of support that the evidence provides for the negation B. The plausibility function, $\mathrm{Pl}(B)$, reflects the maximum amount of belief in B that is allowable by the current evidence.

$$\mathrm{Pl}(B) = 1 - \mathrm{Dou}(B) = \sum_{A \cap B \neq \emptyset} m(A) \tag{4.39}$$

It is trivial to prove that $\mathrm{Bel}(B) + \mathrm{Bel}(\overline{B}) \leq 1$, which leads to $\mathrm{Bel}(B) \leq \mathrm{Pl}(B)$. The width of the *belief interval* $[\mathrm{Bel}(B), \mathrm{Pl}(B)]$ reflects the amount of belief that is committed to elements that have non-empty intersections with B but are not its subsets.

Dempster's Rule of Combination

For a given frame of discernment Θ, it is possible to alter prior beliefs by incorporating new evidences from different sources. Consider two belief functions Bel_1 and Bel_2 defined on Θ and their corresponding basic probability assignments m_1 and m_2. Let A_i and B_j denote the focal elements of Bel_1 and Bel_2 respectively. Then m_1 and m_2 can be combined to obtain the belief mass committed to $C \subset \Theta$ following Dempster's rule of combination (Shafer, 1976), formulated as follows:

$$m(C) = \frac{\sum_{i,j,A_i \cap B_j = C} m_1(A_i) m_2(B_j)}{1 - \sum_{i,j,A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)} \tag{4.40}$$

Alternatively, the combination can be expressed as the orthogonal sum ($\oplus$), allowing Equation 4.40 to be rewritten in a simpler form:

$$\mathrm{Bel}_3 = \mathrm{Bel}_1 \oplus \mathrm{Bel}_2 \tag{4.41}$$

The numerator of Equation 4.40 is the sum over intersections between the hypotheses being combined. The denominator is a normalization factor representing the non-contradictory portions of the hypotheses. The *weight of conflict* between the belief functions, $\mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2)$, is also determined from the denominator.

$$\mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2) = \ln \frac{1}{1 - \sum\limits_{i,j,A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)} \tag{4.42}$$

Dempster's rule of combination is a symmetrical function, and can be easily extended to combine multiple belief functions through repeated pairwise operations. The combination of n belief functions $\mathrm{Bel}_1, \mathrm{Bel}_2, \mathrm{Bel}_2, \ldots, \mathrm{Bel}_n$ can be formulated as:

$$\bigoplus_{i=1}^{n} \mathrm{Bel}_i = ((\mathrm{Bel}_1 \oplus \mathrm{Bel}_2) \oplus \mathrm{Bel}_3) \ldots \oplus \mathrm{Bel}_n \tag{4.43}$$

Individual evidences may be weighted (Guan & Bell, 1993; Bell, Guan, & Shapcott, 1998; Bean, 2004) to reflect their relative strength. Let m be a basic probability assignment defined on the frame of discernment Θ. Let $\alpha \in [0,1]$ denote the discounting factor. The discounted basic probability assignment m^{α} can be formulated as:

$$m^{\alpha}(A) = (1-\alpha)m(A), \forall A \subset \Theta \tag{4.44}$$

$$m^{\alpha}(\Theta) = (1-\alpha)m(\Theta) + \alpha; \tag{4.45}$$

As shown above, the discounting operation effectively transfers belief from focal elements to Θ, which stands for ignorance.

A general condition for the applicability of Dempster's rule of combination is that the evidences being combined come from independent sources. The concept of independence has been the subject of many research projects. For example, Liu and Hong (2000) makes the distinction between original information level and target information level and asserts that independence needs to be assured at the original information level. However, a recent study by Altinçay (2006) suggests that the independence requirement may be relaxed for practical multi-classifier systems. While the evidences being combined in this study can be generally considered independent, Altinçay's experiment results allow us to be less concerned about the issue of independence.

Combination of Gender/Number Heuristics

Instead of using words or proper names as the minimum unit of gender/number information, the system developed in this study operates on evidences. Each evidence is represented by a tuple $e = < l, i, w, m >$, where l, i, w, and m denote respectively the lexical source, the origin of the information used for decision, a predefined weight associated with the information source, and the basic probability assignment. This design allows evidences to be safely combined at both individual word level and for coreferential clusters. When information is needed for a specific word or proper name, evidences pertaining to the entity are first discounted by applying equations 4.44 and 4.45, and then combined using Dempster's rule of combination. In the cases where clusters of entities need to be evaluated, duplicate evidences (identified on the basis of lemmas and information sources) are removed prior to combination. For example, in a cluster containing both '*John Doe*' and '*John*', web-based evidences for *John* are only counted once.

The most common use of gender/number information is assessing whether a discourse entity can join a specific cluster. This is usually done by computing the weight of conflict (Con) between the combined evidences from the cluster and the evidences available to the individual entity. In alias matching, the Con statistic is used to rank competing aliases having the same Levenshtein distance to a previous mention (e.g. '*Mr. Doe*' vs '*Mrs. Doe*' given the antecedent '*John Doe*'). For pronominal anaphora resolution, the statistic is compared to a fixed threshold – if the value of Con is larger than the threshold (currently set to $\ln 2$), the pronoun is considered unfit for the cluster due to gender/number mismatch.

4.4 Pronominal Anaphora Resolution

In a nut shell, the system approaches pronominal anaphora resolution with a rule-based algorithm that operates on coreference clusters. Besides gender/number agreement, other major factors considered by the system include syntax-based salience, guidance provided by the centering theory, and semantic-based selectional restrictions. The basic operations of the system can be described as follows:

1. Initialize. During the initialization process, the system identifies word-level gender/number information for all nominal entities processed by the system, including common nouns, proper names, relative pronouns, and personal pronouns, and identifies word-level type information for proper names. Various text-level relationships, such as apposition, predication, and antecedents of NP-* elements[17], are also identified.
2. Establish coreference clusters for proper names and resolve relative pronouns.
3. Assign types to proper names based on information in the text, as illustrated in Figure 4.2.
4. Identify pleonastic *it* instances. This process is detailed in Chapter 5.
5. Propose potential antecedents for anaphoric personal pronouns and assign a salience score to each potential antecedent based on its relative position to the anaphor on the dependency structure.
6. Apply gender/number constraints to the identified antecedents.
7. If one of the highest-ranked potential antecedents is also a personal pronoun, it is selected as the antecedent. This can be seen as a technical realization of the centering theory's preferences for pronominalizing the backward-looking center and center-continuation. If the antecedent is

[17] As mentioned earlier, the system does not attempt to resolve all NP-* elements, many of which also do not have any antecedent.

not resolved yet, its list of potential antecedents are reduced to the intersection of the original list and that of the anaphor.

8. Perform additional gender check for plural pronouns (discussed in Section 4.4.2).
9. Trim the potential antecedents so that only the highest-ranked ones are retained.
10. Apply selectional restrictions to the remaining potential antecedents. This process is modeled with possessor-possessee and verb-argument queries against the web.
11. Finalize. This step assigns antecedents to any pronouns that have more than one potential antecedents or have no potential antecedent. If the competing potential antecedents are located in one of the preceding sentences, the one that is closer to the top of the dependency structure and is functioning as surface subject is preferred. If they are located in the same sentence as the anaphor, the one having the shortest linear distance towards the anaphor is preferred. If the anaphor has no potential antecedent, a linear search is performed linearly towards the left of the anaphor, until an antecedent with matching gender/number is identified.

4.4.1 Antecedent Identification

As discussed in Section 2.3, while most recent studies adopt either factor-based or machine-learning based approaches to pronominal anaphora resolution, Hobbs' (1978) naive algorithm remains competitive performance-wise (Mitkov & Hallett, 2007; Tetreault, 2001). The most significant advantage of the Hobbs algorithm is its simplicity – the algorithm follows the simple heuristics that antecedents functioning as surface subjects are preferred over those functioning as objects and adjuncts, and that local antecedents are preferred over remote ones, with the 'closeness' of an antecedent expressed in terms of clause hierarchy instead of linear distance. This study adopts an algorithm that follows the same heuristics for proposing antecedent candidates. Unlike Hobbs' original algorithm, the algorithm proposed in this study works on dependency structures, which makes the information hierarchy in a sentence more obvious, and does not strictly follow the left-to-right, breadth-first search order as proposed by Hobbs. As the system detects and resolves certain empty categories (see Section 4.3.2), the algorithm also has more accurate information regarding the distance between a potential antecedent and the anaphor. For example, consider the previously discussed (4.7), repeated here with additional phrase structure information:

(4.7) The SEC documents describe those chips, which are made of gallium arsenide, as (S$_1$ (NP$_2$ *)$_2$ (VP$_3$ being (ADJP$_4$ (ADJP$_5$ so fragile and minute)$_5$ (SBAR$_6$ (S$_7$ *they* will require special robotic handling equipment)$_7$)$_6$)$_4$)$_3$)$_1$.

Resolving the NP-* element before *being* is crucial for the resolution of the pronoun *they*. If the NP-* is not identified or is incorrectly resolved to *The SEC documents*, the algorithm will identify *The SEC documents* as being more salient than the correct antecedent, *those chips*[18]. Like the original

[18] According to the results of web queries, *document* is also more closely related to the verb *require* than *chip* is.

Hobbs algorithm, the algorithm proposed in this study is also able to resolve cataphoric cases like (4.16)[19], but without the need for a dedicated treatment.

(4.16) As *he* stands on a hill at the beginning of a six-day motor expedition ... Stevens surveys the view and thereby provides a self-portrait ... WSJ 2149:9

The algorithm mainly targets third-person personal pronouns. A special module is also constructed to handle first-person pronouns in quoted speech. Since the corpus consists of mainly written proses, second-person pronouns are not considered as they are mostly generic.

Given a pronoun, the algorithm walks up the dependency tree to find the nearest verb. Once the verb is located, the system collects its subject, object(s), and preposition phrase complements/adjuncts, and iterates the process by navigating further up the dependency tree. It is not always possible to find the subject under the verb, as there might be a few levels of cascading verb phrases between the subject and the object, as in (4.17):

(4.17) He added: "(S$_1$ (NP$_2$ Every paper company management)$_2$ (VP$_3$ has (S$_4$ (NP$_5$ *)$_5$ (VP$_6$ to (VP$_7$ be (VP$_8$ saying (PP$_9$ to *itself*)$_9$, '(S$_a$ Before someone comes after me, I'm going to go after somebody)$_a$)$_8$)$_7$)$_6$)$_4$)$_3$)$_1$.' " WSJ 317:63

The dependency structure for the relevant portion of the sentence is illustrated in Figure 4.3. In the dependency structure generated by head percolation, the head nodes of the VP elements are stacked together and the subject *Every paper company management* is attached under the top verb, *has*. The system uses a simple finite state machine (Table 4.5) to find its way up the chain of verbs in order to locate the subject. The navigation stops when no further transition can be made; or when the parent node has an object or predicate, as in the case of '*It* (VP$_1$ *is* (NP-PRD$_2$ *fun*)$_2$ (S$_3$ (NP$_4$ *)$_4$ (VP$_5$ *competing as a private company*)$_5$)$_3$)$_1$', where navigation from *competing* to *is* is interrupted because of the presence of the predicate nominal *fun*. Applying the rules in Table 4.5 to (4.17), the verbs *saying-be-to-has* are identified as a single chain, with the subject attached to the top verb. This process also implicitly resolves the NP-* element between *has* and *to* that is present in the original CFG parse tree. While the fixed expression *have to* is not exactly a modal auxiliary, considering it as having no internal structure has little consequence for the purpose of this study. In fact, most of the other semi-auxiliaries (Quirk et al., 1985, section 3.40) can be processed in the same manner.

Subjects and objects (as well other complements/adjuncts, which are treated as objects) collected from the same verb chain are given different weights. The exact salience value of a potential antecedent j with regard to the anaphor i is calculated as $s_{i,j} = w - \delta \times \mathrm{D}(i,j)$, where $w \in \{w_{\text{subject}}, w_{\text{object}}\}$ is the weight constant chosen based on the surface function of the potential antecedent, δ is the constant discounting factor, and $\mathrm{D}(i,j)$ is the distance function that returns the vertical distance between the two nodes, expressed in the number of intervening verbs between

[19] Mitkov and Hallett (2007) are probably mistaken when they mention that the Hobbs algorithm does not resolve cataphoric cases (cf. Hobbs, 1976, page 19).

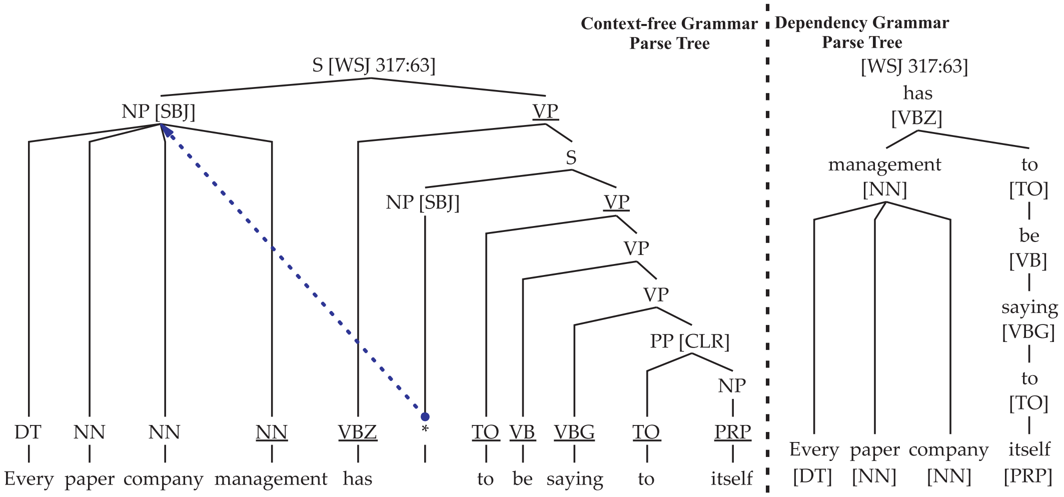

Figure 4.3: Structures of the relevant fraction of example (4.17)

Current Node	Higher-level Node				
	MD	TO	*be*	*have*	*do*
VB	X	X			X
VBG			X		
VBN			X	X	
TO	X			X	

Table 4.5: Transition table for bottom-up verb chain navigation. Headers in CAPITAL, such as VB or MD, refer to part-of-speech tags. Rules for the verbs *be*, *have*, and *do* also apply to their inflected forms.

them. More specifically, given any node, the system identifies the verb chain associated with it and use the count of verbs dominating the top verb, l, to represent the 'level' of the node. The distance function simply calculates the difference in levels between the anaphor and the potential antecedent, or $\mathrm{D}(i,j) = l_i - l_j$.

Unlike the Hobbs algorithm, which performs a breadth-first search, the system gives the same salience values to the subjects/objects and their modifiers. For example, consider (4.18):

(4.18) Not only is (NP$_1$ development (PP$_2$ of (NP$_3$ (NP$_4$ the new company's)$_4$ initial machine)$_3$)$_2$)$_1$ tied directly to Mr. Cray, so is *its* balance sheet. WSJ 18:2

The dependency structure of (4.18) is illustrated in Figure 4.4. The system extract the modifying components to the head (*development*) of the subject NP, including both *initial machine* and *the new company*, and gives all three elements the same salience value, $s = w_{\text{subject}} - \delta \times \mathrm{D}(its,development) = w_{\text{subject}} - \delta \times 1 = w_{\text{subject}} - \delta$.

The system identifies conjunctive clauses during head percolation and assigns custom tags to them. As the algorithm walks up the dependency tree, it also looks for conjunctive clauses that appear as left-siblings to the node being visited. These conjunctive clauses are processed as if

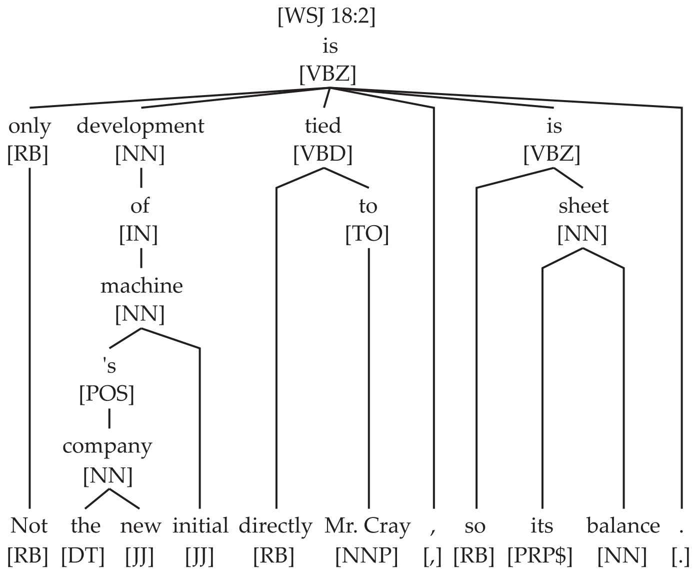

Figure 4.4: Dependency structure of example (4.18)

they are located on the path. For example, in (4.19), *the U.S.* is assigned a salience value of $s = w_{\text{object}} - \delta \times \mathrm{D}(it,U.S.) = w_{\text{object}} - \delta \times 0 = w_{\text{object}}$. Similarly, in (4.20), *Mr. Cray* is assigned a salience value of w_{subject}. The dependency structures of (4.19) and (4.20) are illustrated in Figure 4.5, with the tags `VBC` and `SBARC` used to mark the corresponding conjunctive clauses.

(4.19) Japan not only outstrips the U.S. in investment flows but also outranks *it* in trade with most Southeast Asian countries (although the U.S. remains the leading trade partner for all of Asia). WSJ 43:34

(4.20) Documents filed with the Securities and Exchange Commission on the pending spinoff disclosed that Cray Research Inc. will withdraw the almost $100 million in financing it is providing the new firm if Mr. Cray leaves or if the product-design project *he* heads is scrapped. WSJ 18:3

The above-mentioned scheme does not provide complete coverage for all potential antecedents within the sentence. For example, in (4.21), the antecedent *Minneapolis-based Cray Research* is embedded deep inside an adjunct clause.

(4.21) While many of the risks were anticipated when Minneapolis-based Cray Research first announced the spinoff in May, the strings *it* attached to the financing hadn't been made public until yesterday. WSJ 18:6

In order to handle cases like (4.21), the system uses another procedure to search for verbs that have not been previously visited and propose their subjects, objects, and pp-complements/adjuncts as potential antecedents. Antecedents collected in this process are all given the same salience value, $s = w_{\text{adjunct}}$.

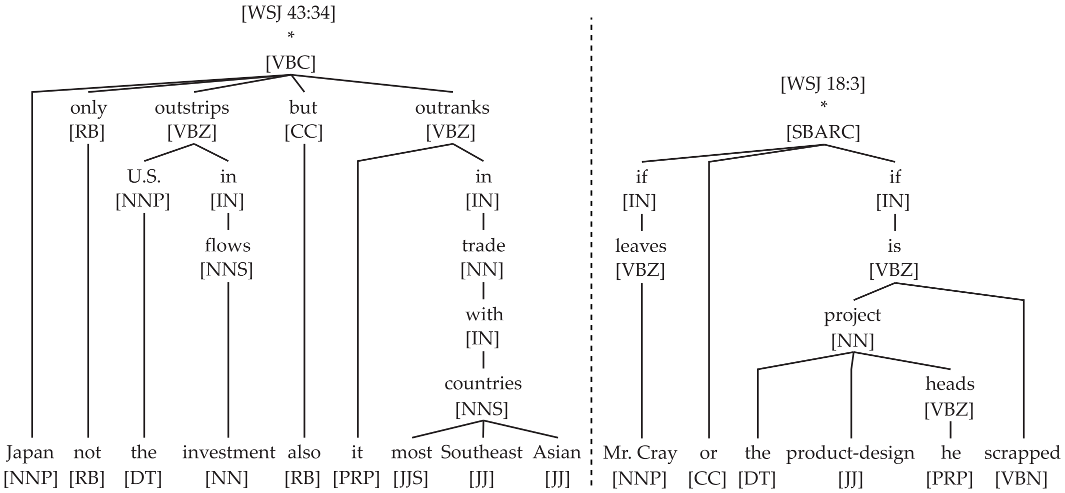

Figure 4.5: Dependency structures of the relevant fragments of examples (4.19) and (4.20)

Finally, the system also attempt to identify potential antecedents of reflexive and possessive pronouns located inside preposition phrases or possessive pronouns in conjunctive noun phrases by searching inside the noun phrases that dominate them. Examples of these configurations include (4.22), (4.45), and (4.24). The algorithm uses essentially the same techniques proposed by Hobbs (1978). However, the Hobbs algorithm relies on a parser that generates $\bar{\mathrm{N}}$ nodes, which are not directly annotated in the WSJ corpus. As noted by Hobbs, the difference between PP arguments and PP adjuncts, the former of which is attached to the $\bar{\mathrm{N}}$ node and the latter to the NP node, is critical to the interpretation of phrases such as '*a driver of his car*' and '*a driver in his car*'. A recent study by Merlo and Ferrer (2006) indicates that the function of NP-attached PPs can be (relatively) reliably determined based on the preposition alone. Following their analysis, the whole WSJ corpus is surveyed and 21 prepositions[20] are selected based on their overall frequency and likelihood of leading PP arguments. Setting aside *of*, which accounts for half of all NP-attached PPs and is used almost exclusively as argument, the remaining 20 prepositions account for about 85% of all (non-*of*) argument uses with a precision of 92%.

(4.22) Video Tip: Before seeing "Sidewalk Stories," take a look at "City Lights," (NP$_1$ (NP$_2$ (NP$_3$ Chaplin's)$_3$ Tramp)$_2$ (PP$_4$ at (NP$_5$ *his* finest)$_5$)$_4$)$_1$. WSJ 39:43

(4.23) A … hero sets off for snow country in search of (NP$_1$ (NP$_2$ an elusive sheep)$_2$ (PP$_3$ with (NP$_4$ (NP$_5$ a star)$_5$ (PP$_6$ on (NP$_7$ *its* back)$_7$)$_6$)$_4$)$_3$)$_1$ … WSJ 37:12

(4.24) Mr. Nixon met (NP$_1$ (NP$_2$ Mr. Bush)$_2$ and (NP$_3$ *his* national security adviser, Brent Scowcroft,)$_3$)$_1$ before coming to China on Saturday. WSJ 93:20

[20] The identified prepositions are *of*, *for*, *on*, *with*, *from*, *as*, *by*, *about*, *over*, *between*, *than*, *against*, *like*, *via*, *into*, *per*, *toward*, *through*, *without*, *out*, and *except*, in the order of their overall frequencies in the WSJ corpus. Some of the prepositions (such as *on*, 67% argument, and *over*, 75%) are less consistently used as arguments than others, but are nevertheless included due to their overall frequency.

Antecedents identified in these configurations are given a salience value of w_{subject}. In other words, they are treated as surface subjects at the same level as their corresponding antecedents.

After walking through the current sentence, the algorithm performs a 'simplified' search on the previous three sentences[21] – only the top-level clauses of the sentences and full clauses (both with or without the complementizer, or led by a *wh*-adverb) are examined. All collected entities are weighted equally regardless of their grammatical functions. The final salience values are calculated as $s_{i,j} = w_{\text{remote}} - \delta \times \mathrm{D}'(i,j)$, where w_{remote} is the weight constant and $\mathrm{D}'(i,j) = \text{SentenceId}_i - \text{SentenceId}_j$ is the distance between the anaphor and the antecedent expressed in number of sentences.

4.4.2 Web-based Shallow Semantic Analysis

There is little doubt that pronominal anaphora resolution could benefit from additional semantic information – the role of semantics can be quite prominent in many cases, such as (4.25):

(4.25) Once inside, she spends nearly four hours measuring and diagramming each room in the 80-year-old house, gathering enough information to estimate what it would cost to rebuild *it*. WSJ 766:15

The last instance of *it* in (4.25) is unambiguous because it is well-understood that a room is much more unlikely to be the patient of *rebuild* in comparison to a house. But the level of reasoning required to resolve a pronoun is often more complex than simple possessor-possessee or predicate-argument relationships. Consider example (1.3) by Mitkov (2002), repeated here as (4.26):

(4.26) a. The soldiers shot at the women and *they* fell.

b. The soldiers shot at the women and *they* missed.

The key information needed to resolve the pronouns in (4.26) is the causal relationship between the actions *shot at*, *fell*, and *missed*.

Since Dagan and Itai (1990), a number of researchers have attempted to implement selectional restrictions using corpus-based or web-based co-occurrence statistics. However, there does not seem to be a consensus on the usefulness of such methods. There are a few obvious issues related to the use of co-occurrence statistics. First, it is very difficult to capture high-level semantic relationships with co-occurrence patterns. Bean (2004) indicates a possible direction for acquiring the kind of knowledge needed to resolve the pronouns in (4.26) – his CFNET captures the co-occurrence statistics of the different verb-argument patterns from a corpus, such as `agent`-*shoot* and `agent`-*miss*. However, as noted by Bean himself, this approach may not work well for articles that are not domain-specific, as the data would be too sparse to produce useful statistics. It is possible to obtain the co-occurrence statistics of verb-argument patterns by querying the web. For example, the Yahoo

[21] If a sentence is composed of multiple full clauses at the top level, each clause is counted as as a separate sentence.

search engine returns a page count of 268 for the query `"shot at her" NEAR "she fell"`, and 8 for `"shot at her" NEAR "he fell"`, indicating that the same kind of technique (exploiting gender/number mismatch) used earlier in the study to determine the subject of *as*-predications may be useful in identifying implicit causal relationships. Second, simple co-occurrence patterns are unable to represent the immediate linguistic contexts of the anaphor. The context in which a verb-argument relationship lies is often more important in deciding the felicity of an antecedent. For example, in (4.27), the decisive factor is that the verb *join*, in the sense it is used for in this sentence, requires that the party being joined is performing some sort of action. The (non)coreferent dependency path by Bergsma and Lin (2006) is a promising step towards correctly representing this kind of knowledge. Given a large-enough corpus, Bergsma and Lin's system may be able to identify the path `join N as Pro Verb` as coreferent and `N join ...as Pro Verb` as non-coreferent[22].

(4.27) The dancers were joined by about 70 supporters as *they* marched around a fountain not far from the Mayor's office, chanting: "Giuliani – scared of sex! Who's he going to censor next?" Kehler et al. (2004, ex. 4)

The third issue is the reliability of verb-argument co-occurrence statistics. Sine the higher-level semantics are not taken into consideration, the verb-argument statistics cannot be seen as a reliable source of antecedent-anaphor agreement. As noted out by Kehler et al. (2004), the term *supporter* is much more likely to co-occur with *march* than *dancer* is[23]. Thus following the statistics leads to the wrong antecedent. In fact, the verb-argument statistics is largely irrelevant in this case – the only useful information provided by the statistics is that dancers can march, since the verb-argument query yields non-zero count. However, even this information is not always reliable. For example, while it is generally understood that supercomputers have chips (as in the sense of integrated circuits), the queries "supercomputer 's chip|chips" or "chip|chips of the supercomputer|supercomputers" yield no hit. Generalizing the concept *supercomputer* to *computer* solves the problem. But it is difficult to determine where should the process of generalization end, how to deal with the different word senses, and how should the results obtained through generalization be compared to the unmodified queries generated from competing antecedents.

The relative low reliability of co-occurrence statistics leads to the most important issue of all – greater care is needed in integrating it with the other heuristics. While examples like (4.27) lead Kehler et al. (2004) to believe that co-occurrence statistics is of little use, later studies such as those by Yang et al. (2005) and Bergsma and Lin (2006) indicate that they might have reached the conclusion too early. The performance gains from co-occurrence statistics reported in these studies suggest that the utility of this feature is closely related with when and how it is applied during the resolution process. In the approach proposed in this study, the strategy is to 'leave room' for co-

[22] Obviously, a prerequisite is that the function of *as* (e.g. serving to indicate purpose or reason, or to indicate temporal relationship) can be reliably determined, which is a difficult problem on its own.

[23] Kehler et al. (2004) report that the normalized score of *dancer-march* obtained using the AltaVista search engine is about 15% of that of *supporter-march*. On the Google N-grams corpus, the score of the former is only about 3% of the latter.

occurrence statistics but at the same time prevent it from overriding more important salience factors. The most common situations where multiple antecedents are tied include candidates obtained from a chain of preposition/possession relationships, as illustrated in Figure 4.4, or when there are multiple candidates in a preceding sentence. In these cases, a pure syntax-based salience model is generally considered less helpful than co-occurrence statistics.

The system gathers two different kinds of co-occurrence statistics. The first one is specific to plural pronouns, which are converted to two groups of singular forms, the personal group *He* | *She* | *he* | *she* | *who*, and gender-neutral group *It* | *it* | *which* and tested against the verb-argument relationships (using the patterns $\texttt{verb}_{\text{Active}}$ and $\texttt{verb}_{\text{Passive}}$ described on page 85) identified from all members of the coreference cluster that the pronoun is assigned to[24]. The result counts are normalized by the counts of the respective pronouns. If one of the relationships[25] demonstrates to a strong tendency of person/neutral preference ($r = \min(n_{person}, n_{neutral}) / \max(n_{person}, n_{neutral}) < 0.25$ in the current implementation), the preference is used to eliminate antecedents whose singular forms exhibit a different gender distribution. The second kind of co-occurrence statistics covers both possession and verb-argument relationships. The same algorithm used to assign type labels to proper names (as illustrated in Figure 4.2) is used to evaluate the semantic compatibility of potential antecedents.

The current implementation does not include the more complex corpus-based co-occurrence models such as the coreferent path or relation co-occurrence. While the former, especially its negative version, is certainly helpful, the relevant examples in the annotated portion of the corpus are too sparse to reveal how it can be applied. On the other hand, analysis of the corpus indicates that there are other important issues that remain unsolved in the current approach. For example, consider the first pronoun *its* in (4.28):

(4.28) Toni Johnson pulls a tape measure across the front of what was once a stately Victorian home. A deep trench now runs along *its* north wall, exposed when the house lurched two feet off its foundation during last week's earthquake. WSJ 766:1-2

The correct antecedent is *a stately Victorian home*[26], embedded deep inside the preceding sentence. The competing concept, *A deep trench*, is not only more salient in terms of position, but is also compatible in terms of semantics. A trench also has more than one walls. In fact, querying the Google N-gram corpus reveals that using normalized counts as standard, *trench* is actually more closely related with the term *wall* than *home* is. To the best of my knowledge, none of the existing shallow semantic approaches help in this case. For example, it is unlikely that enough samples can

[24] It is possible that the pronoun being examined has been identified as the antecedent of one or more pronouns during the previous step.

[25] All antecedents are retained if multiple relationships are found and they demonstrate different preferences.

[26] More precisely, it should be *what was once a stately Victorian home*, or simply *what*. The system recognizes the predication relationship and uses the more informative part as the potential antecedent.

be collected to support a non-coreferent path such as N *run along* Pro’s *walls*[27]. It would also be unwise to punish non-definite antecedents, as they often serve as legitimate antecedents in bound anaphora cases or in generic sentences.

The centering theory appears to offer a solution to this puzzle – provided that clauses, not full sentences, are recognized as units of utterances. Had *A deep trench* been chosen as the antecedent, the clause ‘*A deep trench now runs along its north wall*’ would be completely irrelevant in the context, which is obviously undesirable. However, this kind of constraint is very difficult to implement technically, since the system has to make sure that *A deep trench* is not actually a realization of one of the previously mentioned entities, i.e. it is not involved in an associative anaphoric relationship. Even then, it is still possible that a seemingly ‘irrelevant’ sentence is actually licensed by higher-level discourse structures. For example, it may be used to elaborate a topic brought up in a previous sentence, as in (4.29).

(4.29) That’s not to say that the nutty plot of “A Wild Sheep Chase” is rooted in reality. It’s imaginative and often funny. A disaffected, hard-drinking, nearly-30 hero sets off for snow country in search of an elusive sheep with a star on its back at the behest of a sinister, erudite mobster with a Stanford degree. *He* has in tow his prescient girlfriend, whose sassy retorts mark her as anything but a docile butterfly. WSJ 37:10-13

4.5 Anaphoricity Determination of Definite Descriptions

It is well-known that, unlike pronouns, a significant portion of definite description uses are ‘discourse-new’ in the sense that they are non-anaphoric and have no coreferential antecedents[28]. As shown in Section 3.3, close to 60% of the definite descriptions annotated for this study are classified as non-anaphoric, over 80% of which are also discourse-new. Similar percentages of discourse-new definite descriptions have been reported in previous studies. For example, Poesio and Vieira (1998) reports that 47% of the definite descriptions studied in their first annotation experiment are discourse-new; Fraurud (1990) finds that in a corpus of Swedish prose, over 60% of the definite descriptions do not have coreferential antecedents.

Given the prevalence of non-anaphoric definite descriptions, the importance of anaphoricity determination is also widely recognized. While there are some disagreement about the utility of currently-available discourse-new detectors for coreference resolution[29], the issue is rather of technical nature than methodological. Depending on the corpus under investigation and the implemen-

[27] If the system allows both terminal nouns to be generalized to part-of-speech tags, a path like N *run along* Pro’s N will capture many coreferent relationships, such as ‘*the river runs along its banks*’.

[28] Note that according to the notion of anaphora developed in Chapter 3, non-anaphoric definite descriptions can nonetheless have coreferential antecedents.

[29] The results reported by Ng and Cardie (2002a) seems to indicate that their discourse-new detector leads to inferior system performance for common noun anaphors. Poesio, Uryupina, Vieira, Alexandrovkabajov, and Goulart (2004) and Poesio et al. (2005) show that a state-of-the-art discourse-new detector does contribute positively to the tasks of coreference resolution.

tation of the coreference resolution system, the chance of the system incorrectly assigning an antecedent for a discourse-new entity may be relatively low. Therefore a coreference resolution system benefits only from a high-precision anaphoricity detector – false-positive discourse-new entities introduced by the detector would easily cancel the limited benefit it offers. On the other hand, however, the recall of an anaphoricity detector can be crucial for the purpose of associative anaphora resolution, primarily due to the lack of effective treatment for associative anaphora[30]. Without anaphoricity detection, a significant amount of the discourse-new entities would be assigned an associative antecedent.

Since Vieira's (1998) original empirical study, a few other researchers have tackled the issue of definite description anaphoricity. Their proposals are briefly reviewed in Section 2.4.1 and are also analyzed by Poesio, Uryupina, et al. (2004). Among these methods, the one by Uryupina (2003) is the closest to the approach proposed in the current study, both in terms of underlying concepts and source of information. Aside from the ±*discourse_new* classifier, Uryupina also trains a ±*unique* classifier to recognize the uniquely-referring expressions, which contains many of the definite descriptions considered by this study as non-anaphoric but nevertheless have coreferential antecedents[31]. Like Uryupina's approach, the system developed in this study also combines structural heuristics and web-based statistics to identify non-anaphoric definite descriptions. However, as the ensuing discussion will show, the system is both simpler in the sense that it only uses a straightforward set of hand-coded rules, and at the same time more elaborate in terms of the employed syntactic heuristics and the design of the query patterns.

4.5.1 Syntactic Indicators of Anaphoricity

As noted by Hawkins (1978) and Vieira and Poesio (2000), among others, non-anaphoric definite descriptions often exhibit certain syntactic structures. For example, when a definite description is post-modified by a restrictive relative clause, it is usually not anaphoric. The system considers the following grammatical heuristics:

PostMod/Complement The definite description is post-modified or has a complement clause

RRC The definite description is modified by a restrictive relative clause or a verbal phrase

(4.30) Alan F. Shugart, currently chairman of Seagate Technology, led the team that developed the disk drives for PCs. WSJ 22:14

(4.31) Because of the difficulty of assessing the damages caused by the earthquake, Aetna pulled together a team of its most experienced ... 766:35

(4.32) He said Mexico could be one of the next countries to be removed from the priority list because of its efforts to craft a new patent law. WSJ 20:17

[30] Existing research on associative anaphora resolution report F-measures in the range of 30%-40%, cf. Section 2.4.2.

[31] Bean's (2004) 'existential' category also covers non-anaphoric coreferential cases. It covers associative cases as well.

(4.33) Yale is one of the few medical institutions conducting privately funded research on fetal-tissue transplants. 47:20

Complement The definite description has a complement clause

(4.34) "We look upon this as a great opportunity to prove the fact that we have a tremendous management team," he said. WSJ 109:38

(4.35) "We have no problem to our freight service at all expect for the fact businesses are shut down." WSJ 1803:22

(4.36) Options give a holder the right, but not the obligation, to buy or sell a security at a set price within a set period of time. 1438:13

(4.37) "The theory is that Seymour is the chief designer of the Cray-3, and without him it could not be completed. WSJ 18:8

(4.38) "Now," says Joseph Napolitan, a pioneer in political television,"the idea is to attack first, last and always." WSJ 41:4

(4.39) The reason: Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold. WSJ 34:27

Prep The definite description is modified by a prepositional phrase (with the exception of *by*)

(4.40) The survival of spinoff Cray Computer Corp. as a fledgling in the supercomputer business appears to depend heavily on the creativity – and longevity – of its chairman and chief designer, Seymour Cray. WSJ 18:1

Close Apposition The definite description is in close apposition with a proper name

(4.41) In the film classic "Twelve Angry Men," the crucible of deliberations unmasks each juror's bias and purges it from playing a role in the verdict. WSJ 1267:23

NNP+NN PreMod The definite description is modified by a combination of common noun and proper name

(4.42) The new company said it believes there are fewer than 100 potential customers for supercomputers priced between $15 million and $30 million – presumably the Cray-3 price range. WSJ 18:22

CD+Head$_{\text{singular}}$ The definite description is composed of a singular head noun and a cardinal number as modifier

(4.43) They fell into oblivion after the 1929 crash. WSJ 34:9

Since it is of less interest to an anaphora resolution system to identify the different subtypes of non-anaphoric definite descriptions, the system subsumes the first three syntactic heuristics into one group. The syntactic patterns proposed by Vieira (1998, section 4.4.2), as listed below (rearranged), already cover the majority of the post-modification cases.

$$\texttt{(NP } \textit{the} \texttt{ premodifiers head (PHRASE ...) ...)} \tag{4.46}$$

$$\texttt{(NP (NP } \textit{the} \texttt{ premodifiers head) (PHRASE ...))} \tag{4.47}$$

$$\texttt{PHRASE} \in \{\texttt{SBARQ, SBAR, S, VP, PP, WHPP}\}$$

There may be zero, one, or more pre-modifying components. Nonrestrictive relative clauses, which are usually marked with a preceding comma, are not considered. These two patterns covers examples (4.30) through (4.36)[32] as well as (4.40). While the system developed in this study works on dependency structures, the underlying syntactic structures recognized by the system are basically the same as those identified by Vieira. With a few notable exceptions, the syntactic filters are fairly robust in recognizing non-anaphoric definite descriptions. The first exception include the definite descriptions only post-modified by a *by*-preposition. As discussed in Section 3.2.3, in general, a prepositional phrase led by *by* is incapable of giving rise to semantic uniqueness effects. In addition, the system checks the head nouns of the definite descriptions only post-modified by one prepositional phrase. If the head noun is derived from nominalization and denotes the action per se, the system attempts to locate an event antecedent for the definite description[33]. Vieira (1998) also considers cases like (4.37) under a broad category named ‘copula constructions’, which covers both definite descriptions functioning as predicate nominals and those occupying the subject position in a copula construction, as long as the complement is not a predicate adjective. In comparison, the system proposed in this study only considers definite description subjects accompanied by nominal clause predicates, which are considered instances of the type II quasi-nonanaphoric definite descriptions discussed in Section 3.2.4. The other quasi-nonanaphoric type involving copula construction[34], type I, is generally processed by the post-modification patterns since it does not cover cases without restrictive modifications. The system also recognizes cases like (4.39), which are marked as fragments (FRAG) in WSJ annotation. They are considered special cases of the type II quasi-nonanaphoric construction.

While Vieira’s (1998) ‘apposition’ heuristics covers both close and loose appositions[35], only the former is considered non-anaphoric in this study. The reason is that close appositions form a single information unit but the appositives in loose appositions are different information units (Quirk et al., 1985, section 17.70). This distinction is important as it allows the appositives in a loose apposition to be interpreted as anaphoric, either direct or associative, to discourse elements outside the apposition – just like the definite descriptions connected with a copula can acquire anaphoric readings. Consider

[32] It is not clear how Vieira (1998) handles the rare cases like (4.36), where complements or restrictive modifications are separated from the head by more than one commas. The system developed in this study counts the number of intervening commas in order to separate the nonrestrictive post-modifications from the restrictive ones.

[33] For details, see example (3.60) on page 59 and the relevant discussions thereby. The system uses the NOMLEX-plus dictionary (Meyers et al., 2004) to identify nominalized words and their types of nominalization.

[34] Type III cases are considered anaphoric in this study.

[35] Quirk et al. (1985, section 17.68) uses the terms ‘restrictive’ and ‘non-restrictive’. In written text, the appositives in the latter are usually separated by commas.

(4.44) [36]:

(4.44) Mr. Dinkins's inner circle of advisers appears to include both ideologues and pragmatists, leaving voters with little clue as to who will be more influential. *The key man* seems to be *the campaign manager*, Mr. Lynch. Vieira and Poesio (2000, ex. 52c, WSJ 765:84-85)

Both *The key man* and *the campaign manager* in (4.44) are anaphoric. The former should be read as '*The key man in Mr. Dinkins's inner circle of advisers*'; and the latter could be interpreted as associative to either *Mr. Dinkins* or direct anaphoric to a previous mention of *Mr. Lynch* a few sentences away[37].

The remaining two syntactic cues both target pre-modifications. The first one is a special subset of definite descriptions pre-modified by proper names. While the larger set shows a strong tendency of non-anaphoricity in general, it also include many cases where the proper name is only loosely related to the head noun, as in the expression '*the SEC documents*' used in reference to some documents filed with the SEC. Definite descriptions with both proper name and common noun pre-modifiers, as in *the Cray-3 price range* in (4.42), usually require the modifiers to be closely related to the head and are therefore more precise indicators of non-anaphoricity. The other grammatical cue is intended for capturing definite descriptions denoting well-known events in specific years, such as *the 1929 crash* in (4.43). In practice, it also captures other expressions like '*The 1988 trade act*', most of which are also non-anaphoric. In contrast, definite descriptions with plural head nouns and cardinal numbers as modifiers, such as '*the five new fields*', are often anaphoric. This fact is also taken into consideration when designing the classification rules.

It is also worth noting that all titles (proper names with leading definite articles) are considered non-anaphoric in this study – they are treated just like the rest of the proper names and are not considered for anaphora resolution. Vieira (1998) (as well as Vieira & Poesio, 2000) treats subsequent mentions of titles as coreferential[38]. As discussed in Chapter 3, this study makes the distinction between coreference and anaphora, and treats them as separate tasks. Aside from the marked (`NNP` and `NNPS`) proper names, the system also identifies other capitalized words, including demonyms (e.g. '*Germans*') and adjectives for countries and continents (e.g. '*Latin American*'), both of which are identified using a simple list (Demonyms, 2009) obtained from Wikipedia and are enriched with the names of the corresponding places. The identified demonyms are considered as de facto proper names[39]; adjectives for locations are considered proper names when the system applies the `NNP+NN PreMod` rule.

[36] WSJ 765:84 is added to provide the context. The example also appears in Vieira's (1998) thesis as ex. 5.15c.

[37] A similar apposition, '*Mr. Dinkins's campaign manager and former chief of staff, Bill Lynch*', is used in sentence 68 as the initial introduction of *Mr. Lynch*. As discussed in Section 3.1.4, neither mentions of *campaign manager* (in sentences 68 and 85) establishes a separate discourse referent, thus the instance in (4.44) should not be considered direct anaphoric to *Mr. Dinkins's campaign manager* in sentence 68.

[38] Vieira (1998, section 4.4.5) uses the term anaphoric to describe such relationships. The subsequent mentions of titles are not anaphoric under the notion developed in this study.

[39] A more appropriate treatment for demonyms should involve a scan to make sure they have not been previously introduced with an indefinite article or cardinals. However, referential uses of demonyms are relatively rare and are ignored in this study.

Finally, the system captures time-related expressions using the WordNet (henceforth the **Time** rule)[40]. Under the `Time` rule, a definite description is considered non-anaphoric if the first sense of its head noun is a hyponym of `time period` or if any of its senses fall under `time unit`. Examples of expressions covered by this rule include '*the year*', '*the next day*'[41], and '*the 1920s*'. Such expressions are non-anaphoric because time periods and units are universally (weakly) familiar to all parties of the conversation.

The syntactic cues identified in this section generally do not overlap, and there is no need to specify a particular order of application. The order adopted in this study is as follows: `Titles`, `Close Apposition`, `NNP+NN PreMod`, `CD+Head`$_{\text{singular}}$, `Time`, `Prep`, `RRC`, and finally `Complement`.

4.5.2 Web-based Tests for Anaphoricity

The syntactic cues described in the earlier section are generally very accurate. However, their coverage is limited. A quick review of the data presented in Table 3.3 reveals that, setting aside the titles, the grammatical rules only cover about 60%-70% of the non-anaphoric definite descriptions (other than titles). Counting titles in pushes the coverage higher to around 80%, which is still not quite satisfactory, especially considering the large percentage of definite descriptions that are not anaphoric.

Uryupina (2003) has shown that web-based statistics, or more specifically the four page count ratios (# "*the Y*")/(# *Y*), (# "*the Y*")/(# "*a Y*"), (# "*the H*")/(# *H*), and (# "*the H*")/(# "*a H*"), where *Y* and *H* denote respectively the original expression without determiner and the head noun of the expression, can be helpful in anaphoricity determination and the more general task of coreference resolution. Later studies by Poesio, Uryupina, et al. (2004) and Poesio et al. (2005) also verify the utility of such measures. As mentioned by Uryupina (2007), behavior changes in commercial search engines makes it difficult to implement the web-based statistics. However, since this study employs the Google N-grams corpus, search engine is no longer a source of problem for anaphoricity determination. The N-grams corpus also has certain other features that are not available in commercial search engines, such as case-sensitivity[42] and punctuation marks. The latter is especially helpful for demarcating the boundaries of the noun phrases being searched.

This study employs four sets of query patterns for different problems. The first set targets definite descriptions functioning as objects of prepositions:

$$\texttt{pp} \qquad \texttt{preposition} \textvisiblespace \texttt{determiner} \textvisiblespace \texttt{pre-modifiers} \textvisiblespace \texttt{head} \textvisiblespace \texttt{tail} \tag{4.48}$$

$$\overline{\texttt{prep}} \qquad \texttt{preposition} \textvisiblespace \texttt{determiner} \tag{4.49}$$

$$\overline{\texttt{np}} \qquad \textit{\texttt{the}} \textvisiblespace \texttt{pre-modifiers} \textvisiblespace \texttt{head} \textvisiblespace \texttt{tail} \tag{4.50}$$

[40] Strictly speaking, this is not a grammatical feature; it is described here for convenience's sake.

[41] See the discussion on the Adjectival class (page 53) for details on the interpretation of expressions like '*the next day*'.

[42] As discussed in Section 4.3.1, the system exploits this feature for gender/number determination.

$$\begin{aligned}\texttt{determiner} &\in \{\texttt{determiner}_{the}, \texttt{determiner}_{a}, \texttt{determiner}_{\text{otherdefdets}}\}\\ &= \{\,\mathtt{the}\,,\, \mathtt{a}\,|\,\mathtt{an}\,,\, \mathtt{this}\,|\,\mathtt{that}\,|\,\mathtt{these}\,|\,\mathtt{those}\,|\,\mathtt{his}\,|\,\mathtt{her}\,|\,\mathtt{its}\,|\,\mathtt{their}\,\}\\ \texttt{tail} &= \mathtt{period}\,|\,\mathtt{comma}\ (\,\mathtt{.}\,|\,\mathtt{,}\,)\end{aligned}$$

According to the patterns, each definite description is used to instantiate seven different queries – three queries of varying determiners for each of the patterns pp and $\overline{\texttt{prep}}$, plus the original definite description modeled by $\overline{\texttt{np}}$. The result counts obtained from these queries are denoted respectively as $n_{\text{pp},the}$, $n_{\text{pp},a}$, $n_{\text{pp,otherdefdets}}$, $\bar{n}_{\text{prep},the}$, $\bar{n}_{\text{prep},a}$, $\bar{n}_{\text{prep,otherdefdets}}$, and $\bar{n}_{\text{np},the}$. For queries instantiated with $\texttt{determiner}_{a}$, the head nouns are converted to singular forms if they are originally in plural. Due to technical constraints imposed by the N-grams corpus, the `tail` component is only added when the queries contain less than five words. In the case that the length of a query exceeds five, the system tries to remove the 'nonessential' pre-modifiers, such as adjectives (other than those denoting countries) and adverbs, from the original noun phrase[43]. If the query is still 'oversize' after the attempt, it is not executed. In addition, the system also checks the value of $\bar{n}_{\text{np},the}$ – if it falls below a predefined threshold ($N = 10^4$ in the current implementation), the 'nonessential' pre-modifiers are also removed and the queries are rebuilt and executed again.

The system further obtains three measures from the counts, namely MI_{pp}, r_{prep}, and r_{det}, as shown below:

$$\text{MI}_{\text{pp}} = \log \frac{n_{\text{pp},the}}{\bar{n}_{\text{prep},the} \times \bar{n}_{\text{np},the}} \tag{4.51}$$

$$r_{\text{prep}} = \frac{n_{\text{pp},the}}{\bar{n}_{\text{np},the}} \tag{4.52}$$

$$r_{\text{det}} = \left(\frac{n_{\text{pp},a}}{\bar{n}_{\text{prep},a}} \bigg/ \frac{n_{\text{pp},the}}{\bar{n}_{\text{prep},the}} \right) \times \left(\frac{n_{\text{pp,otherdefdets}}}{\bar{n}_{\text{prep,otherdefdets}}} \bigg/ \frac{n_{\text{pp},the}}{\bar{n}_{\text{prep},the}} \right) \tag{4.53}$$

The MI_{pp} statistic is the point-wise mutual information for the noun phrase and the preposition under the context of the definite article *the*. The statistic is used to capture fixed expressions such as '*off the hook*' and '*out of the question*', and certain 'semi-fixed' expressions such as '*around the world*'. The r_{prep} ratio is used to assess the relation between the preposition and the definite description itself. The system deems a small r_{prep} as a sign that the preposition is not essential to the interpretation of the definite description and leaves the noun phrase to later stages of processing. Finally, the r_{det} ratio indicates the acceptability of alternative determiners in the original context of the definite description. A small r_{det} means that the expression is incompatible with neither an indefinite article nor definite determiners other than *the*, which in turn hints that it is most likely non-anaphoric. Expressions captured by this heuristics are mostly (semi-)fixed expressions and generic uses, such as '*on the road*', '*in the developing world*', and '*in the future*'[44].

For definite descriptions outside preposition phrases and those attached to 'nonessential' prepo-

[43] The current implementation removes all modifiers up to the first modifier that is either a noun or a capitalized adjective.

[44] The head noun *future* is not processed by the `Time` rule discussed in the previous section.

sitions, a slightly different set of patterns are applied:

$$\texttt{np} \qquad \texttt{determiner} ␣ \texttt{pre-modifiers} ␣ \texttt{head} ␣ \texttt{tail} \tag{4.54}$$

$$\texttt{np-}\textit{of} \qquad \textit{the} ␣ \texttt{pre-modifiers} ␣ \texttt{head} ␣ \textit{of} \tag{4.55}$$

$$\begin{aligned}\texttt{determiner} &\in \{\texttt{determiner}_{\varnothing}, \texttt{determiner}_{the}, \texttt{determiner}_{a}, \texttt{determiner}_{\text{otherdefdets}}\} \\ &= \{\varnothing, \textit{the}, \textit{a}|\textit{an}, \textit{this}|\textit{that}|\textit{these}|\textit{those}|\textit{his}|\textit{her}|\textit{its}|\textit{their}\} \\ \texttt{tail} &= \textit{period}|\textit{comma}\ (\texttt{.}|\texttt{,})\end{aligned}$$

Overall, the patterns are similar to the ones used for definite descriptions serving as objects of prepositions. The main differences include the additional null determiner (i.e. the queries represent 'simple' NPs) and the np-*of* pattern, both of which are nonapplicable to definite descriptions under preposition. Five queries are instantiated for each definite description. Let $n_{\text{np},\varnothing}$, $n_{\text{np},the}$, $n_{\text{np},a}$, $n_{\text{np,otherdefdets}}$, and $n_{of,the}$ denote the respective result counts of these queries, the system further calculates four ratios, namely r_{the}, r_a, $r_{\text{otherdefdets}}$, and r_{of} according to the following formulae:

$$r_{the} = \frac{n_{\text{np},the}}{n_{\text{np},\varnothing}},\ r_a = \frac{n_{\text{np},a}}{n_{\text{np},\varnothing}},\ r_{\text{otherdefdets}} = \frac{n_{\text{np,otherdefdets}}}{n_{\text{np},\varnothing}},\ r_{of} = \frac{n_{of,the}}{n_{\text{np},the}} \tag{4.56}$$

For the two queries used to obtain r_a, the system ensures that the head nouns are converted to singular forms if they are originally in plural. The first three ratios directly assess the likelihood of the noun phrase co-occurring with the respective determiners. The last one, r_{of}, is used to determine if the noun phrase usually takes a relational reading. A large r_a, $r_{\text{otherdefdets}}$, or r_{of} is taken as the sign that the definite description under investigation is anaphoric.

As discussed in the previous section, definite descriptions with pre-modifying proper names are more likely to be non-anaphoric in general, but there are also a large number of exceptions. The same principles used to identify the other non-anaphoric cases also apply to this category. However, since these cases are uniformly restrictively pre-modified, $\texttt{determiner}_{\text{otherdefdets}}$ is no longer applicable. Therefore, yet another set of patterns is designed specifically for them:

$$\texttt{rmnp}_{the} \qquad \textit{the} ␣ \texttt{pre-modifiers} ␣ \texttt{head} \tag{4.57}$$

$$\texttt{rmnp}'_{the} \qquad \textit{the} ␣ \texttt{pre-modifiers} ␣ \texttt{head}' \tag{4.58}$$

$$\texttt{rmnp}_{a} \qquad \textit{a}|\textit{an} ␣ \texttt{pre-modifiers} ␣ \texttt{head} \tag{4.59}$$

$$\texttt{rmnp-}\textit{of} \qquad \textit{the} ␣ \texttt{pre-modifiers} ␣ \texttt{head} ␣ \textit{of} \tag{4.60}$$

The $\texttt{rmnp}_{the}$ and $\texttt{rmnp}_a$ patterns are the direct counterparts of the previously discussed np pattern, with the tail component removed due to the overall low result count of the cases in this category. Similarly, the rmnp-*of* pattern is same as np-*of*. The new $\texttt{rmnp}'_{the}$ pattern takes the same form as $\texttt{rmnp}_{the}$, but uses a different head, $\texttt{head}'$, obtained by inflecting the head noun of the original phrase to the opposite number. For example, if the original head noun is the singular '*document*', the inflected `head`$'$ is '*documents*', and vice versa. Let $n_{\text{rmnp},the}$, $n'_{\text{rmnp},the}$, $n_{\text{rmnp},a}$, and $n_{of,the}$ denote the respective result counts of these queries, the system calculates the ratios r_{indef}, r_{althead}, and r_{of}:

$$r_{\text{indef}} = \frac{n_{\text{rmnp},a}}{n_{\text{rmnp},the}},\ r_{\text{althead}} = \frac{n'_{\text{rmnp},the}}{n_{\text{rmnp},the}},\ r_{of} = \frac{n_{of,the}}{n_{\text{rmnp},the}} \tag{4.61}$$

For non-anaphoric expressions, it is expected that all three r values are sufficiently small.

Finally, the system use a combination of syntactic information, the WORDNET, and web queries to identify definite descriptions that acquire functional readings via pre-modifying adjectives such as *next*, *only*, ordinal numbers, and superlatives. As discussed in Section 3.2, these cases do not form a uniform category; neither do they behave uniformly with regard to anaphoricity. Despite the analyses offered in Sections 3.2.3 and 3.2.4, there does not seem to be a reliable practical method to determine the anaphoricity of such expressions. The system simply picks them out so that they do not interfere with the normal anaphoricity determination process. In the WSJ corpus, the superlatives are marked with the part-of-speech tags `JJS` and `RBS`. Instead of constructing a list, the ordinal numbers are obtained by querying the WORDNET for hyponyms of the sense `rank (2)`. In addition, the system also queries the N-gram corpus using the patterns ***`the`*** `adjective` and ***`a|an`*** `adjective`. If the ratio of the result counts of the two patterns, $r = n_a/n_{the}$, is less than a predefined threshold ($T_r = 0.1$ in the current implementation), the adjective is deemed as functional.

Figure 4.6 illustrates the overall algorithm for the web-based anaphoricity determination process. In the current implementation, the threshold for point-wise mutual information is set at $T_{\text{MI}} = \log 10^{-9}$, and the threshold for the various ratios is set at $T_r = 0.1$. It is worth noting that the web-based method is only designed to supplement the syntactic heuristics, and the thresholds are selected for maximum precision instead of coverage.

```
input  : a definite description $dd$,
         a threshold for point-wise mutual information $T_{\mathrm{MI}}$,
         a universal threshold for ratios $T_r$.
output: anaphoricity of $dd$.
// to be executed after syntactic heuristics are applied
if DD is pre-modified by a functional adjective then
    return Functional-Unknown;
else if DD is pre-modified by proper name then
    $n_{\mathrm{rmnp},the} \leftarrow$ ExecuteQuery(rmnp$_{the}$,$dd$);
    if $n_{\mathrm{rmnp},the} > 0$ then
        instantiate and execute patterns rmnp$_{the}$, rmnp$'_{the}$, and rmnp$_a$;
        calculate $r_{\mathrm{indef}}$, $r_{\mathrm{althead}}$, and $r_{of}$;
        if $r_{\mathrm{indef}} < T_r$ and $r_{\mathrm{althead}} < T_r$ and $r_{of} < T_r$ then
            return Non-anaphoric;
        else
            return Anaphoric;
        end
    end
    return Unknown;
else
    if DD is object of preposition then
        $\bar{n}_{\mathrm{np},the} \leftarrow$ ExecuteQuery($\overline{\mathrm{np}}$,$dd$);
        if $\bar{n}_{\mathrm{np},the} > 0$ then
            instantiate and execute patterns pp and $\overline{\mathrm{prep}}$;
            calculate $\mathrm{MI}_{\mathrm{pp}}$, $r_{\mathrm{prep}}$, and $r_{\mathrm{det}}$;
            if $\mathrm{MI}_{\mathrm{pp}} > T_{\mathrm{MI}}$ then  return Non-anaphoric;
            else if $r_{\mathrm{prep}} > T_r$ then
                if $r_{\mathrm{det}} < T_r$ then return Non-anaphoric;
                else return Anaphoric;
            end
        end
    end
    $n_{\mathrm{np},the} \leftarrow$ ExecuteQuery(np,'the',$dd$);
    if $n_{\mathrm{np},the} > 0$ then
        instantiate and execute patterns np and np-$of$;
        calculate $r_{the}$, $r_a$, $r_{\mathrm{otherdefdets}}$, and $r_{of}$;
        if $r_{the} > T_r$ and $(r_a + r_{\mathrm{otherdefdets}})/r_{the} < T_r$ and $r_{of} < T_r$ then
            return Non-anaphoric;
        end
    end
    return Unknown;
end
```

Figure 4.6: Algorithm for web-based definite description anaphoricity determination

4.6 Definite Description Anaphora Resolution

While Vieira's (1998) (also Vieira & Poesio, 2000) pioneering research provides much insight into definite description anaphora, it also leaves a few issues unaddressed. The first issue is terminology-related: Vieira essentially uses the terms anaphora and coreference interchangeably – as discussed extensively in the previous chapter, the two concepts are essentially different despite that the phenomena they denote often coincide. Another terminology issue is the artificial distinction Vieira makes between the so-called 'direct anaphora' and the rest of the directly anaphoric cases based on whether the anaphor and the antecedent have the same head noun. The latter are grouped together with the associative cases under the name 'bridging'. This practice breaks the semantically-uniform group of directly anaphoric cases into two and at the same time creates another inconsistent category. Furthermore, the proper name titles, which are non-anaphoric[45] and do not necessarily share the same head with their coreferential antecedents (e.g. acronyms and 'shortened' aliases), are put into the 'direct anaphora' category. Vieira's mainly focuses on the 'direct anaphora' cases and only include some preliminary findings on using the WORDNET to resolve bridging cases. However, the directly-anaphoric, different-head (henceforth 'indirect') cases are by no means ignorable – Vieira estimates that these cases represent about 11% of all definite descriptions in the data set used for the second annotation experiment. This figure should be compared to the weight of the 'direct anaphora' cases in the same data set, which is 33%. Omitting a quarter of the directly anaphoric cases not only limits a system's utility but also poses processing difficulties to the rest of the cases. For example, in an article describing the interaction of multiple companies, the author may choose to interleave proper names with definite descriptions sharing the same head, i.e. '*Company A ... the company ... Company B ... the company*'. If a system bases its decision solely on the head nouns of the definite descriptions, it would face many long-range resolutions as well as a high risk for resolving to wrong antecedents.

The system extends on Vieira's (1998) proposal and targets both 'direct' and 'indirect' anaphora for definite descriptions. Instead of invoking the 'indirect anaphora' processing only upon a failed attempt for same-head antecedent, the system treats the two (largely) equally in the same process. Same-head antecedents are still preferred, as they afford 'perfect' head semantic compatibility with the anaphor. However, this preference does not override the distance-based salience measure. The main difference between the treatments received by same-head and 'indirect' antecedents is that the latter are generally processed within a more limited window. This choice is motivated by both the observation from the corpus annotation that subsequent mentions using a different head usually closely follows its antecedent and the limited precision offered by currently-available approaches.

4.6.1 Segmentation

Like the approach of Vieira (1998), this study also adopts a loose, window-based segmentation and recency-based salience. As the term 'loose' indicates, certain antecedents outside the predefined window are still considered by the system. Based on Vieira's research and data gathered during annotation, a four-sentence window is used.

[45] Vieira (1998, section 3.2, footnote 8) also recognize that they are "not strictly anaphoric".

Vieira (1998) indicates that when strict segmentation is followed, in comparison to one-sentence window, using a four-sentence window only results in minimum precision loss of around 1% but offers much a higher recall of 58%, almost double the figure obtained using the one-sentence window. Increasing the window size to eight sentences leads to a further 10% gain in recall, but also an additional precision drop of around 4%. Annotated data used in this study suggest a similar pattern – about 80% of the directly anaphoric cases have their annotated antecedents within a four-sentence window, while the figures for one-sentence and eight-sentence windows are 45% and 90%, respectively.[46] Although extending the window size beyond four sentences may further increase a system's recall, it is not compatible with the observations made during annotation.

Most of the long-range anaphora cases in the corpus exhibit either (or both) of the following two patterns: (a) the discourse referent is central to the topic of the article, or (b) there are explicit or implicit indications that the author is returning to a previous discussion. Obviously, neither pattern can be properly captured by a window-based method. For example, in an article featuring stories of claims adjusters in the 1989 San Francisco earthquake (WSJ 0766), one of the subsequent mentions '*the earthquake*' has its closest antecedent as far as 35 sentences away. Vieira (1998) seeks to address the long-distance anaphora problem with two simple rules under the name 'loose segmentation heuristic': an out-of-the-window antecedent is still considered if it is a subsequent mention or identical to the anaphor (including the definite article). Although Vieira does not discuss the motivations behind the heuristic, it is shown (on the training data set) to make significant contributions to the system's recall without deteriorating its precision.

The system developed in this study implements a different set of loose segmentation heuristics[47] that directly target the aforementioned patterns. The new rules allow the system to consider antecedents beyond the fixed window if: (a) the head of the anaphor have appeared in the first three sentences of the article, or (b) one or more proper names are found in the vicinity of the anaphor. A successful match of heuristic (a) indicates that the discourse referent of the anaphor may be central to the entire discussion, and should be exempted from the window-based constraints. Heuristic (b) identifies potential anchors (i.e. the proper names) for returning to previous discussions. Since proper name aliases can be resolved with relative high confidence, the coreference chains formed by the aliases provide good anchors for different discourse segments[48].

The above-mentioned approach primarily applies to 'simple' definite descriptions, i.e. those having no additional descriptive contents, which represent the majority (about 75%) of the directly anaphoric cases in the corpus. Different treatments are given for 'complex' definite descriptions such as '*the new company*' and '*the SEC documents*'. The system tries first to identify potential an-

[46] The figures are not directly comparable to Vieira's (1998). Aside from the fact that her figures are results of a live system, the underlying data sets are also different. However, the most significant factor for the much higher percentages obtained from the annotation is most-likely caused by terminology differences, i.e. the inclusion of 'indirect' antecedents and the exclusion of subsequent proper name mentions, which are generally not subject to window-based salience.

[47] This is motivated partly by the fact that Vieira's (1998) heuristic is not accompanied by an explanation of its intended functions, and partly by the unexplained 15% drop in recall when the system was evaluated against the test data set.

[48] The utility of this heuristic is limited to news stories and other genres with non-trivial proper name uses.

chors contained in these definite descriptions. If an anchor is located, sentences containing previous mentions of the anchor are tried first. In addition, the search for antecedents of the complex definite descriptions are not restricted by segmentation windows. When the anchor-guided search fails, all sentences preceding the anaphor are visited in reverse order[49].

4.6.2 'Indirect' Anaphora Resolution

Since Vieira's (1998) original empirical work, a number of other studies have tackled on the subject of 'indirect' anaphora. For example, Vieira et al. (2006) use semantic tagging, which assigns to each noun a semantic prototype from a limited selection, to help recognize and rank potential antecedents. Garera and Yarowsky (2006) employ an unsupervised model to extract different-head coreferential pairs from large corpora based on their co-occurrence statistics. Markert and Nissim (2005) compare the performance of the WORDNET and a web-based co-occurrence statistic for both *other*-anaphora resolution and definite description anaphora resolution, and report that the web-based approach outperforms the WORDNET-based one for both tasks[50]. The performance patterns of the two knowledge sources are similar in both tasks: the web-based method generates more false-positives, which is well-compensated by its much higher coverage. As noted by Bunescu (2003), the behavior of web-based approaches can be adjusted by varying the mutual information threshold. Therefore, it is also possible to use web-based co-occurrence patterns for situations where high-precision is more desirable. Finally, Yang and Su (2007) propose methods to automatically induce and evaluate text patterns. The algorithm starts with a set of seeds consisting of known coreferential pairs and searches the corpus for co-occurrences of the pairs. The intervening words are then proposed as text patterns. The patterns mined from Wikipedia are further applied to the task of coreference resolution and is shown to improve the system's performance on proper names.

Encouraged by Markert and Nissim's (2005) results, this study relies exclusively on the web for resolving 'indirect' anaphora with nominal antecedents. Five patterns are used for this purpose, including three patterns for common noun candidates and proper names of known types – the *and other* pattern originally used by Markert and Nissim (2005), a modified version of Bunescu's (2003) $\texttt{n}_t\texttt{.}\ \textit{The}\ \texttt{n}_a\ \texttt{Verb}$ pattern, and an additional *or* pattern, plus two patterns specifically targeting proper name candidates, as shown below:

$$\texttt{rel}_{other} \qquad \texttt{head}_{1,\text{singular}}|\texttt{head}_{1,\text{plural}} \text{␣} \textit{and} \text{␣} \textit{other} \text{␣} \texttt{head}_{2,\text{plural}} \tag{4.62}$$

$$\overline{\texttt{rel}}_{other} \qquad \{\texttt{head}_{1,\text{singular}}|\texttt{head}_{1,\text{plural}} \text{␣} \textit{and} \text{␣} \textit{other}, \textit{and} \text{␣} \textit{other} \text{␣} \texttt{head}_{2,\text{plural}}\} \tag{4.63}$$

$$\texttt{rel}_{or} \qquad \texttt{head}_{1,\text{plural}} \text{␣} \textit{or} \text{␣} \texttt{head}_{2,\text{plural}} \tag{4.64}$$

$$\overline{\texttt{rel}}_{or} \qquad \{\texttt{head}_{1,\text{plural}} \text{␣} \textit{or}, \textit{or} \text{␣} \texttt{head}_{2,\text{plural}}\} \tag{4.65}$$

$$\texttt{evoke} \qquad \texttt{head}_{\text{candidate}} \text{␣} \textit{,} \text{␣} \textit{the} \text{␣} \texttt{head}_{\text{anaphor}} \tag{4.66}$$

[49] The news articles in the corpus are usually not very long. It may be necessary to impose a hard limit on the search for proses of other genres.

[50] As noted by Versley (2007), the same may not hold across different languages.

$$\texttt{rel}_{\text{app}} \qquad \texttt{head}_{\text{anaphor}} \sqcup \texttt{proper name} \sqcup \texttt{tail} \tag{4.67}$$

$$\overline{\texttt{rel}}_{\text{app}} \qquad \{\texttt{head}_{\text{anaphor}}, \texttt{proper name} \sqcup \texttt{tail}\} \tag{4.68}$$

$$\texttt{rel}_{\text{pred}} \qquad \texttt{proper name} \sqcup \texttt{be} \sqcup \texttt{det} \sqcup \texttt{head}_{\text{anaphor}} \tag{4.69}$$

$$\overline{\texttt{rel}}_{\text{pred}} \qquad \{\texttt{proper name} \sqcup \texttt{be} \sqcup \texttt{det}, \texttt{be} \sqcup \texttt{det} \sqcup \texttt{head}_{\text{anaphor}}\} \tag{4.70}$$

$$\texttt{tail} = \textit{\texttt{period}}|\textit{\texttt{comma}}\ (\texttt{.}\,|\,\texttt{,}),\ \texttt{be} = \textit{\texttt{is}}\,|\textit{\texttt{was}},\ \texttt{det} = \textit{\texttt{a}}\,|\textit{\texttt{an}}$$

As shown by Markert and Nissim (2005), the $\texttt{rel}_{other}$ pattern works reasonably well for hyponym/synonym relationships. The additional $\texttt{rel}_{or}$ is included with the intention to strengthen the system's ability to identify coreferential pairs composed of sister terms, such as *home* and *house*. Both patterns are undirected. Given a pair of anaphor and potential antecedent, two queries are generated from each pattern – one obtained by substituting $\texttt{head}_1$ with the head noun of the anaphor and $\texttt{head}_2$ with that of the candidate, and the other one obtained by alternating the terms being replaced. For example, instantiating $\texttt{rel}_{or}$ with the candidate-anaphor pair $\langle$*home*, *house*$\rangle$ yields both "homes or houses" and "houses or homes". Similarly, two different sets of normalization queries, {"homes or", "or houses"} and {"houses or", "or homes"}, are also generated from the accompanying normalizing pattern $\overline{\texttt{rel}}_{or}$. Let $n_{or,\text{can,ana}}$, $\{\bar{n}_{1,or,\text{can,ana}}, \bar{n}_{2,or,\text{can,ana}}\}$, $n_{or,\text{ana,can}}$, $\{\bar{n}_{1,or,\text{ana,can}}, \bar{n}_{2,or,\text{ana,can}}\}$ denote respectively the result counts obtained from these queries, the system calculates two mutual information measures $\text{MI}_{\text{can,ana}}$ and $\text{MI}_{\text{ana,can}}$ and uses the larger of the two as the final result, MI_{or}.

$$\text{MI}_{or} = \max\left(\log\frac{n_{or,\text{can,ana}}}{\bar{n}_{1,or,\text{can,ana}}\cdot\bar{n}_{2,or,\text{can,ana}}}, \log\frac{n_{or,\text{ana,can}}}{\bar{n}_{1,or,\text{ana,can}}\cdot\bar{n}_{2,or,\text{ana,can}}}\right) \tag{4.71}$$

MI_{other} is obtained through a similar process. For the directional patterns $\texttt{rel}_{\text{app}}$ and $\texttt{rel}_{\text{pred}}$, the MI measures are simply calculated as $\log(n/\bar{n}_1/\bar{n}_2)$. It is worth noting that the evoke pattern does not have accompanying normalization patterns. Since the system does not impose a particular order for the instantiation of the $\texttt{rel}_{other}$ pattern[51] and also uses the more generic $\texttt{rel}_{or}$ pattern, the evoke pattern is used as a minimum safety measure to ensure that pairs like $\langle$*catastrophe*, *earthquake*$\rangle$ are not recognized as coreferential. If a query generated by evoke returns a non-zero count[52], the underlying pair is considered admissible. Figure 4.7 illustrates the algorithm used to determine the semantic compatibility between an anaphor and a common noun antecedent candidate. In the current implementation, the thresholds for MI_{other} and MI_{or} are set at $T_{other} = T_{or} = \log 10^{-7}$.

While much effort has been devoted to assigning types to named entities (cf. Section 4.3), many proper names remain unrecognized prior to the anaphora resolution process. In a final attempt, the system uses the $\texttt{rel}_{\text{app}}$ and $\texttt{rel}_{\text{pred}}$ patterns to identify potential anaphoric relationships to the 'un-typed' proper names. The former pattern is intended to capture close appositions involving both the anaphor's head noun and the name, as in '*composer Bach*'. The tail is an important component that helps filter out unwanted instances such as '*the woman Bach loved*'. The latter

[51] Markert and Nissim (2005) only allow the candidate to instantiate $\texttt{head}_1$ if it is a common noun.

[52] Note that the Google N-grams corpus has a cut-off count of 40, which can be seen as an implicit constraint.

```
function: Head-Compatible
input    : a pair of head nouns $n_c$ (candidate), $n_a$ (anaphor)
           a distance measure $dist$ expressed in number of sentences
output   : $c \in [-\infty, 0]$ denoting the semantic compatibility of the heads
$SAMEHEAD \leftarrow 0$;
$INCOMPATIBLE \leftarrow -\infty$;
if $n_c = n_a$ then
    return $SAMEHEAD$;                              // same-head antecedent
else if Lemmatize($n_c$) = Lemmatize($n_a$) then
    return $INCOMPATIBLE$;                          // same-head but number mismatch
else if $dist <= 1$ then
    $n_{\text{evoke}} \leftarrow$ ExecuteQuery(evoke, $n_c$, $n_a$);
    if $n_{\text{evoke}} > 0$ then
        instantiate and execute patterns $\texttt{rel}_{other}$ and $\overline{\texttt{rel}}_{other}$;
        calculate $\text{MI}_{other}$;
        if $\text{MI}_{other} > T_{other}$ then
            return $\text{MI}_{other}$;
        else
            instantiate and execute patterns $\texttt{rel}_{or}$ and $\overline{\texttt{rel}}_{or}$;
            calculate $\text{MI}_{or}$;
            if $\text{MI}_{or} > T_{or}$ then
                return $\text{MI}_{or}$;
            end
        end
    end
end
return $INCOMPATIBLE$;
```

Figure 4.7: Algorithm for common noun antecedent candidate semantic filtering

pattern captures explicit copular constructions such as '*Bach was a composer*'. Compared with the apposition pattern, queries instantiated from the copula pattern generally return less results and does not seem to provide better coverage. However, Versley's (2007) study shows that the German version of the pattern is one of the better-performing patterns among those evaluated, which indicates it is probably unwise to dismiss the utility of the pattern based on the limited experience of this study alone. The algorithm for assessing the compatibility between the anaphor and a proper name antecedent candidate is illustrated in Figure 4.8. The `GetTypes` function (line 3) not only returns the identified type of the proper name but also includes educated guesses (e.g. '*Amfac Hotel*' → *hotel*) and other nouns that are used predicatively (e.g. predication, apposition, or *as*-preposition) against any instance of the name (including aliases). In the current implementation, the thresholds for MI_{pred} and MI_{app} are set at $T_{\text{pred}} = \log 10^{-7}$ and $T_{\text{app}} = \log 10^{-10}$, respectively.

When a nominal antecedent cannot be located, the system also checks for potential event antecedents if the head of the anaphor is derived from nominalization. The process begins by searching the NOMLEX-plus dictionary (Meyers et al., 2004) for the original verb form of the anaphor[53].

[53] The system only considers entries marked as 'verb-nom', i.e. where the nominalized form refers directly to the state or action of the verb.

```
function: NE-Compatible
input    : a proper name $ne$
           the head noun of the anaphor $n_a$
output   : $c \in [-\infty, 0]$ denoting the semantic compatibility of the heads
$SAMEHEAD \leftarrow 0$;
$INCOMPATIBLE \leftarrow -\infty$;
$Types \leftarrow$ GetTypes($ne$);
foreach $type \in Types$ do
    if $type = n_a$ then
        return $SAMEHEAD$;                          // same-head antecedent
    else if Lemmatize($type$) $=$ Lemmatize($n_a$) then
        return $INCOMPATIBLE$;             // same-head but number mismatch
    else
        instantiate and execute patterns $\texttt{rel}_{other}$ and $\overline{\texttt{rel}}_{other}$;
        calculate $\mathrm{MI}_{other}$;
        if $\mathrm{MI}_{other} > T_{other}$ then
            return $\mathrm{MI}_{other}$;
        end
    end
end
instantiate and execute patterns $\texttt{rel}_{\text{pred}}$ and $\overline{\texttt{rel}}_{\text{pred}}$;
calculate $\mathrm{MI}_{\text{pred}}$;
if $\mathrm{MI}_{\text{pred}} > T_{\text{pred}}$ then
    return $\mathrm{MI}_{\text{pred}}$;
else
    instantiate and execute patterns $\texttt{rel}_{\text{app}}$ and $\overline{\texttt{rel}}_{\text{app}}$;
    calculate $\mathrm{MI}_{\text{app}}$;
    if $\mathrm{MI}_{\text{app}} > T_{\text{app}}$ then
        return $\mathrm{MI}_{\text{app}}$;
    end
end
return $INCOMPATIBLE$;
```

Figure 4.8: Algorithm for named entity antecedent candidate semantic filtering

If the search is successful, the system identifies candidate verbs in the vicinity and matches the lemmatized candidates against the verb form(s) of the anaphor. When direct string matching fails, the system further obtains the nominalizations of the candidates, pairs them with the anaphor, and instantiates the $\texttt{rel}_{or}$ pattern with the pairs. An additional undirected pattern, $\texttt{vrel}_{or}$, is also used to expand the system's coverage:

$$\texttt{vrel}_{or} \qquad \textit{to} \;\text{␣}\; \texttt{lemmatized verb}_1 \;\text{␣}\; \textit{or} \;\text{␣}\; \texttt{lemmatized verb}_2 \tag{4.72}$$

The pattern is similar to the $\texttt{rel}_{or}$ pattern but works with lemmatized verbs. However, it does not have any accompanying normalization pattern – a query instantiated with this pattern is considered successful if it returns a non-zero count.

4.6.3 The Algorithm

Given an article, the overall procedure for definite description anaphora resolution is as follows:

1. Identify all definite descriptions and put them in a list.
2. Determine the anaphoricity of each definite description as detailed in Section 4.5, and remove the non-anaphoric ones.
3. Iterate through the list, identify the simple definite descriptions and try to resolve them.
4. Resolve the rest of the definite descriptions.
5. Iterate through the unresolved definite descriptions, classify those pre-modified by proper names or restrictively post-modified as non-anaphoric, and the rest as associative. The system does not attempt to identify the antecedents of the associative definite descriptions.

Figure 4.9 illustrates the algorithm for simple definite description resolution. The function `SemanticCompatibility` (line 5) is the combination of the `Head-Compatible` function (Figure 4.7) and the `NE-Compatible` function (Figure 4.8), plus the procedure for identifying event antecedents outlined in Section 4.6.2. As mentioned in Section 4.6.1, the system identifies the concepts that are mentioned in the first three sentences of the article. Prior to anaphora resolution, all noun phrases in these sentences are scanned and the head nouns of common noun phrases and types of proper names (if one has been identified) are added to the *KeyConcepts* list. When the system fails to resolve a simple definite description within the segmentation window, it checks if the head of the anaphor can be found in the list (line 10). Upon successful match, the system relaxes the segmentation constraint and searches the entire preceding text for potential antecedents. The block beginning at (line 19) implements the other loose segmentation heuristic discussed in Section 4.6.1. It identifies the nearby proper names and uses the (preceding) coreferential mentions of the proper names as a starting points for new searches. Finally, when all previous methods fail, the system attempts to identify an event antecedent. If multiple candidates are identified, the system ranks them by their semantic compatibility scores and recency (in that order) expressed in the number of intervening sentences. Since same-head antecedents are assigned a high compatibility score of 0, they are always preferred over other candidates. In the rare cases when a tie is observed, i.e. when there are multiple highest-ranked candidates in the same sentence, the system gives preference to the candidates that are definite or restrictively modified, and as a last resort, the one closest to the anaphor.

```
input  : a simple definite description $dd$
output: antecedent of $dd$, or $\varnothing$ if none can be identified.

// to be executed after anaphoricity determination
1  $INCOMPATIBLE \leftarrow -\infty$;
2  $Candidates \leftarrow \emptyset$;
   // get preceding NPs in a four-sentence window in inverse order
3  $NPs \leftarrow$ GetNPs($dd$,4);
4  foreach $np \in NPs$ do
5      $c \leftarrow$SemanticCompatibility($np, dd$);
6      if $c > INCOMPATIBLE$ then
7          add $\langle np, c \rangle$ to $Candidates$;
8      end
9  end
10 if $Candidates = \emptyset$ and Head($dd$) $\in KeyConcepts$ then
11     $NPs \leftarrow$ GetNPs($dd$);                          // get all preceding NPs
12     foreach $np \in NPs$ do
13         $head \leftarrow np$ is proper name ? Type($np$) : Head($np$);
14         if $head =$ Head($dd$) then
15             add $\langle np, c \rangle$ to $Candidates$;
16         end
17     end
18 end
19 if $Candidates = \emptyset$ then
       // identify proper names in a one-sentence window
       // pronouns realizing proper names are also considered
20     $Anchors \leftarrow$ GetProperNames($dd$,1);
21     foreach $anchor \in Anchors$ do
22         $AnchorPoints \leftarrow$ GetCoreferentProperNames($anchor$);
23         foreach $propername \in AnchorPoints$ do
               // get NPs preceding $dd$ in a four-sentence window
               // beginning with the sentence containing $propername$
24             $NPs \leftarrow$ GetNPs(SentenceId($propername$),$dd$,4);
25             repeat lines 4 to 9;
26         end
27     end
28 end
29 if $Candidates = \emptyset$ and $np$ is derived from nominalization then
       // get preceding verbs in a four-sentence window in inverse order
30     $Verbs \leftarrow$ GetVerbs($dd$,4);
31     repeat lines 4 to 9 with $Verbs$ in place of $NPs$;
32 end
33 if $Candidates = \emptyset$ then
34     return $\varnothing$;
35 else
36     sort $Candidates$ by score;
37     return $Candidates[0]$;
38 end
```

Figure 4.9: Algorithm for definite description anaphora resolution (simple DDs)

The algorithm for complex definite description resolution is presented in Figure 4.10. Except for the additional requirement for modifier matching, the algorithm shares the same basic components with the one used for simple definite descriptions, although the order of execution is different. Given a complex definite description, the algorithm first identifies the named entities and other definite descriptions that either directly modify it or are embedded in its modifiers. These previous mentions of these entities are then used as anchor points to initiate searches. Unlike the similar process used for simple definite descriptions, the searches are confined to the same sentence where the anchors appear. If the above method fails, the system searches the entire preceding text and gathers all 'head-compatible' entities as candidates. For each identified candidate, the system collects the pre-modifiers from the entire coreference chain and matches them against the pre-modifiers of the antecedent using the `ModifierCompatibility` function (line 26). The currently implemented method for modifier matching is based on string comparison and does not involve any semantic checking. The modifiers are divided into several groups – functional adjectives, normal adjectives, cardinal numbers, nouns, and proper names[54] – and matching is performed for each group. A complete match in each group receives a reward of 2, and a partial match receives a reward of 1. Zero matches are generally not punished, except when the anaphor is modified by a functional adjective, or when the anaphor and the candidate both have proper name modifiers but the two sets have no intersection. In either of these cases the candidate is eliminated.

4.7 Evaluation

Three performance measures are used throughout the section: precision, recall, and the balanced F-measure (Rijsbergen, 1979). Precision is defined as the ratio of correctly classified instances in a specific category (or a collection of categories) to the number of instances identified by the system as belonging to the category (categories). In other words, precision is calculated as $P = \frac{TP}{TP+FP}$, where TP and FP are the number of true positives and false positives respectively. Recall is defined as the ratio of correctly classified instances in a specific category (or a collection of categories) to the total number of instances in the category (categories), or $R = \frac{TP}{TP+FN}$, where FN denotes the number of false negatives. Finally, the F-measure is the weighted harmonic mean of precision and recall used to indicate a system's overall performance. When precision and recall are weighted equally, as used in this study, the balanced F-measure is defined as $F = \frac{2PR}{P+R}$. When it is necessary to examine the contribution of a specific factor, approximate randomization test (Noreen, 1989) is used to determine the statistical significance of the differences in performance. The significance level $\alpha = 0.05$ and number of shuffles $R = 9999$, both chosen arbitrarily, are used where significance tests are performed.

[54] Adjectives for locations are also treated as proper names for this purpose.

4.7.1 Pronominal Anaphora Resolution

The performance of the pronominal anaphora resolution system is summarized in Table 4.6. The data set includes 22 non-anaphoric instances of pronouns – 5 generic/deictic first-person instances, 16 instances of pleonastic *it*, and 1 idiomatic *it*. As the system only employs a simple quoted-speech identifier, many anaphoric instances of first person pronouns are missed and misclassified as non-anaphoric. On the other hand, the pleonastic *it* detector (cf. Chapter 5) is highly-effective: it only missed one instance of pleonastic cases and introduced one false-positive[55] (Precision=Recall= 94%). Because each pleonastic instance of *it* can be further assigned an antecedent if it is not correctly identified, removing the anaphoricity detector from the system immediately results in 14 andditional incorrect resolutions. This lowers the system's accuracy from 91% to 87%, which is statistically significant ($p < 0.001$). Only removing the various web-based selectional restrictions (in which case the system falls back to syntax-guided recency) also leads to a decrease of performance – with the addition of 8 incorrectly resolved cases, the system's overall accuracy drops to 89%. The difference in performance is not statistically significant ($p = 0.09$). However, further removing the components for determining *as*-predication and the antecedents for subject NP-* elements, both of which also heavily rely on the web, introduce another three incorrectly resolved cases, causing the system's performance to drop to 88% and the difference in system performance becomes statistically significant ($p = 0.03$).

Category	**Items**		**Correct**	
First-person		28		23 (82%)
Anaphoric	23		17	
Generic	3		3	
Deictic	2		2	
Third-person: *it*		108		93 (86%)
Anaphoric	91		78	
Extraposition[a]	14		14	
Cleft[a]	1		1	
Local situation[a]	1		0	
Idiom[a]	1		0	
Third-person: others		175		168 (96%)
Singular	116		115	
Plural	59		53	
Anaphoric		289		264 (91%)
Non-anaphoric		22		20 (91%)
Total		**311**		**284 (91%)**

Table 4.6: Performance of pronominal anaphora resolution

[a] See Chapter 5 for detailed discussion.

[55] The false-positive pleonastic case can be correctly resolved by the system had it not been marked as pleonastic. The case is a missing-object construction. As discussed in Section 5.3.3, we have identified potential methods to address these cases but they were not implemented due to the sparseness of data.

Since the system does not have a backup method for acquiring gender and number information, it is not possible to remove the web-based gender/number determination module. However, as shown by Bergsma (2005b), the web-based approach to gender/number information acquirement provides both higher accuracy and better coverage than corpus-based approaches do. Interestingly, despite being one of the best available methods, the web-based gender/number determination module still caused one case to be resolved incorrectly and is also responsible for another case as it failed to provide any guidance. Both cases are listed below:

(4.45) A disaffected, hard-drinking, nearly-30 hero sets off for snow country in search of an elusive sheep with a star on *its* back at the behest of a sinister, erudite mobster with a Stanford degree. WSJ 37:12

(4.46) The 40-year-old Mr. Murakami is a publishing sensation in Japan. A more recent novel, "Norwegian Wood"(every Japanese under 40 seems to be fluent in Beatles lyrics), has sold more than four million copies since Kodansha published it in 1987. But *he* is just one of several youthful writers – Tokyo's brat pack – who are dominating the best-seller charts in Japan. WSJ 37:15-17

In (4.46), none of the queries generate enough hit count for *Kodansha*, thus leaving it available to all pronouns. Since the real antecedent, *Mr. Murakami*, is located in a sentence further away, the system has no choice but to settle on *Kodansha*. In (4.45), counts gathered from the web strongly suggest a plural reading for *sheep*, which eliminates it from the candidates for the pronoun *it*. The case of (4.45) is particularly interesting, not only because *sheep* belongs to a small but important group of nouns for which gender/number information cannot be confidently determined from external sources, but also because it serves as a reminder of the difficulty in reliably integrating information from different sources – while we could add a new rule to the gender/number determination module to specify that nouns with the indefinite articles *a* or *an* are more likely to be interpreted as singular, it would be difficult to decide how much weight should be given to such a rule. On one hand, the web-based approach for gender/number determination is already quite accurate, making the utility of such a rule questionable. On the other hand, it is also possible to use expressions such as '*a people and their past*', which indicates that the rule cannot be used to override information gained from other sources.

The *sheep* issue is strongly reminiscent of one of the key problems we faced while designing the pronominal anaphora system: there does not seem to be a suitable place for selectional restrictions. Although there are many cases that clearly suggest their role, they are generally less helpful than gender/number information and syntactic cues. The main issue here seems to be that selectional patterns are too simplistic to capture the level of reasoning required to resolve an anaphor (Kehler et al., 2004 are certainly right in this respect). More specifically, in many situations, for a human reader, all that is needed is that two concepts can be used together[56], instead of whether they are more likely to occur together. For example, consider (4.47), a case in which the selectional restrictions preferred *Fans* over the correct antecedent, *players*:

(4.47) But other than the fact that besuboru is played with a ball and a bat, it's unrecognizable: Fans politely return foul balls to stadium ushers; the strike zone expands depending on the size of the hitter; ties are permitted – even welcomed – since they honorably sidestep the shame of defeat; players must abide by strict rules of conduct even in *their* personal lives – players for the Tokyo Giants, for example, must always wear ties when on the road. WSJ 37:24

Most human readers would agree that in this sentence, the question of which group is more likely to have a life is essentially irrelevant.

Although we have not been able to find a pattern capable of reliably predicting the situations under which selectional restrictions become absolutely necessary, the inferior results caused by the removal of web-based selectional restrictions indicate that we might have identified a reasonably good niche for their application – in the gaps where syntactic cues usually do not offer strong guidance. Perhaps the most obvious example of such gaps is the situations where a candidate is modified by prepositions or genitive constructions, as illustrated by example (4.25), repeated below:

(4.25) Once inside, she spends nearly four hours measuring and diagramming each room in the 80-year-old house, gathering enough information to estimate what it would cost to rebuild *it*.

Since the entities in such relationships are usually closely related, it seems more reasonable to assign them he same syntactic salience (as opposed to the original approach of Hobbs, 1978). In addition, such entities generally refer to different discourse referents and are highly likely to have different semantic types. Both these characteristics make the constructions ideal candidates for the application of selectional restrictions.

Finally, we would like to present (4.48), a case that was correctly resolved 'by accident':

(4.48) 53. Many of the adjusters employed by Aetna and other insurers have some experience with construction work or carpentry.

54. But such skills were alien to Toni Johnson.

55. Four years ago, she was managing a film-processing shop and was totally bored.

56. A friend mentioned that *she* might want to look into a position at Aetna, if she was interested in a job that would constantly challenge her.

57. She signed up, starting as an "inside" adjuster, who settles minor claims and does a lot of work by phone. WSJ 766:53-57

[56] In certain situations, as in reading a fiction, we would even accept things that are not generally possible, just out of cooperation. Blindly applying selectional restrictions in such situations would certainly lead to unwanted effects.

Had the gender/number module provided correct information about *friend*, which should be made available to both feminine and masculine pronouns, the system would have resolved the pronoun *she* in sentence 56 (and the entire chain of pronouns follows) to *A friend*. This case closely resembles example (4.28) on page 99, for which we identified centering as the solution. However, the case of (4.48) is more complex in that one must recognize that the indefinite description *A friend* is actually a form of realization of the previously mentioned Toni Johnson. In other words, *A friend* is equivalent to '*one of Toni Johnson's friends*', or simply '*one of her friends*', using the pronoun *she* in sentence 55. Once this is done, the case would be easily resolvable because the algorithm would find a pronoun among the most salient entities and assigns it as the antecedent. However, there is currently no reliable way of determining whether a non-definite description should be interpreted as anaphoric, not to mention assigning the correct antecedent to it. Examples like (4.28) and (4.48) suggest that further research along the lines of how theories of discourse structures can be more effectively integrated into the process of anaphora resolution are likely to yield fruitful results.

4.7.2 Definite Description Anaphoricity Determination and Resolution

Table 4.7 outlines the system's performance on definite description anaphoricity determination. As shown in Table 3.3, about 30% of the annotated non-anaphoric definite descriptions are titles (proper names definite articles). Titles are of less interest to this study, since they are essentially the same as the 'article-less' counterparts and are relatively easy to identify. Compared to the other tasks involved in this study, they are also relatively easy to resolve. With the assistance of gender/number information, the word-based editing distance approach introduced in Section 4.3.2 correctly resolved almost all named entity aliases appeared in annotated portion of the corpus, with the only exception of '*Fannie Mae*', used as an alias to '*the U.S. Federal National Mortgage Association*'. Titles are included in Table 4.7 as a separate item for the sake of completeness. Unless otherwise noted, the ensuing discussion only refers to the 211 definite descriptions that cannot be identified through part-of-speech tags or string comparison.

As shown in Table 4.7, the majority of the non-anaphoric definite descriptions are identified through syntactic means or the WORDNET. With the insight gained from Chapter 3, we were able reexamine the rules offered by Poesio and Vieira (1998), discard some of the rules while fine-tune a few others. These efforts increased the performance of the syntactic cues considerably. For example, the relevant discussions in Chapter 3 motivated the decision to not consider functional adjectives (e.g. *next*, *same* etc.) per se as indicators of non-anaphoricity. This decision alone avoided 9 false-positives – among the 14 definite descriptions[57] in the corpus that contain functional adjectives but without further restrictive modification, only 5 are actually non-anaphoric.

The addition of web-based anaphoricity detection significantly increased the coverage of the system. Judging from the figures reported in Table 4.7, it is the second most powerful component

[57] Those with head nouns denoting time-related concepts are not counted.

Rules	Identified	Correct	Incorrect	
			Direct	Associative
Syntactic Cues				
R. PostMod/Complement	115	111	1	3
Quasi-Nonanaphoric	2	2		
CD+Single Head	6	6		
NNP+NN Premod	8	8		
Time	12	12		
Web	42	41		1
Sub Total (P=97%, R=85%[a], F=91%[a])	**185**	**180**		
Latent[b]	5	4	1	
Titles[c]	92	92		
Total (P=98%, R=91%, F=94%)	**282**	**276**	**2**	**4**
Total Annotated Non-anaphoric Instances		**303**		

Table 4.7: Performance of anaphoricity determination

[a] Calculated based on the number of non-anaphoric definite descriptions exclusive of titles (i.e. $303 - 92 = 211$ DDs).

[b] Items identified as non-anaphoric after anaphora resolution.

[c] Including both proper name titles and demonyms.

in identifying non-anaphoric cases, only after the syntactic cues of restrictive post-modification and complement clauses. Removing the web-based features lowers the system's recall to 66% and F-measure to 79%; the deterioration of both measures are statistically significant ($p < 0.001$).

Since the design of the anaphoricity detector put more emphasis on precision and is generally conservative in classifying entities as non-anaphoric, it is still possible to gain higher coverage by examining the definite descriptions for a second time after the process of anaphora resolution. The current implementation classifies the antecedent-less definite descriptions that are either post-modified by prepositional phrases or pre-modified by proper names as non-anaphoric. Including these cases further increase recall to 87% and the F-measure to 92%, but the difference is not statistically significant ($p = 0.15$).

Finally, the overall performance of the definite description anaphora resolution system is presented in Table 4.8. The total number of cases for each category are listed in the column headed by (#); the number of correctly identified or resolved cases are shown in the column (+); the number of incorrect results are listed in the column (-), followed by the performance measures precision, recall, and balanced F-measure. The system correctly resolved 132 directly anaphoric definite noun phrases, 87 (66%) of which share the same head noun with the antecedent. The percentage of same-head pairs is slightly lower (106 out of 172, or 62%) if we also consider the incorrectly resolved cases. As one would expect, same-head pairs are recognized with higher precision (82%) than the other directly anaphoric cases (68%). There are 5 cases in which the system correctly identifies the coreferential antecedent of a non-anaphoric entity (that is missed by the anaphoricity detector).

Task	#	+	-	P	R	F
Direct Anaphora (classification)	157	147	25	85%	94%	89%
Direct Anaphora (resolution)	157	132	40	77%	84%	80%
Associative Anaphora (classification)	51	30	27	53%	59%	56%
Anaphoricity Detection	211	184	6	97%	85%	91%

Table 4.8: Summary of system performance on definite description processing

These cases are marked as incorrect resolutions in Table 4.8. In over 50% of the correctly resolved different-head cases the antecedent is a proper name or title, but many would not have been correctly resolved without the assistance of the web. A particularly interesting case is '*Hong Kong . . . the colony*', which is unlikely to be (and should not be) encoded in any reasonably up-to-date ontology database.

Although the study uses the same data set as Poesio and Vieira's (1998) test corpus[58], the differences in terminologies adopted in the two studies make it difficult to perform a direct and quantitative comparison. Vieira's 'direct anaphora' category covers subsequent mentions of titles (proper names), but does not cover the directly anaphoric cases where the anaphor and the antecedent do not share the same head noun. The latter cases are instead merged with the associative cases under the name 'bridging'. In addition, the notion of 'discourse-new' adopted in Vieira's study excludes the non-anaphoric cases that also have coreferential antecedents. As mentioned earlier in Chapter 1, the inadequacy of these practices are part of the motivations for this study.

Having said that, some qualitative observations can still be made. While a system only processing same-head mentions can achieve relatively good precision (Vieira reports precision of 83% on the test data and 88% for the training data), it faces more severe problems with irregular long-distance anaphor-antecedent relationships. For example, Vieira shows that increasing the window size to 8 increases the system's recall from 62% to 67%. However, our own annotation indicates that approximately 80% of the definite descriptions have their antecedents within a 4-sentence window when different-head antecedents are also considered. To this end, it may be actually advantageous to consider both same-head and different-head antecedents.

[58] As mentioned earlier, we cannot identify the source of discrepancy in the total number of definite descriptions as reported by Vieira (1998) and our study.

```
   input  : a complex definite description $dd$
   output: antecedent of $dd$, or $\varnothing$ if none can be identified.
   // to be executed after anaphoricity determination
1  $INCOMPATIBLE \leftarrow -\infty$;
2  $Candidates \leftarrow \emptyset$;
   // identify pre-modifying proper names and
   // definite descriptions/proper names embedded in post-modification
3  $Anchors \leftarrow$ GetDefiniteModifiers($dd$);
4  foreach $anchor \in Anchors$ do
5      $AnchorPoints \leftarrow$ GetCoreferentNPs($anchor$);
6      foreach $ap \in AnchorPoints$ do
           // get NPs in the same sentence as $ap$
7          $NPs \leftarrow$ GetNPs(SentenceId($ap$),$dd$,0);
8          foreach $np \in NPs$ do
9              $c \leftarrow$SemanticCompatibility($np, dd$);
10             if $c > INCOMPATIBLE$ then
11                 add $\langle np, c \rangle$ to $Candidates$;
12             end
13         end
14     end
15 end
16 if $Candidates = \emptyset$ and $np$ is derived from nominalization then
17     repeat lines 4 to 15 for verbs;
18 end
19 if $Candidates = \emptyset$ then
       // get the pre-modifying components of the anaphor
20     $AMods \leftarrow$ GetModifyingComponents($dd$);
       // get all preceding NPs in inverse order
21     $NPs \leftarrow$ GetNPs($dd$);
22     repeat lines 8 to 13;
23     foreach $candidate \in Candidates$ do
24         $CandidateCorefChain \leftarrow$ GetCoreferentNPs($candidate$);
           // get the pre-modifying components from all entities on
           // the coreference chain of the candidate
25         $CMods \leftarrow$ GetModifyingComponents($CandidateCorefChain$);
           // assess the compatibility of the two sets of modifiers
26         $m \leftarrow$ModifierCompatibility($CMods, AMods$);
27         add $m$ to $candidate$'s score;
28     end
29 end
30 if $Candidates = \emptyset$ then
31     return $\varnothing$;
32 else
33     sort $Candidates$ by score;
34     return $Candidates[0]$;
35 end
```

Figure 4.10: Algorithm for definite description anaphora resolution (complex DDs)

Chapter 5

Web-assisted Pleonastic Pronoun Identification[1]

As noted in Chapter 1, not all pronoun uses are anaphoric. Typical non-anaphoric pronoun uses are illustrated by the following examples:

(5.1) But *it* doesn't take much <u>to get burned</u>†. WSJ 34:20

(5.2) And most disturbing, *it* is <u>educators, not students</u>†, who are blamed for much of the wrongdoing. WSJ 44:26

Pronouns used without an antecedent, often referred to as being pleonastic or structural, pose a serious problem for anaphora resolution systems. Many anaphora resolution systems underestimate the issue and choose not to implement a specific module to handle pleonastic pronouns but instead have their input 'sanitized' manually to exclude such cases. However, the high frequency of pronoun usage in general and pleonastic cases in particular warrants that the phenomenon deserves more serious treatment. The pronoun *it*, which accounts for most of the pleonastic pronoun usages, is by far the most frequently used of all pronouns in the British National Corpus (BNC). In the WSJ corpus, upon which this study is based, *it* accounts for more than 30% of personal pronoun usage. The percentage of cases where *it* lacks a nominal antecedent is also significant: previous studies have reported figures between 16% and 50% (J. Gundel, Hedberg, & Zacharski, 2005) while our own analysis based upon the WSJ corpus results in a value around 25%, more than half of which are pleonastic cases.

In this chapter, a novel approach is proposed to identify the pleonastic uses of *it*. The chapter is organized as follows: first, Section 5.1 offers an overview of the various uses of the pronoun *it*; Section 5.2 briefly surveys related work toward both classification of *it* and identification of pleonastic *it*; Section 5.3 proposes a web-based approach for identification of pleonastic *it*; Section 5.4 demonstrates the proposed method with a case study; Section 5.5 follows with evaluation; and finally, Section 5.7 discusses the findings and presents ideas for further improvements.

[1] A version of this chapter has been published. Yifan Li, Petr Musilek, Marek Reformat, Loren Wyard-Scott 2009. J. Artif. Intell. Res. (JAIR) 34: 339-389.

5.1 Uses of *It*

Applying criteria similar to those established by J. Gundel et al. (2005), the usage of *it* can be generally categorized as follows.

1. Referential with nominal antecedent

   (5.3) The thrift holding company said *it* expects to obtain regulatory approval and complete the transaction by year-end. WSJ 6:2

   where *it* refers to the thrift holding company.

2. Referential with clause antecedent

   (5.4) He was on the board of an insurance company with financial problems, but he insists he made no secret of *it*. WSJ 41:29

   where *it* refers to the fact that the person was on the board of an insurance company.

   (5.5) Everyone agrees that most of the nation's old bridges need to be repaired or replaced. But there's disagreement over how to *do it*. WSJ 102:2-3

   where *it*, together with *do*, refers to the action of repairing or replacing the bridge.

3. No antecedent – Pleonastic

   (a) Extraposition

   (5.6) But *it* doesn't take much to get burned†. WSJ 34:20

   where the infinitive clause to get burned is extraposed and its original position filled with an expletive *it*. The equivalent non-extraposed sentence is '*But to get burned doesn't take much.*'

   (5.7) *It*'s a shame their meeting never took place†. WSJ 37:34

   The equivalent non-extraposed sentence is '*That their meeting never took place is a shame.*'

   (b) Cleft[2]

   (5.8) And most disturbing, *it* is educators, not students†, who are blamed for much of the wrongdoing. WSJ 44:26

   The equivalent non-cleft version is '*And most disturbing, educators, not students, are blamed for much of the wrongdoing.*'

   (5.9) *It* is partly for this reason† that the exchange last week began trading in its own stock "basket" product ... WSJ 591:21

   The equivalent non-cleft version is '*The exchange last week began trading in its own stock basket product partly for this reason.*'

[2] Some claim that cleft pronouns should not be classified as expletive (J. K. Gundel, 1977; Hedberg, 2000). Nevertheless, this does not change the fact that the pronouns do not have nominal antecedents; hence clefts are included in this analysis.

(c) Local Situation

(5.10) *It* was not an unpleasant evening … WSJ 207:37

This category consists of *it* instances related to weather, time, distance, and other information about the local situation. Since the texts reviewed in this study lack instances of other subtypes, only weather and time cases are discussed.

4. Idiomatic

(5.11) The governor couldn't make *it*, so the lieutenant governor welcomed the special guests. WSJ 10:10

This chapter focuses on pleonastic cases (the third category), where each subclass carries its unique syntactic and/or semantic signatures. The idiomatic category, while consisting of non-anaphoric cases as well, is less coherent and its identification is much more subjective in nature, making it a less attractive target.

5.2 Previous Work

As Evans (2001) pointed out, usage of *it* is covered in most serious surveys of English grammar, some of which (e.g. Sinclair, 1995) also provide classifications based on semantic categories. In a recent study, J. Gundel et al. (2005) classify third-person personal pronouns into the following comprehensive hierarchy:

- Noun phrase (NP) antecedent
- Inferrable
- Non-NP antecedent
  - Fact
  - Proposition
  - Activity
  - Event
  - Situation
  - Reason
- Pleonastic
  - Full extraposition
  - Full cleft
  - Truncated cleft
  - Truncated extraposition
  - Atmospheric
  - Other pleonastic
- Idiom
- Exophoric
- Indeterminate

Without going into the details of each category, it is apparent from the length of the list that the phenomenon of pleonastic *it*, and more generally pronouns without explicit nominal antecedents, have been painstakingly studied by linguists. However, despite being identified as one of the open issues of anaphora resolution (Mitkov, 2001), work on automatic identification of pleonastic *it* is relatively scarce. To date, existing studies in the area fall into one of two categories: one wherein a rule-based approach is used, and the other using a machine-learning approach.

5.2.1 Rule-based Approaches

Paice and Husk (1987) together with Lappin and Leass (1994) provide examples of rule-based systems that make use of predefined syntactic patterns and word lists. The Paice and Husk approach employs bracketing patterns such as *it ... to* and *it ... who* to meet the syntactic restrictions of extraposition and cleft. The matched portions of sentences are then evaluated by further rules represented by word lists. For example, the *it ... to* rule prescribes that one of the 'task status' words, such as *good* or *bad*, must be present amid the construct. In order to reduce false positives, general restrictions are applied on sentence features such as construct length and intervening punctuation.

Lappin and Leass' (1994) approach employs a set of more detailed rules such as *It is* Cogv-ed *that* S and *It is* Modaladj *that* S, where Cogv and Modaladj are predefined lists of cognitive verbs (e.g. *think* and *believe*) and modal adjectives (e.g. *good* and *useful*), respectively. Compared to Paice and Husk's (1987) approach, this method is much more restrictive, especially in its rigidly-specified grammatical constraints. For example, it is not clear from the original Lappin and Leass paper whether the system would be able to recognize sentences such as (5.12), despite its claim that the system takes syntactic variants into consideration.

(5.12) *It* isn't clear, however, whether support for the proposal will be broad enough to pose a serious challenge to the White House's acid-rain plan†. WSJ 146:14

Lappin and Leass' (1994) approach is part of a larger system, and no evaluation is provided. The Paice and Husk (1987) approach, on the other hand, evaluates impressively. It has an accuracy of 93.9% in determining pleonastic constructs on the same data used for rule development, without using part-of-speech tagging or parsing.

Both rule-based systems rely on patterns to represent syntactic constraints and word lists to represent semantic constraints. This makes them relatively easy to implement and maintain. However, these features also make them less scalable – when challenged with large and unfamiliar corpora, their accuracies deteriorate. For example, Paice and Husk (1987) noticed nearly a 10% decrease in accuracy when rules developed using one subset of the corpus are applied to another subset without modifications. Boyd, Gegg-Harrison, and Byron (2005) also observed a significant performance penalty when the approach was applied to a different corpus. In other words, rule-based systems can only be as good as they are designed to be. Denber (1998) suggested using WORDNET (Fellbaum, 1998) to extend the word lists, but it is doubtful how helpful this would be considering the enormous number of possible words that are not included in existing lists and the number of inapplicable words that will be identified by such an approach.

5.2.2 Machine-learning Approaches

Recent years have seen a shift toward machine-learning approaches, which shed new light on the issue. Studies by Evans (2001, 2000) and Boyd et al. (2005) are examples of this class. Both systems employ memory-based learning on grammatical feature vectors; Boyd et al.'s approach also includes a decision tree algorithm that produces less ideal results. In his attempt to place uses of *it* into seven categories, including pleonastic and nominal anaphoric among others, Evans uses 35 features to encode information such as position/proximity, lemmas, and part-of-speech, related to both the pronoun and other components of interest, such as words and noun phrases, in the sentence. Evans reported 73.38% precision and 69.25% recall for binary classification of pleonastic cases, and an overall binary classification accuracy of 71.48%. In a later study featuring MARS, a fully automatic pronoun resolution system that employs the same approach, Mitkov et al. (2002) reported a significantly higher binary classification accuracy of 85.54% when the approach is applied to technical manuals.

Boyd et al.'s (2005) approach targets pleonastic *it* alone. It uses 25 features, most of which concern lengths of specific syntactic structures; also included are part-of-speech information and lemmas of verbs. The study reports an overall precision of 82% and recall of 71%, and, more specifically, recalls on extrapositional and cleft constructs of 81% and 45%, respectively.

In addition, Clemente, Torisawa, and Satou (2004) used support vector machines with a feature-set similar to that proposed by Evans (2001) to analyze biological and medical texts, and reported an overall accuracy of 92.7% – higher than that of their own memory-based learning implementation. Ng and Cardie (2002a) built a decision tree for binary anaphoricity classification on all types of noun phrases (including pronouns) using the C4.5 induction algorithm. Ng and Cardie reported overall accuracies of 86.1% and 84.0% on the MUC-6 and MUC-7 data sets. Categorical results, however, are not reported and it is not possible to determine the system's performance on pronouns. Using automatically induced rules, Müller (2006) reported an overall accuracy of 79.6% when detecting non-referential *it* in spoken dialogs. An inter-annotator agreement study conducted in the same paper indicates that it is difficult even for humans to classify instances of *it* in spoken dialogs. This finding is supported by our own experiences.

Machine-learning approaches are able to partly circumvent the restrictions imposed by fixed word lists or rigid grammatical patterns through learning. However, their advantage also comes with a price – training is required in the initial development phase and for different corpora re-training is preferable since lemmas are part of the feature sets. Since the existing approaches fall within the area of supervised learning (i.e. training data need to be manually classified), the limited number of lemmas they gather from training may lead to degraded performance in unfamiliar circumstances. Moreover, the features used during learning are unable to reliably capture the subtleties of the original sentences, especially when considering non-technical documents. For example, the quantitative features frequently used in machine-learning approaches, such as position and distance, become less reliable when sentences contain a large number of adjuncts. Additionally, the meanings of lemmas are often domain-dependent and can vary with their local structural and lexical environment – such nuances cannot be captured by the lemma features alone. In short, while machine-learning approaches generally deliver better performance classifying *it* than their rule-based counterparts do, they have their own inherent problems.

5.3 A Web Based Approach

Both syntactic patterns and semantics of various clause constituents play important roles in determining if a third-person personal pronoun is pleonastic. The role of grammar is quite obvious since both extrapositions and clefts must follow the grammatical patterns by which they are defined. For example, the most commonly seen type of *it*-extraposition follows the pattern:

> *it* + copula + status + subordinate clause
> *It* ␣ is ␣ easy ␣ to see why the ancient art is on the ropes†. WSJ 89:17

In contrast, the role semantics plays here is a little obscure until one sits down and starts to "dream up exceptions" (Paice & Husk, 1987) analogous to (5.13) and (5.14), where referential and pleonastic cases share the same syntactic structure.

(5.13) … *it* has taken measures to continue shipments during the work stoppage. WSJ 74:5

(5.14) … *it* didn't take a rocket scientist to change a road bike into a mountain bike† … WSJ 367:44

Despite its less overt role, failure to process semantic information can result in a severe degradation of performance. This observation is supported by the word-list-based systems' dramatic decay in accuracy when they are confronted with text other than that they obtained their word lists from.

Like every other classification system, the system developed in this study strives to cover as many cases as possible and at the same time perform classification as accurately as possible. To achieve this, it attempts to make good use of both syntactic and semantic information embedded in sentences. A set of relaxed yet highly relevant syntactic patterns is first applied to the input text to filter out the syntactically inviable cases. Unlike the matching routines of some previous approaches, this process avoids detailed specification of syntactic patterns. Instead, it tries to include every piece of text containing a construct of possible interest. Different levels of semantic examinations are performed for each subtype of pleonastic constructs. For reasons discussed later in Section 5.3.2, semantic analysis is not performed on clefts. A WordNet-based analysis is used to identify weather/time cases because among the samples examined during the system's development stage, cases pertaining to this class are relatively uniform in their manner of expression. For the most complex and populous class, the extrapositions, candidates are subjected to a series of tests performed as queries against the web. Results of the queries provide direct evidence of how a specific configuration of clause constituents is generally used.

The reason that such a corpus-based approach is chosen versus applying manually constructed knowledge sources, such as a word list or WordNet, is fourfold:

1. Manually constructed knowledge sources, regardless of how comprehensive they are, contain only a small portion of general world knowledge. In the particular settings of this study, general world knowledge is used for making judgements such as which words are allowed to serve as the matrix verb of an extraposition, and even more subtle, which specific sense of a word is permitted.

2. Manually compiled knowledge sources are subject to specific manners of organization that may not satisfy the system's needs. Taking WORDNET as an example, it identifies a large number of various relationships among entities, but the information is mainly organized along the axes of synonyms, hypernyms (kind-of relationship), and holonyms (part-of relationship) etc., while it is the surroundings of a particular word that are of more interest to this study.

3. Natural languages are evolving quickly. Taking English as an example, each year new words are incorporated into the language[3] and the rules of grammar have not been immune to changes either. Using a large and frequently-updated corpus such as the web allows the system to automatically adapt to changes in language.

4. Most importantly, corpora collect empirical evidence of language usage. When the sample size is large enough, as in the case of the web, statistics on how a specific construct is generally used in corpora can be employed as an indicator of its speaker's intention.

The approach proposed in this study is also inspired by Hearst's (1992) work on mining semantic relationships using text patterns, and many other quests that followed in the same direction (Berland & Charniak, 1999; Poesio et al., 2002; Markert, Nissim, & Modjeska, 2003; Cimiano, Schmidt-Thieme, Pivk, & Staab, 2005). Unlike these investigations that focus on the semantic relationship among noun phrases, the pleonastic pronoun identification problem mandates more complex queries to be built according to the original sentences. However, the binary nature of the problem also makes it simpler to apply comparative analysis on results of multiple queries, which, in turn, leads to better immunity to noise.

Figure 5.1 illustrates the general work flow of the system. A sentence is first preprocessed to obtain a dependency tree with part-of-speech tags, which is then passed on to the syntactic filtering component to determine whether minimum grammatical requirements of the pleonastic constructs are met. It is also during the syntactic filtering process that clefts and weather/time expressions are identified using syntactic cues and the WORDNET respectively. The candidate extrapositions are thereafter used to instantiate various queries on search engines; the results returned from the queries serve as parameters for the final decision-making mechanism.

[3] Metcalf and Barnhart (1999) have compiled a chronicle of many important additions to the vocabulary of American English.

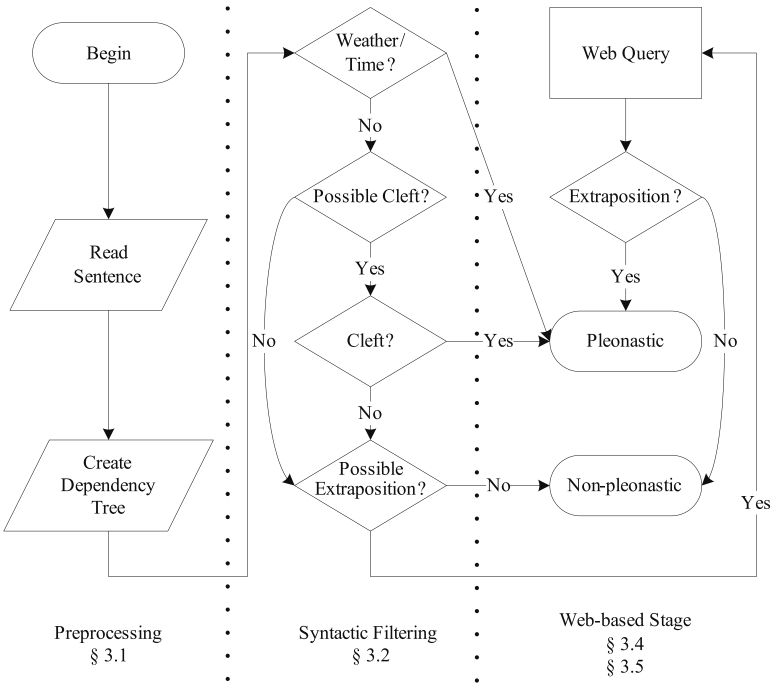

Figure 5.1: Illustration of the system work flow broken into three processing stages – preprocessing, syntactic filtering, and web-based analysis.

5.3.1 Preprocessing

The pleonastic *it* identification system uses the same preprocessing unit introduced in Section 4.2. The unit converts the tagged and parsed sentences in the WSJ corpus to dependency structures using a set of head percolation rules. Figure 5.2 illustrates the syntactic structure of example of *it*-extraposition (5.7) before and after head percolation. Head entities are underlined in the CFG diagram and circled in the DG diagram.

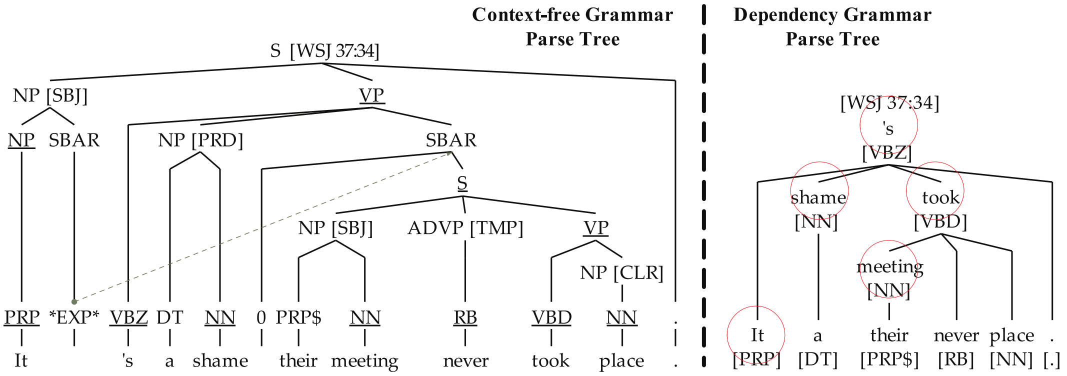

Figure 5.2: Illustration of a sentence's syntactic structure, both as annotated in the WSJ corpus (left) and after head percolation (right).

5.3.2 Syntactic Filtering

The syntactic filtering process determines whether a clause meets the grammatical requirements of an extraposition or cleft construct by matching the clause against their respective syntactic patterns.

Extrapositions

It-extrapositions occur when a clause is dislocated out of its ordinary position and replaced with *it*. An *it*-extraposition usually follows the pattern:

$$\overbrace{\mathtt{it}_{subject} + \underbrace{\left\{ \begin{array}{l} \mathtt{be} + \left\{ \begin{array}{l} \texttt{noun phrase} \\ \texttt{adjective phrase} \\ \texttt{prepositional phrase} \end{array} \right\} \\ \\ \texttt{verb phrase} \end{array} \right\}}_{\texttt{matrix verb phrase}}}^{\texttt{matrix clause}} + \texttt{extraposed clause} \tag{5.1}$$

This pattern summarizes the general characteristics of subject *it*-extrapositions, where the pronoun *it* assumes the subject position. When the matrix verb (the verb following *it*) is the main copula *to be*, which serves to equate or associate the subject and an ensuing logical predicate, it must be followed by either a noun phrase, an adjective phrase, or a prepositional phrase.[4] There is no special requirement for the matrix verb phrase otherwise. Similarly, there is almost no restriction placed upon the extraposed clause except that a full clause should either be introduced without a complementizer (e.g. 5.7) or led by *that*, *whether*, *if*, or one of the *wh*-adverbs (e.g. *how*, *why*, *when*, etc.). These constraints are developed by generalizing a small portion of the WSJ corpus and are largely in accordance with the patterns identified by Kaltenböck (2005). Compared to the patterns proposed by Paice and Husk (1987), which also cover cases such as *it ... to*, *it ... that* and *it ... whether*, they allow for a broader range of candidates by considering sentences that are not explicitly marked (such as 5.7). The above configuration covers sentences such as:

(5.15) Since the cost of transporting gas is so important to producers' ability to sell it, *it* helps to have input and access to transportation companies†. WSJ 529:9

(5.7) *It*'s a shame their meeting never took place†.

(5.16) *It* is insulting and demeaning to say that scientists "needed new crises to generate new grants and contracts† ...[5] WSJ 360:36

(5.17) *It* won't be clear for months whether the price increase will stick†. WSJ 336:19

[4] Other copula verbs do not receive the same treatment. This arrangement is made to accommodate cases where verbs such as *to seem* and *to appear* are immediately followed by an extraposed clause.

[5] Neither *insulting* nor *demeaning* is in Paice and Husk's (1987) list of 'task status words' and therefore cannot activate the *it ... to* pattern.

Except in the case of the last sentence, the above constructs are generally overlooked by the previous rule-based approaches identified in Section 5.2.1. As the last sample sentence illustrates, the plus sign (+) in the pattern serves to indicate a forthcoming component rather than suggest two immediately adjacent components.

Some common grammatical variants of the pattern are also recognized by the system, including questions (both direct and indirect), inverted sentences, and parenthetical expressions (Paice & Husk, 1987). This further expands the pattern's coverage to sentences such as:

(5.18) I remembered how hard *it* was for an outsider to become accepted† … WSJ 772:6

(5.19) "The sooner our vans hit the road each morning, the easier *it* is for us to fulfill that obligation†." WSJ 562:15

(5.20) Americans *it* seems have followed Malcolm Forbes's hot-air lead and taken to ballooning in a heady way†. WSJ 239:9

Aside from being the subject of the matrix clause, extrapositional *it* can also appear in the object position. The system described here captures three flavors of object extraposition. The first type consists of instances of *it* followed by an object complement:

(5.21) Mrs. Yeargin was fired and prosecuted under an unusual South Carolina law that makes *it* a crime to breach test security†. WSJ 44:14

In this case the system inserts a virtual copula *to be* between the object *it* and the object complement (*a crime*), making the construct applicable to the pattern of subject extraposition. For example, the part of (5.21) marked by wavy underline translates into '*it* is a crime to breach test security†'.

The other two kinds of object extraposition are relatively rare:

- Object of verb (without object complement)

  (5.22) Speculation had *it* that the company was asking $100 million for an operation said to be losing about $20 million a year† … WSJ 114:7

- Object of preposition

  (5.23) They should see to *it* that their kids don't play truant† … WSJ 1286:54

These cases cannot be analyzed within the framework of subject extraposition and thus must be approached with a different pattern:

$$\texttt{verb + [preposition]}\ \textit{it}_{object}\ \texttt{+ full clause} \tag{5.2}$$

The current system requires that the full clauses start with a complementizer *that*. This restriction, however, is included only to simplify implementation. Although in object expositions it is more common to have clauses led by *that*, full clauses without a leading complementizer are also acceptable.

According to Kaltenböck's (2005) analysis there are special cases in which noun phrases appear as an extraposed component, such as:

(5.24) *It*'s amazing the number of theologians that sided with Hitler.
Kaltenböck (2005, ex. S1A-053-201)

He noted that these noun phrases are semantically close to subordinate interrogative clauses and can therefore be considered a marginal case of extraposition. However, no such cases were found in the corpus during the annotation process and they are consequently excluded from this study.

Cleft

It-clefts are governed by a slightly more restricted grammatical pattern. Following Hedberg (1990), *it*-clefts can be expressed as follows:

$$\mathtt{it}_{subject} + \mathtt{copula} + \mathtt{clefted\ constituent} + \mathtt{cleft\ clause} \tag{5.3}$$

The cleft clause must be finite (i.e. a full clause or a relative clause); and the clefted constituents are restricted to either noun phrases, clauses, or prepositional phrases.[6] Examples of sentences meeting these constraints include:

(5.25) "*It*'s the total relationship† that is important." WSJ 296:29

(5.26) *It* was also in law school† that Mr. O'Kicki and his first wife had the first of seven daughters. WSJ 267:30

(5.27) "If the market goes down, I figure *it*'s paper profits† I'm losing." WSJ 121:48

In addition, another non-canonical and probably even marginal case is also identified as a cleft:

(5.28) I really do not understand how† *it* is that Filipinos feel so passionately involved in this father figure that they want to dispose of and yet they need. WSJ 296:37

Text following the structure of this sample, where a *wh*-adverb immediately precedes *it*, is captured using the same syntactic pattern by appending a virtual prepositional phrase to the matrix copula (e.g. '*for this reason*'), as if the missing information has already been given.

Each of the examples above represents a possible syntactic construct of *it*-clefts. While it is difficult to tell the second and the third cases apart from their respective extrapositional counterparts, it is even more difficult to differentiate the first case from an ordinary copula sentence with a restrictive relative clause (RRC). For example, the following sentence,

[6] Adjective and adverb phrases are also possible but they are relatively less frequent and are excluded from this analysis.

(5.29) "*It*'s precisely the kind of product that's created the municipal landfill monster," the editors wrote. WSJ 62:12

and its slightly modified version,

(5.29′) "*It*'s <u>this kind of product</u>† that's created the municipal landfill monster," the editors wrote.

are similar in construction. However, the latter is considered a cleft construct while the first is an RRC construct. To make things worse, and as pointed out by many (e.g. Boyd et al., 2005, ex. 5), sometimes it is impossible to make such a distinction without resorting to the context of the sentence.

Fortunately, in the majority of cases the syntactic features, especially those of the clefted constituent, provide some useful cues. In an *it*-cleft construct, the cleft clause does not constitute a head-modifier relationship with the clefted constituent, but instead forms an existential and exhaustive presupposition[7] (Davidse, 2000; Hedberg, 2000; Lambrecht, 2001). For example, '*I figure it's paper profits I'm losing.*' implies that in the context there is something (and only one thing) that the speaker is going to lose, and further associates *paper profits* with it. This significant difference in semantics often leaves visible traces on the syntactic layer, some of which, such as the applicability of proper nouns as clefted constituents, are obvious. Others are less obvious. The system utilizes the following grammatical cues when deciding if a construct is an *it*-cleft[8]:

- For the clefted constituent:
  - Proper nouns[9] or pronouns, which cannot be further modified by an RRC;
  - Common nouns without determiner, which generally refer to kinds[10];
  - Plurals, which violate number agreement;
  - Noun phrases that are grounded with demonstratives or possessives, or that are modified by RRCs, which unambiguously identify instances, making it unnecessary in most cases to employ an RRC;
  - Noun phrases grounded with the definite determiner *the*, and modified by an *of*-preposition whose object is also a noun phrase grounded with *the* or is in plural. These constructs are usually sufficient for introducing uniquely identifiable entities (through association), thus precluding the need for additional RRC modifiers. The words *kind*, *sort*, and their likes are considered exceptions of this rule;

[7] This applies to canonical clefts, which do not include the class represented by (5.26).

[8] A construct is considered an *it*-cleft if any of the conditions are met.

[9] There are exceptional cases where proper names are used with additional determiners and RRC modifiers, such as in '*the John who was on TV last night*', cf. Sloat's (1969) account.

[10] The validity of this assertion is under debate (Krifka, 2003). Nevertheless, considering the particular syntactic setting in discussion, it is highly unlikely that bare noun phrases are used to denote specific instances.

- Adverbial constructs that usually do not appear as complements. For example, phrases denoting location (*here*, *there* etc.) or a specific time (*today*, *yesterday* etc.), or a clause led by *when*; and
- Full clauses, gerunds, and infinitives.

- For the subordinate clause:
    - Some constructs appear awkward to be used as an RRC. For example, one would generally avoid using sentences such as (5.30), as there are better alternatives.

      (5.30) * *it* is a place that is dirty

      In the current implementation two patterns are considered inappropriate for RRCs, especially in the syntactic settings described in Equation 5.3: A) the subordinate verb phrase consists of only a copula verb and an adjective; and B) the subordinate verb phrase consists of no element other than the verb itself.
- Combined:
    - When the clefted constituent is a prepositional phrase and the subordinate clause is a full clause, such as in the case of example (5.26), the construct is classified as a cleft[11].

Some of these rules are based on heuristics and may have exceptions, making them less ideal guidelines. Moreover, as mentioned earlier, there are cleft cases that cannot be told apart from RRCs by any grammatical means. However, experiments show that these rules are relatively accurate and provide appropriate coverage, at least for the WSJ corpus.

Additional Filters

Aside from the patterns described in earlier sections, a few additional filters are installed to eliminate some semantically unfit constructs and therefore reducing the number of trips to search engines. The filtering rules are as follows:

- For a clause to be identified as a subordinate clause and subsequently processed for extraposition or cleft, the number of commas, dashes and colons between the clause and *it* should be either zero or more than one, a rule adopted from Paice and Husk's (1987) proposal.
- Except the copula *to be*, sentences with matrix verbs appearing in their perfect tense are not considered for either extraposition or cleft.
- When *it* is the subject of multiple verb phrases, the sentence is not considered for either extraposition or cleft.
- Sentences having a noun phrase matrix logical predicate together with a subordinate relative clause are not considered for extraposition.

[11] In case it is not a cleft, chances are that it is an extraposition. This assumption, therefore, does not affect the overall binary classification.

- Sentences having both a matrix verb preceded by modal auxiliaries *could* or *would* and a subordinate clause led by *if* or a *wh*-adverb are not considered for extraposition. For example, (A.16) is not considered for extraposition.

  (5.31) … *it* could complete the purchase by next summer if its bid is the one approved by the bankruptcy court. WSJ 13:17

Except for the first, these rules are optional and can be deactivated in case they introduce false-negatives.

5.3.3 Design of Search Engine Queries

As discussed in previous sections, *it*-extrapositions cannot be reliably identified using syntactic signatures alone or in combination with synthetic knowledge bases. To overcome the artificial limitations imposed by knowledge sources, the system resorts to the web for the necessary semantic information.

The system employs three sets of query patterns: the *what*-cleft, the comparative expletive test, and the missing-object construction. Each set provides a unique perspective of the sentence in question. The *what*-cleft pattern is designed to find out if the sentence under investigation has a valid *what*-cleft counterpart. Since *it*-extrapositions and *what*-clefts are syntactically compatible (as shown in Section 5.3.3) and valid readings can usually be obtained by transformations from one construct to the other, the validity of the *what*-cleft is indicative of whether or not the original sentence is extrapositional. The comparative expletive test patterns are more straightforward – they directly check whether the instance of *it* can be replaced by other entities that cannot be used expletively in the same context as that of an extrapositional *it*. If the alternate construct is invalid, the original sentence can be determined as expletive. The third set of patterns are supplemental. They are intended only for identifying the relatively rare phenomenon of missing-object construction, which may not be reliably handled by the previous pattern sets.

Designing the appropriate query patterns is the most important step in efforts to exploit large corpora as knowledge sources. For complex queries against the web, it is especially important to suppress unwanted uses of certain components, which could result from different word senses, different sentence configuration, or a speaker's imperfect command of the language. For example, the query "*it is a shame that*" could return both a valid extrapositional construct and an RRC such as '*It is a shame that is perpetuated in his life*'; and the query "*what is right is that*" could return both valid *what*-clefts and sentences such as '*Why we ought to do what is right is that* … ' This study employs three different approaches to curb unwanted results:

- The first and most important measure is comparative analysis – pairs of similarly-constructed queries are sent out to the search engine and the ratios of result counts are used for the decision. This method is effective for problems caused by both different sentence configuration and bad language usage, since generally neither contribute a fraction of results large enough to significantly affect the ratio. The method also provides a normalized view of the web because what is of interest to this study is not exactly how frequently a specific construct is used, but whether it is more likely to carry a specific semantic meaning when it is used.

- The second measure is to use stubs in query patterns, as detailed in the following sections. Stubs help ensure that the outcomes of queries are syntactically and semantically similar to the original sentences and partly resolve the problems caused by word sense difference.
- Finally, when it is infeasible to use comparative analysis, part of the query results are validated to obtain an estimated number of valid results.

Query Pattern I: The *What*-cleft

The first query pattern,

What + verb phrase + copula + stub (5.4)

is a *what*-(pseudo-)cleft construct that encompasses matrix-level information found in an *it*-extraposition. The pattern is obtained using a three-step transformation as illustrated below:

it + verb phrase + clause
It ␣ is easy ␣ <u>to see why the ancient art is on the ropes</u>†. WSJ 89:17
1) ⇓
clause + verb phrase
To see why the ancient art is on the ropes ␣ is easy.
2) ⇓ (5.5)
What + verb phrase + copula + clause
What ␣ is easy ␣ is ␣ to see why the ancient art is on the ropes.
3) ⇓
What + verb phrase + copula + stub
What ␣ is easy ␣ is ␣ to

Step 1 transforms the original sentence (or clause) to the corresponding non-extraposition form by removing the pronoun *it* and restoring the information to the canonical subject-verb-complement order. In the above example, the clause *to see* ... is considered the real subject and is moved back to its canonical position. The non-extraposition form is subsequently converted during step 2 to a *what*-cleft that highlights its verb phrase. Finally, in step 3, the subordinate clause is reduced into a stub to enhance the pattern's coverage. The choice of stub depends on the structure of the original subordinate clause: *to* is used when the original subordinate clause is an infinitive, a gerund, or a *for* ... infinitive construct[12]. For the rest of the cases, the original complementizer, or *that*, in the case where there is no complementizer, is used as stub. The use of a stub in the pattern imposes a syntactic constraint, in addition to the ones prescribed by the pronoun *what* and the copula *is*, that demands a subordinate clause be present in query results. The choice of stubs also reflects, to

[12] According to Hamawand (2003), the *for* ... infinitive construct carries distinct semantics; reducing it to the infinitive alone changes its function. However, with only a few exceptional cases, we find this reduction generally acceptable. i.e. The lost semantics does not affect the judgment of expletiveness.

a certain degree, the semantics of the original texts and therefore can be seen as a weak semantic constraint.

Below are two additional examples of the *what*-cleft transformation:

- *It* remains unclear whether the bond issue will be rolled over†. WSJ 59:14

⇒ *What* remains unclear is whether

- *It*'s a shame their meeting never took place†. WSJ 37:34

⇒ *What* is a shame is that

The *what*-cleft pattern only identifies whether the matrix verb phrase is capable of functioning as a constituent in an *it*-extraposition. Information in the subordinate clauses is discarded because this construct is used relatively infrequently and adding extra restrictions to the query will prohibit it from yielding results in many cases.

Some *it*-extraposition constructs such as '*it appears that* ... ' and '*it is said that* ... ' do not have a valid non-extraposition counterpart, but the *what*-cleft versions often bear certain degrees of validity and queries instantiated from the pattern will often yield results (albeit not many) from reputable sources. It is also worth noting that although the input and output constructs of the transformation are syntactically compatible, they are not necessarily equivalent in terms of givenness (whether and how information in one sentence has been entailed by previous discourse). Kaltenböck (2005) noted that the percentage of extrapositional *it* constructs carrying new information varies greatly depending on the category of the text. In contrast, a *what*-cleft generally expresses new information in the subordinate clause. The presupposed contents in the two constructs are different, too. *What*-clefts, according to J. K. Gundel (1977), from which the *it*-clefts are derived, have the same existential and exhaustive presuppositions carried by their *it*-cleft counterparts. On the other hand, the *it*-extrapositions, which are semantically identical to their corresponding non-extrapositions, lack such presuppositions or, at most, imply them at a weaker strength (Geurts & Sandt, 2004). These discrepancies hint that a derived *what*-cleft is a 'stronger' expression than the original extraposition, which may have been why queries instantiated from the pattern tend to yield considerably less results.

Another potential problem with this pattern is its omission of the subordinate verb, which occasionally leads to false positives. For example, it does not differentiate between '*it helps to have input and access to transportation companies*' and '*it helps expand our horizon*'. This deficiency is accommodated by additional query patterns.

Query Pattern II: Comparative Expletiveness Test

The second group of patterns provides a simplified account of the original text in a few different flavors. After execution, the results from individual queries are compared to assess the expletiveness of the subject pronoun. This set of patterns takes the following general form:

$$\texttt{pronoun + verb phrase + simplified extraposed clause} \qquad (5.6)$$

The only difference among individual patterns lies in the choice of the matrix clause subject pronoun: *it*, *which*, *who*, *this*, and *he*. When the patterns are instantiated and submitted to a search engine, the number of hits obtained from the *it* version should by far outnumber that of the other versions combined if the original text is an *it*-extraposition; otherwise the number of hits should be at least comparable. This behavior reflects the expletive nature of the pronoun in an *it*-extraposition, which renders the sentence invalid when *it* is replaced with other pronouns that have no pleonastic use.

A simplified extraposed clause can take a few different forms depending on its original structure:

Original Structure	**Simplified**
infinitive (*to meet you*)	infinitive + stub
for … infinitive[13](*for him to see the document*)	infinitive + stub
gerund (*meeting you*)	gerund + stub
full clause led by complementizer (*it is a shame that their meeting never took place*)	complementizer + stub
full clause without complementizer (*it is a shame their meeting never took place*)	*that* + stub

Table 5.1: Simplification of extraposed clause

Similar to the case of Pattern I, the stub is used both as a syntactic constraint and a semantic cue. Depending on the type of search engine, the stub can be either *the*, which is the most widely used determiner, or a combination of various determiners, personal pronouns and possessive pronouns, all of which indicate a subsequent noun phrase. In the case that an infinitive construct involves a subordinate clause led by a *wh*-adverb or *that*, the complementizer is used as stub. This arrangement guarantees that the results returned from the query conform to the original text syntactically and semantically. A null value should be used for stubs in an object position if the original text lacks a nominal object. To illustrate the rules of transformation, consider the following sentence:

(5.32) "My teacher said *it* was OK for me to use the notes on the test†," he said. WSJ 44:10

The relevant part of the sentence is:

`it` + `verb phrase` + `clause`
it ␣ was OK ␣ for me to use the notes on the test†

Applying the clause simplification rules, the first query is obtained:

`it` + `verb phrase` + `simplified clause`
it ␣ was OK ␣ to use the

The second query is generated by simply replacing the pronoun *it* with an alternative pronoun:

`alternative pronoun` + `verb phrase` + `simplified clause`
he ␣ was OK ␣ to use the

[13] The *for* … passive-infinitive is transformed into active voice (e.g. '*for products to be sold*'→'*to sell products*').

Google reports 94,200 hits for the *it* query, while only one page is found using the alternative query. Since the pronoun *it* can be used in a much broader context, replacing *it* with *he* alone hardly makes a balanced comparison. Instead, the combination of *which*, *who*, *this*, and *he* is used, as illustrated in the following examples:

- "My teacher said *it* was OK for me to use the notes on the test†," he said. WSJ 44:10

$\Rightarrow \left\{\begin{array}{l} \textit{it} \\ \textit{which/who/this/he} \end{array}\right\}$ was ok to use the

- *It* is easy to see why the ancient art is on the ropes†. WSJ 89:17

$\Rightarrow \left\{\begin{array}{l} \textit{it} \\ \textit{which/who/this/he} \end{array}\right\}$ is easy to see why

A special set of patterns is used for object extrapositions[14] to accommodate their unique syntactic construct:

$$\texttt{verb + [preposition] pronoun + } \textit{that} \texttt{ + stub} \tag{5.7}$$

Stubs are chosen according to the same rules for the main pattern set, however only one alternative pronoun – *them* – is used.

- Speculation had *it* that the company was asking \$100 million for an operation said to be losing about \$20 million a year† ... WSJ 114:7

$\Rightarrow$ had $\left\{\begin{array}{c} \textit{it} \\ \textit{them} \end{array}\right\}$ *that* the

Query Pattern III: Missing-object Construction

One search engine annoyance is that they ignore punctuation marks. This means one can only search for text that matches a specific pattern string, but not sentences that end with a pattern string. The stubs used in Pattern II are generally helpful for excluding sentences that are semantically incompatible with the original from the search results. However, under circumstances where no stub is attached to the queries (where the query results should ideally consist of only sentences that end with the query string), the search engine may produce more results than needed. Sentences conforming to the pattern `it + copula + missing-object construction`, such as (referring to a book) '*it is easy to read*', present one such situation. What is unique about the construction – and why special treatment is needed – is that a missing-object construction usually has an *it*-extraposition counterpart in which the object is present, for example '*it is easy to read the book*'. Since the missing-object constructions are virtually the same (only shorter) as their extrapositional counterparts, there is a good chance for them to be identified as extrapositions. The following are some additional examples of the missing-object construction:

(5.33) Where non-violent civil disobedience is the centerpiece, rather than a lawful demonstration that may only attract crime, *it* is difficult to justify. WSJ 290:25

[14] Instances containing object complements are treated under the framework of subject extraposition and are not included here.

(5.34) No price for the new shares has been set. Instead, the companies will leave *it* up to the marketplace to decide. WSJ 18:24-25

(5.35) He declined to elaborate, other than to say, "*It* just seemed the right thing to do at this minute. WSJ 111:5

Two sets of patterns are proposed[15] to identify the likes of the foregoing examples. The first pattern, the compound adjective test, is inspired by Nanni's (1980) study considering the *easy*-type adjective followed by an infinitive (also commonly termed *tough* construction) as a single complex adjective. The pattern takes the form

$$\texttt{stub + adjective}_{base}\texttt{-}\textit{to}\texttt{-verb} \tag{5.8}$$

where the stub, serving to limit the outcome of the query to noun phrases, takes a combination of determiners or *a/an* alone; the original adjective is also converted to its base form $\texttt{adjective}_{base}$ if it is in comparative or superlative form. Expanding on Nanni's original claims, the pattern can be used to evaluate all adjectives[16] as well as constructs furnished with *for* ... infinitive complements. The following example demonstrates the pattern's usage:

- The machine uses a single processor, which makes *it* easier to program than competing machines using several processors. WSJ 258:24

⇒ an easy-to-program

The second set consists of two patterns used for comparative analysis with the same general profile:

$$\textit{that}\ \texttt{+ verb}_{gerund}\ \texttt{+ stub} \tag{5.9}$$

where $\texttt{verb}_{gerund}$ is the gerund form of the original infinitive. The complementizer *that* is used for the sole purpose of ensuring that $\texttt{verb}_{gerund}$ appears as the subject of a subordinate clause in all sentences returned by the queries. In other words, phrases such as '*computer programming*' and '*pattern matching*' are excluded. For the first pattern, the stub is a combination of prepositions (currently *in* and *from* are chosen); for the second one, a combination of determiners or *the* alone is used. For example:

- The machine uses a single processor, which makes *it* easier to program than competing machines using several processors. WSJ 258:24

⇒ *that* programming $\left\{ \begin{array}{l} \textit{in|from} \\ \textit{the} \end{array} \right\}$

This set of patterns tests the transitivity of the verb in a semantic environment similar to that of the original sentence. If the verb is used transitively more often, the pattern with determiners should

[15] Preliminary experiments have confirmed the effectiveness of the patterns. However, due to sparseness of samples belonging to this class, they are not included in the reported evaluation.

[16] This is based on the observation that compounds such as *ready-to-fly* (referring to model aircrafts) exist, and that it is hard to obtain a complete enumeration of the *easy*-type adjectives.

yield more results, and vice versa. As supported by all preceding sample sentences, a usually-transitive verb used without an object[17] is a good indicator of missing-object construction and the sentence should be diagnosed as referential.

Query Instantiation

Patterns must be instantiated with information found in original sentences before they are submitted to a search engine. Considering the general design principles of the system, it is not advisable to instantiate the patterns with original texts – doing so significantly reduces the queries' coverage. Instead, the object of the matrix verb phrase is truncated and the matrix verb expanded in order to obtain the desired level of coverage.

The truncation process provides different renditions based on the structure of the original object:

- Adjective phrases:

  Only the head word is used. When the head word is modified by *not* or *too*, the modifier is also retained in order to better support the *too* … *to* construct and to maintain compatibility with the semantics of the original text.

- Common noun phrases:

  - with a possessive ending/pronoun, or an *of*-preposition:

    The phrase is replaced by `$PRPS$` plus the head word. `$PRPS$` is either a list of possessive pronouns or one of those more widely used, depending on caliber of the search engine used. For example, '*his location*' can be expanded to '*its* | *my* | *our* | *his* | *her* | *their* | *your location*'.

  - with determiners:

    The phrase is replaced by a choice of `$DTA$`, `$DTTS$`, `$DTTP$`, or a combination of `$DTA$` and `$DTTS$`, plus the head word. `$DTA$` is a list of (or one of the) general determiners (i.e. *a*, *an*, *any* etc.). `$DTTS$` refers to the combination of the definite article *the* and the singular demonstratives *this* and *that*. `$DTTP$` is the plural counterpart of `$DTTS$`. The choice is based on the configuration of the original text so as to maintain semantic compatibility.

  - without determiner:

    Only the head word is used.

- Proper nouns and pronouns:

  The phrase is replaced by `$PRP$`, which is a list of (or one of the) personal pronouns.

- Prepositional phrases:

  The object of the preposition is truncated in a recursive operation.

[17] An omitted object of a preposition (e.g. '*It is difficult to account for.*') has the same effect, but it is identifiable through syntactic means alone.

- Numeric values:

  The phrase '*a lot*' is used instead.

Matrix verbs are expanded to include both the simple past tense and the third person singular present form with the aid of WORDNET and some generic patterns. Where applicable, particles such as *out* and *up* also remain attached to the verb.

Generally speaking, truncation and expansion are good ways of boosting the patterns' coverage. However, the current procedures of truncation are still crude, especially in their handling of complex phrases. For example, the phrase '*a reckless course of action*' (WSJ 198:11) yields '`$PRPS$` course', which results in a total loss of the original semantics. Further enhancements of the truncation process may improve the performance but the improvement will likely be limited due to the endless possibilities of language usage and constraints imposed by search engines.

Aside from truncating and expanding the original texts, a stepped-down version of Pattern II, denoted Pattern II$'$, is also provided to further enhance the system's coverage. The current scheme is to simply replace the extraposed clause with a new stub – *to* – if the original extraposed clause is an infinitive, a *for* … infinitive, or a gerund construct. For example,

- *It* is easy to see why the ancient art is on the ropes†. WSJ 89:17

$$\Rightarrow \left\{ \begin{array}{l} \textit{it} \\ \textit{which/who/this/he} \end{array} \right\} \text{ is easy to}$$

In other situations, no downgraded version is applied.

5.3.4 Binary Classification of *It*-extraposition

Five factors are taken into consideration when determining whether the sentence in question is an *it*-extraposition:

Estimated popularity of the *what*-cleft construct (query Pattern I)

denoted as

$$W = n_w \times v_w$$

where n_w is the number of results reported by the search engine, and v_w is the percentage of valid instances within the first batch of snippets (usually 10, depending on the search engine service) returned with the query. Validation is performed with a case-sensitive regular expression derived from the original query. Since the *what*-cleft pattern is capitalized at the beginning, the regular expression only looks for instances appearing at the beginning of a sentence. It is particularly important to validate the results of *what*-cleft queries because some search engines can produce results based on their own interpretation of the original query. For example, Google returns pages containing "*What's found is that*" for the query "*What found is that*", which might be helpful for some but is counterproductive for the purpose of this study.

Result of the comparative expletiveness test (query Pattern II)

denoted as

$$r = \frac{n_X}{n_{it}}$$

where n_{it} is the number of results obtained from the original *it* version of the query, and n_X is the total number of results produced by replacing *it* with other pronouns such as *which* and *who*. The smaller the ratio r is, the more likely that the sentence being investigated is an extraposition. Extrapositional sentences usually produce an r value of 0.1 or less. When both versions of the query yield insufficient results ($max(n_{it}, n_X) < N_{min}$), r takes the value $R_{scarce} = 1000$. Since *it*-extrapositions are relatively rare, it is better to assume that a sentence is not extrapositional when there is insufficient data to judge otherwise. In the case where n_X is sufficient but the *it* version of the query produces no result ($n_X >= N_{min}$ AND $n_{it} = 0$), r takes the value $R_{zero} = 100$. Values of R_{zero} and R_{scarce} are large numbers chosen arbitrarily, mainly for visualization purposes. In other words both R_{zero} and R_{scarce} hint that the sentence is probably not extrapositional, however neither indicates the degree of likelihood.

Result of the stepped-down comparative expletiveness test

denoted as $r' = \frac{n'_X}{n'_{it}}$, where n'_{it} and n'_X are the number of results returned from the *it* version and the alternate version of the stepped-down queries (cf. Section 5.3.3, Page 146). The stepped-down queries are 'simplified' versions of the queries used to calculate r. Due to this simplification, r' is usually more sensitive to extrapositions. However not all queries have stepped-down versions, in which case the original queries are reused, causing $r' = r$. Similar to the way r is defined, r' also takes the values R_{scarce} and R_{zero} in special situations.

Synthesized expletiveness

A new variable R is defined based on the values of r, n_{it}, n_X , and r':

$$R = \begin{cases} r, & \text{if } \max(n_{it}, n_X) \geq N_{min}, \\ r', & \text{if } \max(n_{it}, n_X) < N_{min}. \end{cases}$$

If the original queries yield enough results, R takes the value of r since the original queries better preserve sentence context and are generally more accurate. However, when original queries fail, the system resorts to the back-up method of using the stepped-down queries and bases its judgement on their results instead. Overall, R can be seen as a synthesized indicator of how the subject pronoun is generally used in a similar syntactic and semantic setting to that of the original sentence.

Syntactic structure of the sentence

denoted as S, a binary variable indicating if the sentence under investigation belongs to a syntactic construct that is more prone to generating false-positives. On average the *what*-cleft queries yield fewer results and are less reliable since they cannot be used to provide comparative ratios. However, they are still useful as the last line of defence to curb the impacts of certain syntactic constructs that repeatedly cause the comparative expletive tests to produce false-positives. Currently only one construct is identified – the `it verb infinitive` construct, as in '*it helps to have input from everyone*' and '*it expects to post the results tomorrow*'. Therefore,

$$S = \begin{cases} \text{TRUE}, & \text{if sentence matches } \texttt{it verb infinitive}, \\ \text{FALSE}, & \text{otherwise.} \end{cases}$$

The final binary classification of it-extraposition, E, is defined as follows:

$$E = \begin{cases} ((R < R_{exp}) \text{ AND } (W > N_{min})), & \text{if } S = \text{TRUE}, \\ (R < R_{exp}), & \text{if } S = \text{FALSE}. \end{cases} \tag{5.10}$$

where N_{min} and R_{exp}, set to 10 and 0.15 respectively in this study, are threshold constants chosen based upon empirical observations. In other words, the system recognizes an instance of *it* as extrapositional if it is unlikely (by comparing R to R_{exp}) that an alternative pronoun is used in its place under the same syntactic and semantic settings. For `it verb infinitive` constructs, it is also required that the sentence has a viable *what*-cleft variant (by comparing W to N_{min}).

It is worth noting that today's major commercial search engines do not return the exact number of results for a query but rather their own estimates. The negative effect of this is somewhat mitigated by basing the final decision on ratios instead of absolute numbers.

5.4 Case Study

To better illustrate the system work flow, two sample sentences are selected from the WSJ corpus to be taken through the whole process. The first sample, (5.36), is classified as an it-extraposition; the other, (5.37), is a referential case with a nominal antecedent. Some particulars of the implementation are also discussed here.

(5.36) A fund manager at a life-insurance company said three factors make *it* difficult to read market direction[†]. WSJ 231:15

(5.37) Her recent report classifies the stock as a "hold." But *it* appears to be the sort of hold one makes while heading for the door. WSJ 331:32-33

5.4.1 Syntactic Filtering

First, the syntactic structures of each sentence are identified and dependencies among the constituents are established, as shown in Figures 5.3 and 5.4.

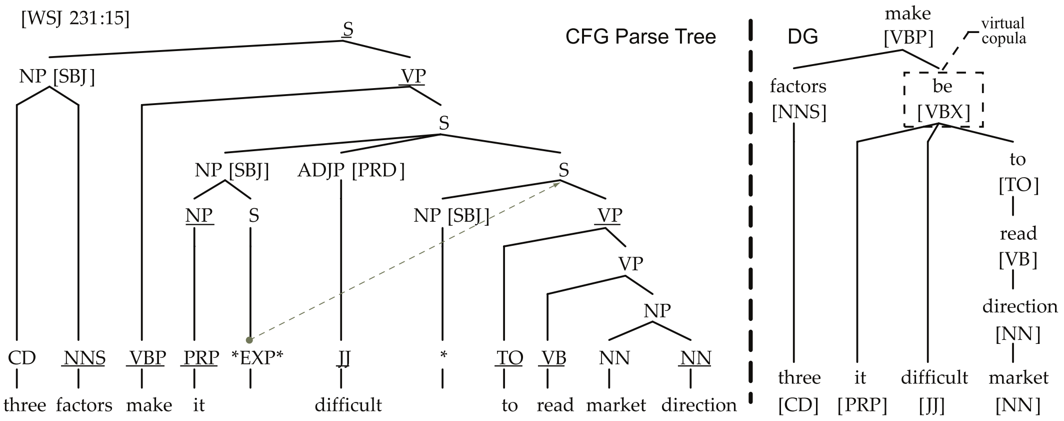

Figure 5.3: Syntactic structure of example (5.36) (fragment)

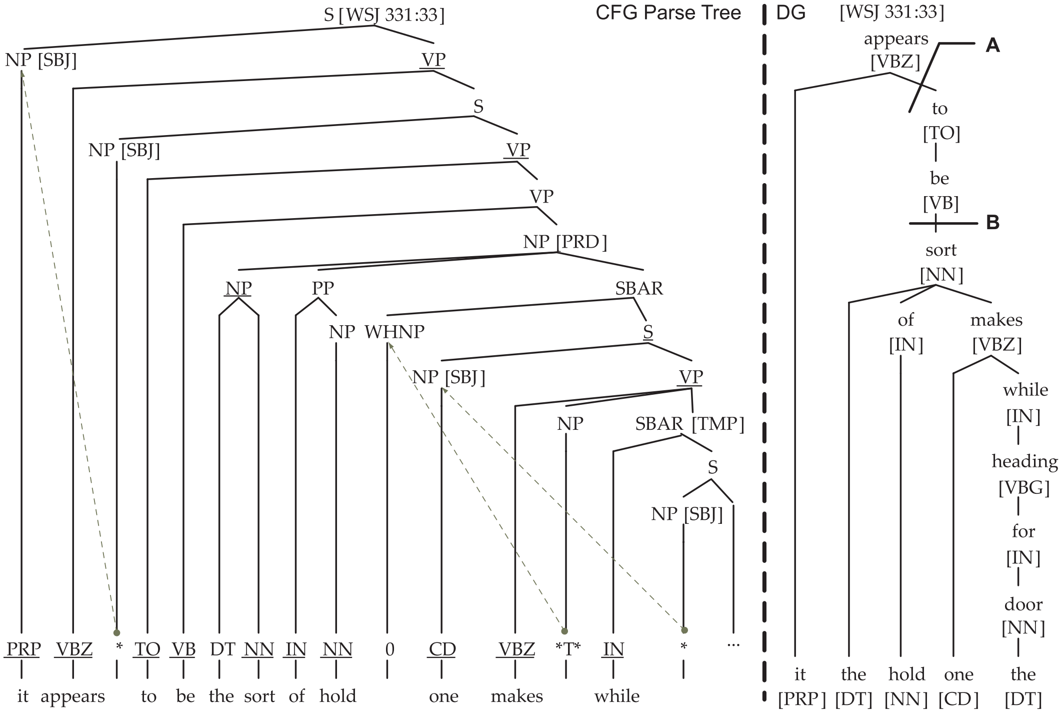

Figure 5.4: Syntactic structure of example (5.37) (fragment). Readings A and B, as indicated in the DG parse tree, are discussed in the text.

In example (5.36), the expletive *it* appears as the object of the verb *makes* and is followed by the object complement *difficult*, therefore a virtual copula (tagged VBX) is created in the dependency tree in order to treat it under the same framework as subject *it*-extrapositions. For (5.37), two different readings are produced – one by assuming *appears* to be the matrix verb (reading A, cf. Figure 5.4), the other by taking *be* (reading B). This is accomplished by 'drilling' down the chain of verbs beginning with the parent verb of the *it* node. Once at the top of the chain, the system starts a recursive process to find verbs and infinitives that are directly attached to the current node and moves down to the newly found node. The process is interrupted if the current verb node is furnished with elements other than verbal or adverbial complements/modifiers.

During the filtering process, various components of the sentences are identified, as listed in Table 5.2.

Sentence	Reading	Matrix		Conjunction	Subordinate		
		Verb	**Object**		**Subject**	**Verb**	**Object**
(5.36)		be	difficult			to read	direction
(5.37)	A	appears				to be	sort
(5.37)	B	be	sort	THAT	One		

Table 5.2: Component breakdown of the case study samples

5.4.2 Pattern Instantiation

Using the components identified in Table 5.2, five queries are generated for each reading, as listed in Tables 5.3-5.5. Patterns II′-*it* and II′-others refer to the stepped-down versions (cf. Section 5.3.3, Page 146) of II-*it* and II-others respectively. The queries shown here are generated specifically for Google and take advantage of features only available in Google. To use an alternative search engine such as Yahoo, the component expansions and determiner lists have to be turned off, and separate queries need to be prepared for individual pronouns. In order to get accurate results, the queries must be enclosed in double quotes before they are sent to search engines.

Pattern	Query	Results
I	what is|was|'s difficult is|was to	1060
II-*it*	it is|was|'s difficult to read the|a|an|no|this|these|their|his|our	3960
II-others	which|this|who|he is|was|'s difficult to read the|a|an|no|this|these|their|his|our	153
II′-*it*	it is|was|'s difficult to	6.3×10^6
II′-others	which|this|who|he is|was|'s difficult to	1.5×10^5

Table 5.3: Queries for example (5.36)

Pattern	Query	Results
I	what appears|appeared is|was to	44
II-*it*	it appears|appeared to be the|a|an|no|this|these|their|his|our	7.5×10^4
II-others	which|this|who|he appears|appeared to be the|a|an|no|this|these|their|his|our	3.2×10^5
II′-*it*	it appears|appeared to	2.2×10^6
II′-others	which|this|who|he appears|appeared to	2.6×10^6

Table 5.4: Queries for example (5.37), Reading A

Pattern	**Query**	**Results**
I	what is|was|’s its|my|our|his|her|their|your sort is|was that	0
II-*it*	it is|was|’s its|my|our|his|her|their|your sort that the|a|an|no|this|these|they|we|he|their|his|our	0
II-others	which|this|who|he is|was|’s its|my|our|his|her|their|your sort that the|a|an|no|this|these|they|we|he|their|his|our	0
II′-*it*	Same as II-*it*	0
II′-others	Same as II-others	0

Table 5.5: Queries for example (5.37), Reading B

5.4.3 Query Results and Classification

For every reading, the number of results for each of the five queries (n_w for Pattern I; n_{it} for II-*it*; n_X for II-others; n'_{it} for II′-*it*; and n'_X for II′-others) is obtained from the search engine; the first 10 results for the *what*-cleft query are also validated to obtain the estimated percentage (v_w) of valid constructs. $W(= n_w \times v_w)$, $r(= n_X/n_{it})$, $r'(= n'_X/n'_{it})$, and R (choosing between either r or r' depending on whether $max(n_{it}, n_X) \geq 10$) are then calculated accordingly, as recorded in Table 5.6.

Query	n_w	v_w	n_{it}	n_X	n'_{it}	n'_X	W	r	r'	R
(5.36)	1060	70%	3960	153	6.3E6	1.5E5	742	0.04	0.02	0.04
(5.37).A	44	0%	7.5E4	3.2E5	2.2E6	2.6E6	0	4.3	1.2	4.3
(5.37).B	0	-	0	0	0	0	0	1000	1000	1000

Table 5.6: Query results for the case study sample sentences

What appears suspicious is that v_w is set to 0 for reading (5.37).A, which means no valid instances are found. A quick look at the returned snippets reveals that, indeed, none of the 10 snippets has the queried contents at the beginning of sentence. Also note that for reading (5.37).B, both r and r', and consequently R have all been set to $R_{scarce} = 1000$ since no query produced enough results.

It can be decided from Table 5.2 that readings (5.36) and (5.37).B do not bear the `it verb infinitive` construct, hence $S =$ FALSE; and for (5.37).A $S =$ TRUE. Applying Equation 5.10 in Section 5.3.4, for (5.36) and (5.37).B, the final classification E is only based on whether R is sufficiently small ($R < 0.15$). For (5.37).A, the system also needs to check whether the *what*-cleft query returned sufficient valid results ($W > 10$). The final classifications are listed in Table 5.7.

Sentence	**Reading**	W	S	R	$E_{reading}$	E
(5.36)	-	742	FALSE	0.04	YES	**YES**
(5.37)	A	0	TRUE	4.3	NO	**NO**
(5.37)	B	0	FALSE	1000	NO	

Table 5.7: Final binary classification of the case study sample sentences

Since neither readings of (5.37) are classified as such, the sentence is not an *it*-extraposition construct.

5.5 First Evaluation

In order to provide a comprehensive picture of the system's performance, a twofold assessment is used. In the first evaluation, the system is exposed to the same sentence collection that assisted its development. Accordingly, results obtained from this evaluation reflect, to a certain degree, the system's optimal performance. The second evaluation (detailed in Section 5.6) aims at revealing the system's performance on unfamiliar texts by running the developed system on a random dataset drawn from the rest of the corpus. Two additional experiments are also conducted to provide an estimation of the system's performance over the whole corpus.

The same set of performance measures as introduced in Section 4.7, namely precision, recall, and the balanced F-measure, are also used in this chapter. To recap, the measures are defined as $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$, and $F = \frac{2PR}{P+R}$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. Following Efron and Tibshirani's (1993) Bootstrap method, 95% confidence intervals are obtained using the 2.5^{th} and 97.5^{th} percentiles of the bootstrap replicates and are provided alongside the system performance figures to indicate their reliability. The number of replicates is arbitrarily set at $B = 9999$, which is much greater than the commonly suggested value of 1000 (e.g., see Davison & Hinkley, 1997; Efron & Tibshirani, 1993) because pleonastic instances are sparse. In the case that a precision or recall value is 100%, the bootstrap percentile method reports an interval of 100%-100%, which makes little sense. Therefore, in this situation the adjusted Wald interval (Agresti & Coull, 1998) is presented instead. When two systems are compared, an approximate randomization test (Noreen, 1989) similar to that used by Chinchor (1992) is performed to determine if the difference is of statistical significance. The significance level $\alpha = 0.05$ and number of shuffles $R = 9999$, both chosen arbitrarily, are used where significance tests are performed.

5.5.1 Development Dataset

For the purpose of this study, the first 1000 occurrences of *it* from the WSJ corpus have been manually annotated. A part of the set has also been inspected in order to determine the values of the constants specified in Section 5.3.4, and to develop the surface structure processor. The annotation process is facilitated by a custom-designed utility that displays each sentence within its context represented by a nine-sentence window containing the six immediately preceding sentences, the original, and the two sentences that follow. Post-annotation review indicates that this presentation of corpus sentences worked well. Except for a few (less than 0.5%) cases, there is no need to resort to broader contexts to understand a sentence; and under no circumstances were valid antecedents located outside the context window while no antecedent was found within it.

Category	**Instances**	**Percentage**
Nominal Antecedent	756	75.60%
Clause Antecedent	60	6.00%
Extraposition	118	11.80%
Cleft	13	1.30%
Weather/Time	9	0.90%
Idiom	18	1.80%
Other	26	2.60%
Grand Total	**1000**	**100.00%**

Table 5.8: Profile of the development dataset according to the author's annotation

Table 5.8 summarizes the distribution of instances in the dataset. The category labeled 'Other' consists mostly of instances that do not fit well into any other categories, e.g. when the identified nominal antecedent is in plural or the antecedent is inferred, as well as certain confusing instances. Out of the twenty-six instances, only two might be remotely recognized as one of the types that interests this study:

(5.38) And though the size of the loan guarantees approved yesterday is significant, recent experience with a similar program in Central America indicates that *it* could take several years before the new Polish government can fully use the aid effectively. WSJ 101:7

(5.39) *It*'s just comic when they try to pretend they're still the master race. WSJ 296:48

Neither instance can be identified as anaphoric. However, the first construct has neither a valid non-extraposition version nor a valid *what*-cleft version, making it difficult to justify as an extraposition, while the *it* in the second case is considered to refer to the atmosphere aroused by the action detailed in the when-clause.

In order to assess whether the pleonastic categories are well-defined and the ability of ordinary language users to identify pleonastic instances, two volunteers, both native English speakers, are invited to classify the *it* instances in the development dataset. To help them concentrate on the pleonastic categories, the volunteers are only required to assign each instance to one of the following categories: referential, extraposition, cleft, weather/time, and idiom. The referential category covers instances with both nominal antecedents and clause antecedents, as well as instances with inferrable antecedents. Table 5.9 outlines both annotators' performance in reference to the author's annotation. The degree of agreement between the annotators, measured by the kappa coefficient (κ; Cohen, 1960), is also given in the same table.

Category	Volunteer 1			Volunteer 2			κ[a]
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	
Referential	99.38%	95.49%	97.40%	96.38%	98.10%	97.23%	.749
Extraposition	82.54%	88.14%	85.25%	88.68%	79.66%	83.93%	.795
Cleft	38.46%	76.92%	51.28%	72.73%	61.54%	66.67%	.369
Weather/Time	66.67%	44.44%	53.33%	75.00%	33.33%	46.15%	-.005
Idiom	39.39%	72.22%	50.98%	50.00%	61.11%	55.00%	.458
Overall Accuracy/κ		**93.50%**			**94.20%**		**.702**

[a] Except for the Weather/Time category ($p = 0.5619$), all κ values are statistically significant at $p < 0.0001$.

Table 5.9: Performance of the volunteer annotators on the development dataset (evaluated using the author's annotation as reference) and the degree of inter-annotator agreement measured by Cohen's kappa (κ). The author's annotations are refitted to the simplified annotation scheme used by the volunteers.

There are many factors contributing to the apparently low κ values in Table 5.9, most notably the skewed distribution of the categories and inappropriate communication of the classification rules. As Di Eugenio and Glass (2004) and others pointed out, skewed distribution of the categories has a negative effect on the κ value. Since the distribution of the *it* instances in the dataset is fairly unbalanced, the commonly-accepted guideline for interpreting κ values ($\kappa > 0.67$ and $\kappa > 0.8$ as thresholds for tentative and definite conclusions respectively; Krippendorff, 1980) may not be directly applicable in this case. In addition, the classification rules are communicated to the annotators orally through examples and some of the not-so-common cases, such as the object *it*-extrapositions, might not have been well understood by both annotators. Another interesting note about the results is that there is a strong tendency for both annotators (albeit on different cases) to classify *it*-clefts as *it*-extrapositions. Rather than taking this as a sign that the cleft category is not well-defined, we believe it reflects the inherent difficulties in identifying instances pertaining to the category.

5.5.2 Baselines

Two baselines are available for comparison – the WSJ annotation, which is done manually and provided with the corpus; and the results from a replication of Paice and Husk's (1987) algorithm (PHA). It should be cautioned that, given the subjectivity of the issues discussed in this study and lack of consensus on certain topics in the field of linguistics, recall ratios of the presented baseline results and the forthcoming results of the system developed in this study should not be compared quantitatively. For example, the original Paice and Husk algorithm does not recognize certain types of object extrapositions and does not always distinguish between individual types of pleonastic *it*; and the WSJ corpus has neither special annotation for parenthetical *it* (e.g. 5.20 on page 135) nor an established annotation policy for certain types of object extrapositions (Bies et al., 1995). No attempts have been made to correct these issues.

Table 5.10 summarizes the performance of the baselines on the development dataset. As expected, Paice and Husk's (1987) algorithm does not perform very well since the WSJ articles are very different from, and tend to be more sophisticated than, the technical essays that the algorithm was designed for. Compared to the originally reported precision of 93% and recall of 96%, the replicated PHA yields only 54% and 75% respectively on the development dataset. The performance of the replica is largely in line with what Boyd et al. (2005) obtained from their implementation of the same algorithm on a different dataset.

	WSJ Annotation		**Replicated PHA**
Measurement	**Extraposition**	**Cleft**	**Overall**[a]
Reference	118	13	140
Identified by Baseline	88	12	194
Baseline True Positives	87[b]	12	105
Precision	98.86%	100%	54.12%
Recall	73.73%	92.31%	75.00%
F-measure	84.47%	96.00%	62.87%

[a] Includes clefts, extrapositions, and time/weather cases.

[b] Based on manual inspection, two cases originally annotated as extrapositional in WSJ are determined as inappropriate. See discussions below.

Table 5.10: Performance of the baselines on the development dataset, evaluated against the author's annotation.

The 31 (118 − 87) extrapositional cases that are not annotated in WSJ can be broken down into the following categories followed by their respective number of instances:

Category	**Items**
Unrecognized	3
Object without complement	1
Parenthetical	2
Inappropriate non-extraposition	18
Agentless passive	9
it seems/appears …	4
it be worth …	2
Others	3
Valid non-extraposition	10
too … to	2
Others	8
Total	**31**

Table 5.11: Profile of the false negatives in the WSJ annotation in reference to the author's annotation

By stating that the 'Characteristic of *it* extraposition is that the final clause can replace it', Bies et al. (1995) define the class in the narrowest sense. Since interpretation of the definition is entirely a subjective matter, there is no way of determining the real coverage of the annotations. However, from the portions of the corpus that have been reviewed, the practice of annotation is not entirely consistent.

Two sentences are marked as extraposition in the corpus but marked otherwise in the author's annotation. Considering the 'golden standard' status of the WSJ corpus, they are also listed here:

(5.40) Moreover, as a member of the Mitsubishi group, which is headed by one of Japan's largest banks, *it* is sure to win a favorable loan. WSJ 277:40

(5.41) *It* is compromises such as this that convince Washington's liberals that if they simply stay the course, this administration will stray from its own course on this and other issues. WSJ 303:6

The first sentence is considered dubious and most likely referring to the company that is a member of the Mitsubishi group. The second one is considered a cleft and is actually also marked as cleft in the corpus. Since it is the only case in the corpus with both annotations, the extraposition marking was considered a mistake and was manually removed.

The Paice and Husk (1987) algorithm suffers from false-positive *it . . . that* and *it . . . to* construct detection, which may be fixed by incorporating part-of-speech and phrase structure information together with additional rules. However, such fixes will greatly complicate the original system.

5.5.3 Results

On the development dataset, results produced by the system are as follows:

Measurement	**Extraposition**	**Cleft**	**Weather/Time**	**Overall**[a]
Reference	118	13	9	140
Identified	116	13	10	139
True Positives	113	13	9	136
Precision	97.41%	100.00%	90.00%	97.84%
95% C.I.[b]	94.07-100.00%	79.74-100.00%	66.67-100.00%	95.21-100.00%
Recall	95.76%	100.00%	100.00%	97.14%
95% C.I.[b]	91.79-99.12%	79.74-100.00%	73.07-100.00%	93.98-99.34%
F-measure	96.58%	100.00%	94.74%	97.49%
95% C.I.	93.98-98.72%	-	80.00-100.00%	95.45-99.21%

[a] Combining extraposition, cleft, and weather/time into one category.
[b] Adjusted Wald intervals are reported for extreme measurements.

Table 5.12: Performance of the system on the development dataset, evaluated using the author's annotation as reference.

Further statistical significance tests reveal more information regarding the system's performance in comparison to that of the two volunteers and the baselines:

- Compared to both volunteer annotators, the system's better performance in all three pleonastic categories is statistically significant.
- In the extraposition category, the difference between the WSJ annotation's (higher) precision and that of the system is not statistically significant.
- Compared to Paice and Husk's (1987) algorithm, the system's higher precision is statistically significant.

Target System	Extraposition	Cleft	Weather/Time
Volunteer 1	F-measure$^+/p < .001$	F-measure$^+/p < .001$	F-measure$^+/p = .033$
Volunteer 2	F-measure$^+/p < .001$	F-measure$^+/p = .007$	F-measure$^+/p = .025$
WSJ Annotation	Precision$^-/p = .630$	F-measure$^+/p = 1.00$	
Replicated PHA	(All Categories) Precision$^+/p < .001$		

Table 5.13: Results of the statistical significance tests presented in the format Test Statistic$^{sign}/p$-value. A plus sign ($^+$) indicates that our system performs better on the reported measurement; otherwise a minus sign ($^-$) is used. If fair comparisons can be made for both precision and recall, the F-measure is used as the test statistic; otherwise the applicable measurement is reported.

Using the author's annotation as reference, the system outperforms both human volunteers. While higher performance is usually desirable, in this particular case, it could indicate possible problems in the design of the experiment. Since the English language is not only used by its speakers but also shaped by the same group of people, it is impractical to have a system that 'speaks better English' than its human counterparts do. One plausible clue to the paradox is that an analytic approach is needed to gain insight into the issue of pronoun classification, but the casual English speakers do not see it from that perspective. As Green and Hecht (1992) and many others indicated, capable users of a language do not necessarily have the ability to formulate linguistic rules. However, these kinds of analytic skills is a prerequisite in order to explicitly classify a pronoun into one of the many categories. Thus, the true performance of casual speakers can only be measured by their ability to comprehend or produce the various pleonastic constructs. In addition, other factors, such as time constraints and imperfections in how the category definitions are conveyed, may also play a role in limiting the volunteers' performance. The author's annotation, on the other hand, is much less influenced by such issues and is therefore considered expert opinion in this experiment. As shown in Section 5.5.2, the WSJ annotation of extrapositions and clefts, which is also considered expert opinion, is highly compatible with that of the author. The differences between the two annotations can mostly be attributed to the narrower definition of extraposition adopted by the WSJ annotators. Therefore, the WSJ annotation's precision of 98.86% for extrapositions (when verified against the author's annotation) is probably a more appropriate hint of the upper-limit for practically important system performance.

In the extraposition category, 279 individual cases passed the syntactic filters and were evaluated by search engine queries. Results of queries are obtained from Google through its web service, the Google SOAP[18] Search API. All three ($116-113$) cases of false-positives are caused by missing-object constructions and can be corrected using the patterns detailed in Section 5.3.3.

The five ($118-113$) false-negative cases are listed below:

(5.42) The newspaper said *it* is past time for the Soviet Union to create unemployment insurance and retraining programs like those of the West. WSJ 283:13

[18] The Simple Object Access Protocol is an XML-based message protocol for web services.

(5.43) “*It*’s one thing to say you can sterilize, and another to then successfully pollinate the plant,” he said. WSJ 209:40

(5.44) Sen. Kennedy said ... but that *it* would be a “reckless course of action” for President Bush to claim the authority without congressional approval. WSJ 198:11

(5.45) Worse, *it* remained to a well-meaning but naive president of the United States to administer the final infamy upon those who fought and died in Vietnam. WSJ 290:49

(5.46) “*It*’s not easy to roll out something that comprehensive, and make it pay,” Mr. Jacob says. WSJ 85:47

Sentence (5.42) is misplaced as weather/time. Sentence (5.43) is not properly handled by the syntactic processing subcomponent. Sentences (5.44) and (5.45) involve complex noun phrases (underlined) at the object position of the matrix verbs – it is very difficult to reduce them to something more generic, such as the head noun only or a pronoun, and still remain confident that the original semantics are maintained. The last case, sentence (5.46), fails because the full queries (containing part of the subordinate clause) failed to yield enough results and the stepped-down versions are overwhelmed by noise.

The last four false-negatives are annotated correctly in the WSJ corpus. The system’s recall ratio on the 87 verified WSJ extraposition annotations is therefore 95.40%, comparable to the overall recall.

5.5.4 System Performance on Parser Output

Thus far, the system has been evaluated based on the assumption that the underlying sentences are tagged and parsed with (almost) perfect accuracy. Much effort has been made to reduce such dependency. For example, tracing information and function tags in the original phrase structures are deliberately discarded; and the system also tries to search for possible extraposed or cleft clauses that are marked as complements to the matrix object. However, deficiencies in tagging and parsing may still impact the system’s performance. Occasionally, even the ‘golden standard’ manual markups appear problematic and happen to get in the way of the task.

It is therefore necessary to evaluate the system on sentences that are automatically tagged and parsed in order to answer the question of how well it would perform in the real world. Two state-of-the-art parsers are employed for this study: the reranking parser by Charniak and Johnson (2005), and the Berkeley parser by Petrov, Barrett, Thibaux, and Klein (2006). The system’s performance on their respective interpretations of the development dataset sentences are reported in Tables 5.14 and 5.15. Table 5.16 further compares the system’s real-world performance to the various baselines.

Measurement	Extraposition	Cleft	Weather/Time	Overall[a]
Reference	118	13	9	140
Identified	114	12	10	136
True Positives	110	12	9	132
Precision	96.49%	100.00%	90.00%	97.06%
95% C.I.[b]	92.68-99.20%	78.40-100.00%	66.67-100.00%	93.92-99.32%
Recall	93.22%	92.31%	100.00%	94.29%
95% C.I.[b]	88.43-97.41%	73.33-100.00%	73.07-100.00%	90.18-97.81%
F-measure	94.83%	96.00%	94.74%	95.65%
95% C.I.	91.60-97.49%	84.62-100.00%	80.00-100.00%	93.08-97.90%

[a] Combining extraposition, cleft, and weather/time into one category.

[b] Adjusted Wald intervals are reported for extreme measurements.

Table 5.14: Performance of the system on the development dataset parsed by the Charniak parser, using the author's annotation as reference.

Measurement	Extraposition	Cleft	Weather/Time	Overall[a]
Reference	118	13	9	140
Identified	114	11	9	134
True Positives	111	10	8	130
Precision	97.37%	90.91%	88.89%	97.01%
95% C.I.	94.07-100.00%	70.00-100.00%	62.50-100.00%	93.81-99.32%
Recall	94.07%	76.92%	88.89%	92.86%
95% C.I.	89.47-98.18%	50.00-100.00%	62.50-100.00%	88.44-96.91%
F-measure	95.69%	83.33%	88.89%	94.89%
95% C.I.	92.75-98.17%	62.50-96.55%	66.67-100.00%	92.02-97.35%

[a] Combining extraposition, cleft, and weather/time into one category.

Table 5.15: Performance of the system on the development dataset parsed by the Berkeley parser, using the author's annotation as reference.

Comparing System Performance On Charniak Parser Output to:

Target System	Extraposition	Cleft	Weather/Time
System w/o Parser	F-measure$^{-}/p = .131$	F-measure$^{-}/p = 1.00$	F-measure$^{=}/p = 1.00$
Volunteer 1	F-measure$^{+}/p = .001$	F-measure$^{+}/p < .001$	F-measure$^{+}/p = .030$
Volunteer 2	F-measure$^{+}/p < .001$	F-measure$^{+}/p = .041$	F-measure$^{+}/p = .021$
WSJ Annotation	Precision$^{-}/p = .368$	F-measure$^{=}/p = 1.00$	
Replicated PHA	(All Categories) Precision$^{+}/p < .001$		

Comparing System Performance On Berkeley Parser Output to:

Target System	Extraposition	Cleft	Weather/Time
System w/o Parser	F-measure$^{-}/p = .380$	F-measure$^{-}/p = .128$	F-measure$^{-}/p = 1.00$
Volunteer 1	F-measure$^{+}/p < .001$	F-measure$^{+}/p = .014$	F-measure$^{+}/p = .061$
Volunteer 2	F-measure$^{+}/p < .001$	F-measure$^{+}/p = .314$	F-measure$^{+}/p = .046$
WSJ Annotation	Precision$^{-}/p = .627$	F-measure$^{-}/p = .374$	
Replicated PHA	(All Categories) Precision$^{+}/p < .001$		

Table 5.16: Results of the statistical significance tests comparing the system's performance on parser output to that of various other systems, presented in the format Test Statistic$^{sign}/p$-value. A plus sign ($^{+}$) indicates that the system developed in this study performs better than the target system on the reported measurement; an equal sign ($^{=}$) indicates a tie; otherwise a minus sign ($^{-}$) is used. If fair comparisons can be made for both precision and recall, the F-measure is used as the test statistic; otherwise the applicable measurement is reported.

Further significance tests reveal that:

- using a parser has no statistically significant influence on the system's performance;
- the system outperforms both volunteer annotators in identifying *it*-extrapositions;
- regardless of the parser used, the difference between the system's performance and that of the WSJ annotation is not statistically significant; and
- regardless of the parser used, the system outperforms the Paice and Husk (1987) algorithm.

5.5.5 Correlation Analysis for Extrapositions

Figures 5.5 through 5.8 illustrate the correlation between the decision factors and the true expletiveness of the pronoun *it* in question. All 279 items that passed the initial syntactic filtering process are included in the dataset with the first 116 being extrapositional and the rest separated by a break on the X-axis. This arrangement is made in order to better visualize the contrast between the positive group and the negative group. In Figures 5.6 through 5.8, different grey levels are used to indicate the number of results returned by queries – the darker the shade, the more popular the construct in question is on the web. The constant $R_{exp} = 0.15$ is also indicated with a break on the Y-axis.

As illustrated, all factors identified in Section 5.3.4 are good indicators of expletiveness. *W* (Figure 5.5) is the weakest of the four factors due to the number of false positives produced by incorrect language usage. This is clear evidence that the web is noisier than ordinary corpora and that the results counts from the web may not be appropriate as the sole decision-making factor.

In comparison, *r* (Figure 5.6) has almost perfect correlation with the expletiveness of instances. However, full versions of the queries usually return fewer results and in many cases yield too few results for expletive cases (unfilled items plotted on top of the graph indicate cases that do not have enough results, cf. Section 5.3.4). The stepped-down versions of the queries (Figure 5.7), while being less accurate by themselves, serve well when used as 'back up', as illustrated by the *R* plot (Figure 5.8). Part of the false-positive outliers on the *R* plot are produced by full queries for expressions that are habitually associated with *it*, such as:

(5.47) The Rockford, Ill., maker of fasteners also said *it* expects to post sales in the current fiscal year that are "slightly above" fiscal 1989 sales of $155 million. WSJ 135:2

When used with a pronoun, these expressions usually describe information quoted from a person or organization already named earlier in the same sentence, making *it* a more natural choice of subject pronoun. Normally the problematic expressions take the form of `verb infinitive-complement`, i.e. $S = \text{TRUE}$. According to the decision process described in Section 5.3.4, W is also considered in this situation, which effectively eliminates such noise.

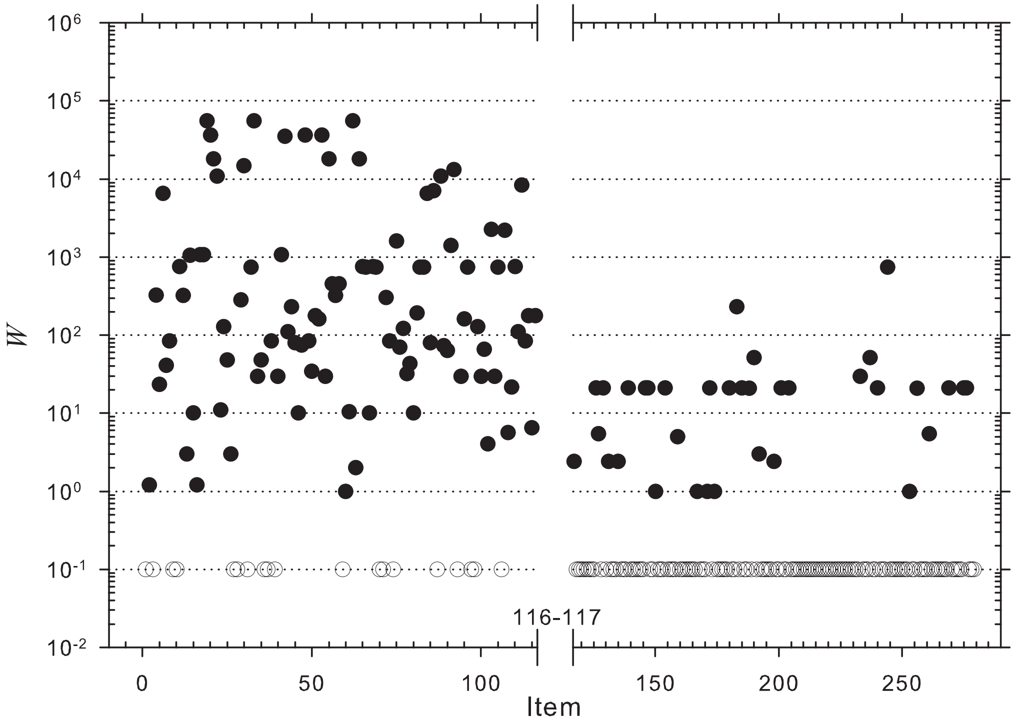

Figure 5.5: A scatter plot illustrating the correlation between W (the estimated number of valid results returned by the *what*-cleft queries) and the expletiveness of the *it* instance. The extrapositional instances are arranged on the left side of the plot and the rest of the cases are on the right. If a query returns no valid results, the corresponding item is shown as a hollow circle on the bottom of the plot.

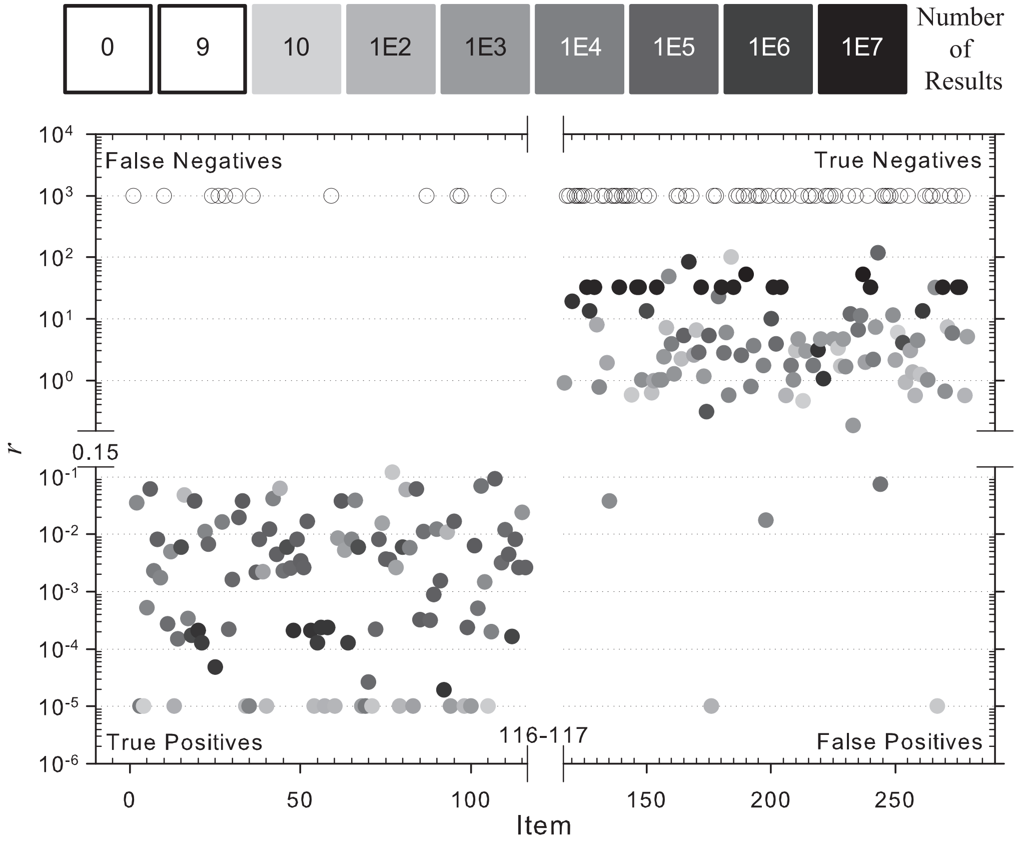

Figure 5.6: A scatter plot illustrating the correlation between *r* (the ratio of the hit count produced by the expression with substitute pronouns to that of the original expression) and the expletiveness of the *it* instance. The extrapositional instances are arranged on the left side of the plot and the rest of the cases are to the right. The items are shaded according to the hit counts produced by the corresponding original expressions. If a query returns insufficient results, the corresponding item is shown as a hollow unshaded circle on the top of the plot.

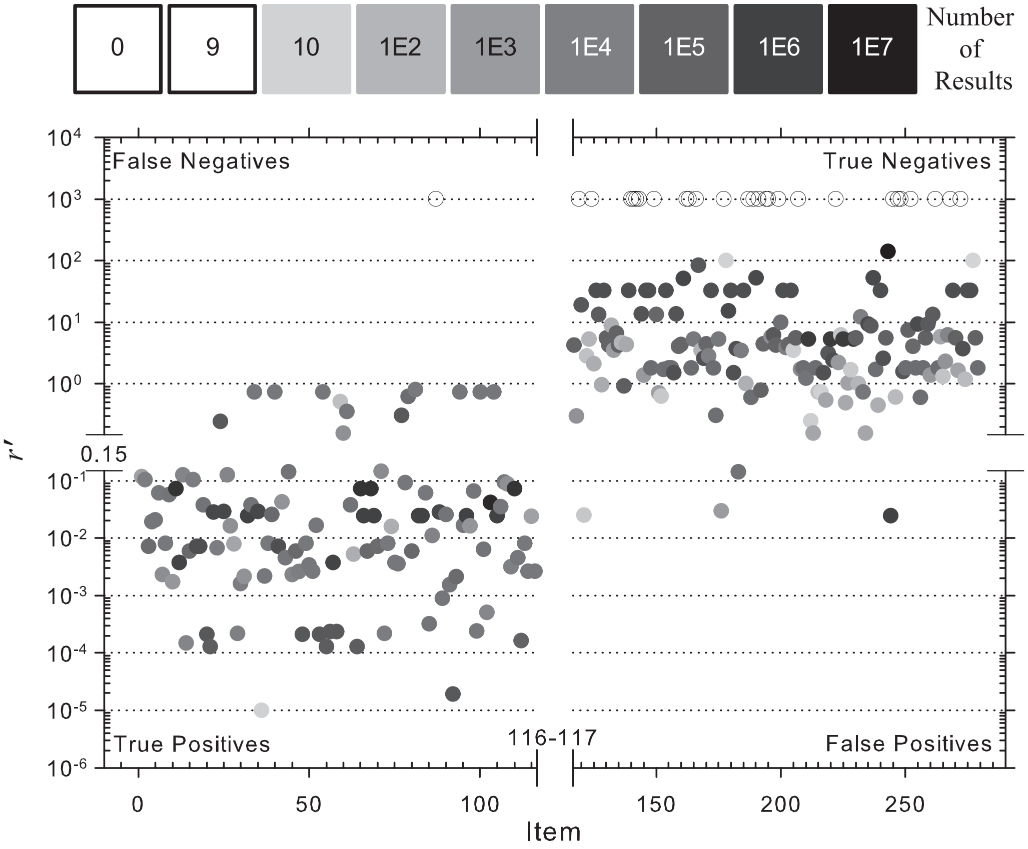

Figure 5.7: A scatter plot illustrating the correlation between r' (similar to r but for the stepped-down queries) and the expletiveness of the *it* instance. The extrapositional instances are arranged on the left side of the plot and the rest of the cases are to the right. The items are shaded according to the hit counts produced by the corresponding original expressions. If a query returns insufficient results, the corresponding item is shown as a hollow unshaded circle on the top of the plot.

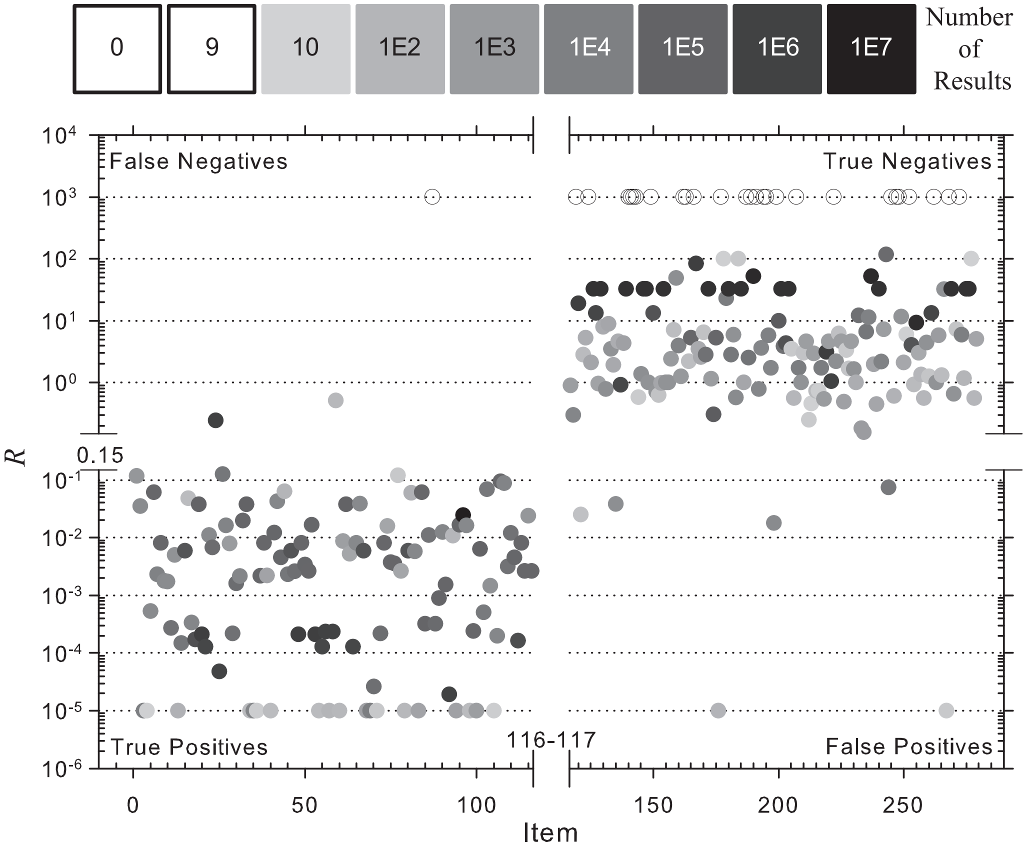

Figure 5.8: A scatter plot illustrating the correlation between R (synthesized expletiveness; it takes the value of r if the more complex queries produce enough results, and takes the value of r' when they fail to do so) and the expletiveness of the *it* instance. The extrapositional instances are arranged on the left side of the plot and the rest of the cases are to the right. The items are shaded according to the hit counts produced by the corresponding original expressions. If a query returns insufficient results, the corresponding item is shown as a hollow unshaded circle on the top of the plot.

5.6 Extended Evaluation

In order to evaluate how well the system generalizes, 500 additional sample sentences are randomly selected from the rest of the WSJ corpus as the test dataset. The distribution of instances is comparable to that of the development dataset, as shown in Table 5.17.

Category	**Instances**	**Percentage**
Nominal Antecedent	375	75.00%
Clause Antecedent	24	4.80%
Extraposition	63	12.60%
Cleft	8	1.60%
Weather/Time	6	1.20%
Idiom	11	2.20%
Other	13	2.60%
Grand Total	**500**	**100.00%**

Table 5.17: Profile of the test dataset according to the author's annotation

As shown in Table 5.18, the overall level of inter-annotator agreement is slightly higher than that of the development dataset. Except for the idiom category, categorical κ values are also higher than their counterparts on the development dataset. This discrepancy is most likely due to chance, since the two volunteers worked independently and started from different datasets (Volunteer 1 started from the development dataset and Volunteer 2 started from the test dataset).

Category	**Volunteer 1**			**Volunteer 2**			κ^a
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	
Referential	98.48%	95.12%	96.77%	97.30%	96.83%	97.07%	.797
Extraposition	87.10%	85.71%	86.40%	80.00%	82.54%	81.25%	.811
Cleft	29.41%	62.50%	40.00%	57.14%	50.00%	53.33%	.490
Weather/Time	100.00%	50.00%	66.67%	100.00%	50.00%	66.67%	.665
Idiom	31.82%	53.85%	40.00%	47.06%	61.54%	53.33%	.280
Overall Accuracy/κ		**91.80%**			**92.80%**		**.720**

[a] All κ values are statistically significant at $p < 0.0001$.

Table 5.18: Performance of the volunteer annotators on the test dataset (evaluated using the author's annotation as reference) and the degree of inter-annotator agreement measured by Cohen's kappa (κ). The author's annotations are refitted to the simplified annotation scheme used by the volunteers.

	WSJ Annotation		**Replicated PHA**
Measurement	**Extraposition**	**Cleft**	**Overall[a]**
Reference	63	8	77
Identified by Baseline	54	6	97
Baseline True Positives	52	6	55
Precision	96.30%	100.00%	56.70%
Recall	82.54%	75.00%	71.43%
F-measure	88.89%	85.71%	63.22%

[a] Includes clefts, extrapositions, and time/weather cases.

Table 5.19: Performance of the baselines on the test dataset, evaluated against the author's annotation.

Table 5.19 summarizes the performance of the baselines on the test dataset. The two $(54-52)$ false-positive extrapositions from the WSJ annotation are listed below together with their respective context:

(5.48) Another solution cities might consider is giving special priority to police patrols of small-business areas. For cities losing business to suburban shopping centers, *it* may be a wise business investment to help keep those jobs and sales taxes within city limits.

WSJ 1450:54-55

(5.49) You think you can go out and turn things around. *It*'s a tough thing when you can't.

WSJ 1996:61-62

The first case is considered referential, and the *it* in the second case is believed to refer to a hypothetical situation introduced by the *when*-clause.

5.6.1 Performance Analysis

On the test dataset, the system is able to maintain its precision; it exhibits slight deterioration in recall but the overall performance is still within expectations. The findings are summarized in Table 5.20.

Measurement	**Extraposition**	**Cleft**	**Weather/Time**	**Overall**[a]
Reference	63	8	6	77
Identified	60	6	7	73
True Positives	58	6	6	70
Precision	96.67%	100.00%	85.71%	95.89%
95% C.I.[b]	91.38-100.00%	64.26-100%	50.00-100.00%	90.77-100.00%
Recall	92.06%	75.00%	100.00%	90.91%
95% C.I.[b]	84.85-98.25%	40.00-100.00%	64.26-100.00%	84.15-97.01%
F-measure	94.31%	85.71%	92.31%	93.33%
95% C.I.	89.60-98.11%	57.14-100.00%	66.67-100.00%	88.75-97.10%

[a] Combining extraposition, cleft, and weather/time into one category.
[b] Adjusted Wald intervals are reported for extreme measurements.

Table 5.20: Performance of the system on the test dataset, evaluated using the author's annotation as reference.

149 instances were evaluated for extraposition using queries, covering 62 of the 63 extrapositions. The excluded case is introduced in the form of a direct question, whose particulars the syntactic processing subsystem is not prepared for. Of the other four false negatives, three involve noun phrases at the matrix object position. One of the two clefts that are not recognized arises out of imperfect processing in the corpus. In addition, the false positive in the weather/time category is caused by the verb '*hail*', which was treated as a noun by the system.

All five $(63-58)$ false-negative extraposition cases are annotated in the corpus and the WSJ annotation agrees with the six clefts identified by the system. Thus the system's recall ratio on the verified WSJ annotations is 90.38% for extraposition and 100% for cleft.

Target System	**Extraposition**	**Cleft**	**Weather/Time**
Volunteer 1	F-measure$^{+}/p = .041$	F-measure$^{+}/p = .005$	F-measure$^{+}/p = .248$
Volunteer 2	F-measure$^{+}/p = .002$	F-measure$^{+}/p = .119$	F-measure$^{+}/p = .254$
WSJ Annotation	Precision$^{-}/p = .697$	F-measure$^{=}/p = 1.00$	
Replicated PHA	(All Categories) Precision$^{+}/p < .001$		

Table 5.21: Results of the statistical significance tests, presented in the format Test Statistic$^{sign}/p$-value. A plus sign ($^{+}$) indicates that our system performs better on the reported measurement; an equal sign ($^{=}$) indicates a tie; otherwise a minus sign ($^{-}$) is used. If fair comparisons can be made for both precision and recall, the F-measure is used as the test statistic; otherwise the applicable measurement is reported.

Performance on Charniak Parser Output

Measurement	**Extraposition**	**Cleft**	**Weather/Time**	**Overall**[a]
Reference	63	8	6	77
Identified	58	7	7	72
True Positives	55	6	6	67
Precision	94.83%	85.71%	85.71%	93.06%
95% C.I.	88.24-100.00%	50.00-100.00%	50.00-100.00%	86.36-98.51%
Recall	87.30%	75.00%	100.00%	87.01%
95% C.I.[b]	78.26-95.08%	37.50-100.00%	64.26-100.00%	78.95-94.12%
F-measure	90.91%	80.00%	92.31%	89.93%
95% C.I.	84.75-95.77%	50.00-100.00%	66.67-100.00%	84.30-94.57%

Performance on Berkeley Parser Output

Measurement	**Extraposition**	**Cleft**	**Weather/Time**	**Overall**[a]
Reference	63	8	6	77
Identified	58	5	7	70
True Positives	56	5	6	67
Precision	96.55%	100.00%	85.71%	95.71%
95% C.I.[b]	91.11-100.00%	59.90-100.00%	50.00-100.00%	90.28-100.00%
Recall	88.89%	62.50%	100.00%	87.01%
95% C.I.[b]	80.60-96.23%	25.00-100.00%	64.26-100.00%	79.22-93.90%
F-measure	92.56%	76.92%	92.31%	91.16%
95% C.I.	87.14-96.97%	40.00-100.00%	66.67-100.00%	85.94-95.52%

[a] Combining extraposition, cleft, and weather/time into one category.

[b] Adjusted Wald intervals are reported for extreme measurements.

Table 5.22: Performance of the system on the test dataset using parser-generated output, evaluated using the author's annotation as reference.

Results of the significance tests, summarized in Table 5.21, reveal the following additional information about the system's performance on the test dataset:

- the system's higher performance in recognizing *it*-extrapositions than both volunteers is statistically significant;
- in the extraposition category, the difference between WSJ annotation's (higher) precision and that of the system is not statistically significant; and
- the system outperforms the Paice and Husk (1987) algorithm, and the difference is statistically significant.

Tables 5.22 and 5.23 outline the system's performance on the test dataset when parsers are used. Again, both parsers cause slight deteriorations in system performance. However, such changes are not statistically significant. With either parser used, the system is able to perform as well as the WSJ annotations.

Comparing System Performance On Charniak Parser Output to:

Target System	Extraposition	Cleft	Weather/Time
System w/o Parser	F-measure$^-/p = .125$	F-measure$^-/p = 1.00$	F-measure$^=/p = 1.00$
Volunteer 1	F-measure$^+/p = .298$	F-measure$^+/p = .013$	F-measure$^+/p = .247$
Volunteer 2	F-measure$^+/p = .022$	F-measure$^+/p = .269$	F-measure$^+/p = .246$
WSJ Annotation	Precision$^-/p = .886$	F-measure$^-/p = 1.00$	
Replicated PHA	(All Categories) Precision$^+/p < .001$		

Comparing System Performance On Berkeley Parser Output to:

Target System	Extraposition	Cleft	Weather/Time
System w/o Parser	F-measure$^-/p = .501$	F-measure$^-/p = 1.00$	F-measure$^=/p = 1.00$
Volunteer 1	F-measure$^+/p = .131$	F-measure$^+/p = .035$	F-measure$^+/p = .256$
Volunteer 2	F-measure$^+/p = .009$	F-measure$^+/p = .308$	F-measure$^+/p = .27$
WSJ Annotation	Precision$^-/p = .809$	F-measure$^-/p = 1.00$	
Replicated PHA	(All Categories) Precision$^+/p < .001$		

Table 5.23: Results of the statistical significance tests comparing the system's performance on parser output to that of various other systems, presented in the format Test Statistic$^{sign}/p$-value. A plus sign ($^+$) indicates that the source system performs better on the reported measurement; an equal sign ($^=$) indicates a tie; otherwise a minus sign ($^-$) is used. If fair comparisons can be made for both precision and recall, the F-measure is used as the test statistic; otherwise the applicable measurement is reported.

5.6.2 Estimated System Performance on the Whole Corpus

The relative sparseness of clefts makes it hard to assess the real effectiveness of the approach proposed in this study. To compensate for this, an approximate study is conducted. First, *it* instances in the whole corpus are processed automatically using the approach. The identified cleft instances are then merged with those that are already annotated in the corpus to form an evaluation dataset of 84 sentences, which is subsequently verified manually. 76 instances out of the 84 are considered to be valid cleft constructs by the author. Respective performances of the approach and the WSJ annotation are reported in Table 5.24; the differences are not statistically significant.

System	Total	Identified	Common	Precision	Recall[a]	F-measure[a]
WSJ	76	66	63	95.45%	82.94%	88.73%
			95% C.I.:	89.55-100.00%	74.32-90.79%	82.86-93.79%
Proposed Approach	76	75	70	93.33%	92.11%	92.72%
			95% C.I.:	87.50-98.65%	85.53-97.40%	87.84-96.65%

[a] The reported recall ratios and F-measures are for the synthetic dataset only and cannot be extended to the whole corpus.

Table 5.24: Estimated system performance on *it*-cleft identification over the entire corpus

Three of the false positives produced by the approach are actually extrapositions[19], which is expected (cf. Footnote 11, Page 138). Thus, in a binary classification of pleonastic *it*, items in the cleft category will have higher contributions to the overall precision than they do for their own

[19] This kind of cleft can be separated from extrapositions using an additional pattern that attaches the prepositional phrase to the subordinate verb. However, the number of samples are too few to justify its inclusion in the study.

category. Until the whole corpus is annotated, it is impossible to obtain precise recall figures of either the WSJ annotations or the proposed approach. However, since the rest of the corpus (other than the synthetic dataset) does not contain any true positives for either system and contains the same number of false-negatives for both systems, the system developed in this study will maintain a higher recall ratio than that of the WSJ annotations on the whole corpus.

A similar experiment is conducted for extrapositions using sentences that are already annotated in the corpus. All 656 annotated extrapositional *it* instances are manually verified and 637 (97.10%) of them turn out to be valid cases. The system produced queries for 623 instances and consequently recognized 575 of them, translating into 90.27% (95% C.I. 89.01-93.56%) recall ratio on the verified annotations. Given the fact that on both the development dataset and the test dataset the system yields slightly higher recall on the whole datasets than it does on the subsets identified by WSJ annotations, its performance for extrapositions on the whole WSJ corpus is likely to remain above 90% in recall.

Similar to the situation in the test based on random cases, a large portion of false-positives are contributed by imperfect handling of both surface structures and noun phrases in the matrix object position, particularly in the form of *it takes/took* ... *to* ... From additional experiments, it seems that this particular construct can be addressed with a different pattern, `what/whatever` *`it takes`* `to verb`, which eliminates the noun phrase. Alternatively, the construct could possibly be assumed as extrapositional without issuing queries at all.

5.7 Discussion

In this chapter a novel pleonastic-*it* identification system is proposed. Unlike its precursors, the system classifies extrapositions by submitting queries to the web and analyzing returned results. A set of rules are also proposed for classification of clefts, whose particular manner of composition makes it more difficult to apply the web-based approach. Components of the system are simple and their effectiveness should be independent of the type of text being processed. As shown in the generalization tests, the system maintains its precision while recall degrades by only a small margin when confronted with unfamiliar texts. This is an indication that the general principles behind the system are not over-fitted to the text from which they were derived. Overall, when evaluated on WSJ news articles – which can be considered a ‘difficult’ type of nonfiction – the system is capable of producing results that are on par with or only slightly inferior to that of casually trained humans.

The system’s success has important implications beyond the particular problem of pleonastic-*it* identification. First, it shows that the web can be used to answer linguistic questions that are based upon more than just simplistic semantic relationships. Second, the comparative study is an effective means to get highly accurate results from the web despite the fact that it is noisier than the manually compiled corpora. In addition, the success of the simple guidelines used in identifying clefts may serve as evidence that a speaker’s intention can be heavily reflected by the surface structures of her utterance, in a bid to make it distinguishable from similarly constructed sentences.

Some problems are left unaddressed in the current study, most notably the handling of complex noun phrases and prepositional phrases. Generally speaking, its approach to query instantiation is somewhat crude. To solve the noun-phrase issue, a finer-grained query downgrading is proposed, viz. first to supply the query with the original noun phrase, then the head noun, and finally the adjective that modifies the head noun, if there is one. The effectiveness of this approach is to be determined. As discussed in Section 5.6.2, a special rule can be used for the verb *take*. This, however, may open the door to exception-based processing, which contradicts the principle of the system to provide a unified approach to pleonastic pronoun identification. Overall, much more data and further experiments are needed before the query instantiation procedures can be finalized.

Aside from the two sets of patterns that are currently in use, other information can be used to assess the validity of a possible extraposition. For example, in extrapositions the matrix verbs are much more likely to remain in present tense than past tense, the noun phrases (if any) at the matrix object position are more likely to be indefinite, and the extraposed clauses are generally longer than the matrix verb phrases. A fuzzy-based decision system with multiple input variables could possibly provide significant performance gains.

Although the system is able to yield reasonable performances on the output of either parser tested, both of them introduce additional errors to the final results. On the combined dataset of development and test items, both parsers cause statistically significant deteriorations in performance at a significance level of 0.1 (Charniak parser: p=0.008 for F-measure on extrapositions; p=0.071 for F-measure on clefts). It is possible that incorporating a pattern-based method will compensate for the problems caused by imperfect parsing and further improve recall ratios; however, more data is needed to confirm this.

Another concern is that the syntactic processing component used in the system is limited. This limitation, caused by the designer's lack of exposure to a large variety of different constructs, is essentially different from the problem imposed by the limited number of patterns in some previous systems. Eventually, for the system developed in this study, this limitation can be eliminated. To illustrate, the current design is not able to correctly process sentences like *what difference does it make which I buy*; however, it only takes minor effort to correct this by upgrading the subsystem so that it recognizes pre-posed objects. Each such upgrade, which may be performed manually or even automatically through some machine-learning approaches, solves one or more syntactic problems and moves the system closer to being able to recognize all grammatically valid constructs. In contrast, it will take considerably more effort to patch the rigidly defined rules or to upgrade the word lists before the rule-based systems can achieve comparable performances.

During the writing of this chapter, Google deprecated their SOAP-based search API. This move makes it technically difficult to precisely replicate the results reported in this study since other search engines lack the ability to process alternate expressions (i.e. Word$_A$ OR Word$_B$) embedded within a quoted query. To use a different search engine, the matrix verbs should not be expanded but should instead be converted to their respective third-person singular present form only. Stubs should also be in their simplest form only, as described in earlier sections. From preliminary experiments it also seems possible to replace the combination of *which/who/this/he* with *they* alone, plus some necessary changes to maintain number agreement among the constituents of the queries. These changes may have some negative effects on the final outcome of the system, but they are unlikely to be severe.

Like most other NLP tasks, classifying the usage of *it* is inherently difficult, even for human annotators who already have some knowledge about the problem – it is one thing to speak the language, and another to then clearly explain the rationale behind a specific construct. Although it is widely accepted that an extrapositional *it* is expletive, the line between extrapositional cases and referential ones can sometimes be very thin. This is clearly manifested by the existence of truncated extrapositions (J. Gundel et al., 2005), which obviously have valid referential readings. Similar things can be said about the relationship among all three pleonastic categories as well as idioms. For example, Paice and Husk classify '*it remains to* . . . ' as an idiom while the same construct is classified as an extraposition in our evaluations. Aside from applying the syntactic guidelines proposed in this study, it is assumed during the annotation process that an extraposition should have either a valid non-extraposed reading or a valid what-cleft reading. It is also assumed that a cleft should generate a valid non-clefted reading by joining the clefted constituent directly to the cleft clause without any leading relative pronoun or adverb. In light of the subjective nature of the problem, our annotations are published on the web to better serve the community.

Chapter 6

Conclusion

This thesis largely evolves around four research questions:

1. Can the web offer more for anaphora resolution?
2. Is it possible to overcome the noise problem and obtain highly accurate result from the web, at least for some tasks?
3. How do the fundamental properties of definite descriptions relate to the notion of anaphora?
4. How does this relationship affect non-anaphoricity determination and anaphora resolution?

Regarding the first two questions, we have shown that the web can provide answers to linguistic questions beyond simple semantic relationships with the studies on pleonastic *it* identification and definite description anaphoricity. In both cases a combination of syntactic and semantic knowledge is needed in order to provide the desired answers. Both systems provide highly accurate predictions and at the same time offer high coverage, and the patterns employed by the systems share a number of design considerations in common: (a) exercise comparative study whenever possible, i.e. constructing queries that represent alternative hypotheses and comparing the query results; (b) retain as much original context as possible; and (c) use stubs to filter out spurious matches. The same principles have also been applied to a number of other patterns performing auxiliary tasks, such as identification of the antecedents of NP-* elements and disambiguation of the subjects of *as*-predications.

As to the latter questions, we have proposed the notion of definite noun phrase anaphora as a device to satisfy the informational uniqueness and weak familiarity presuppositions of definiteness (Roberts, 2003). The notion both explains the perceived 'need' for anaphoric interpretation and bridges the gap between the practice of definite description anaphora resolution and the theoretical studies on the essence of definiteness. We have also examined the closely-related, but fundamentally different concept of coreference and pointed out that the coreferential relationship between two semantically unique entities are essentially different from anaphoric relationships and should be treated differently. We have further studied the various uses of definite descriptions observed in the corpus and examined their behaviors with regard to the familiarity and uniqueness presuppositions and developed a new classification scheme for definite descriptions. The insight gained from this analysis have been particularly useful for the design of the anaphoricity detector. It not only allowed us to identify some of the inadequate syntactic patterns proposed in previous studies but also helped us design better patterns for web-based queries.

In addition, we have designed and implemented an anaphora resolution system that uses the web as its primary information source. The web is used for virtually all tasks, ranging from gender/number determination to semantic relationship discovery and sentence structure disambiguation. We have shown that when the system's design explicitly makes space for web-based features, they can bring statistically significant performance gains to an otherwise knowledge-poor (but still highly effective) pronominal anaphora resolution system. We have also shown that the pleonastic *it* identification module offers tangible benefits for an anaphora resolution system – the addition of pleonastic *it* detection alone brings a (statistically significant) 4% performance gain.

While the study provides answers to the original research questions, it also elicits many more. During the development of the pronominal anaphora resolution system, the issue that surfaces most frequently is to decide when does a specific feature becomes relevant. The issue of relevancy is neither associated with the frequency of a feature's utility nor with its general reliability, but rather has to do with external conditions under which a particular factor 'suddenly' becomes more important than others. Taking verb-argument co-occurrence statistics for example, although there are definitive cases demonstrating the importance of its role, it is actually dormant in majority of the situations. There are numerous factors that could potentially influence the process of anaphora resolution, and many of them behave the same way. Despite that we have identified a few specific situations where co-occurrence may become relevant, we have not found a general solution to the larger problem. Due to the sheer number of potential factors (cf. Uryupina, 2007, for a comprehensive survey), it is impossible to handle the task manually. The commonly used machine-learning approaches for anaphora resolution may not be helpful either, as evidenced by the study of Kehler et al. (2004). We believe that a different model is needed to tackle the issue of feature relevance conditions, with one of the possible directions being manually annotating the most significant factor(s) in each case and use a machine-learning algorithm to discover the conditions.

Another potential area for future research is to study the behaviors of the non-anaphoric definite descriptions that also have coreferential antecedents. In Chapter 3, we have identified a few particular situations in which they often occur. This leads us to believe that these definite descriptions are usually used to perform some kind of discourse function and that it may be possible to explore this point for the purpose of coreference resolution.

Finally, we would like to conclude the thesis by mentioning that although the web can be extremely helpful even with the limited tools available today, there is an urgent need for better utilities geared towards natural language processing. In this study we have used both the live web with commercial search engines and the Google N-gram corpus, and feel that neither offers enough power to fully realize the web’s potential.

References

Abbott, B. (2001). Definiteness and identification in english. In N. T. Enikö (Ed.), Pragmatics in 2000: Selected papers from the 7th international pragmatics conference (Vol. 2, pp. 1–15). Antwerp, Belgium: International Pragmatics Association. 62

Agresti, A., & Coull, B. A. (1998, May). Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician, 52(2), 119–126. 152

Altinçay, H. (2006, August). On the independence requirement in dempster-shafer theory for combining classifiers providing statistical evidence. Applied Intelligence, 25(1), 73–90. 90

Asher, N., & Lascarides, A. (1998). Bridging. Journal of Semantics, 15(1), 83–113. 55

Baker, M. C. (2003). Lexical categories : Verbs, nouns and adjectives (No. 102). Cambridge, UK: Cambridge University Press. Paperback. 185

Baldwin, J. (1895). Old greek stories. American Book Company. Available from `http://www.gutenberg.org/etext/11582` (Retrieved Feb.18, 2009) 28

Barbu, C. (2003). Bilingual pronoun resolution: Experiments in english and french. Unpublished doctoral dissertation, School of Humanities, Languages and Social Sciences, University of Wolverhampton, Wolverhampton, UK. 74

Bean, D. L. (2004). Acquisition and application of contextual role knowledge for coreference resolution. Unpublished doctoral dissertation, University of Utah. 4, 18, 19, 90, 97, 101

Bean, D. L., & Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics (ACL99) (pp. 373–380). Morristown, NJ, USA: Association for Computational Linguistics. 21

Bean, D. L., & Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics (HLT-NAACL04) (p. 297-304). 18, 19, 21

Bell, D., Guan, J., & Shapcott, C. (1998, May). Using the dempster-shafer orthogonal sum for reasoning which involves space. Kybernetes: The International Journal of Systems & Cybernetics, 27(5), 511–526. 90

Bergsma, S. (2005a). Automatic acquisition of gender information for anaphora resolution. In Proceedings of the 18th conference of the canadian society for computational studies of intelligence (Canadian AI2005) (pp. 342–353). 73, 75, 76

Bergsma, S. (2005b). Corpus-based learning for pronominal anaphora resolution. Unpublished master's thesis, University of Alberta. 120

Bergsma, S., & Lin, D. (2006). Bootstrapping path-based pronoun resolution. In Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics (pp. 33–40). Morristown, NJ, USA: Association for Computational Linguistics. 19, 20, 98

Berland, M., & Charniak, E. (1999). Finding parts in very large corpora. In Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics (pp. 57–64). 132

Bies, A., Ferguson, M., Katz, K., & MacIntyre, R. (1995, January). Bracketing guidelines for Treebank II style (Tech. Rep. No. MS-CIS-95-06). Department of Computer and Information Science, University of Pennsylvania. 13, 154, 155

Birner, B., & Ward, G. (1994). Uniqueness, familiarity, and the definite article in English. In Proceedings of the annual meeting of the Berkeley linguistic society (pp. 93–102). 29

Bonzi, S., & Liddy, E. (1989). The use of anaphoric resolution for document description in information retrieval. Information Processing & Management, 25(4), 429–441. 1

Boyd, A., Gegg-Harrison, W., & Byron, D. (2005, June). Identifying non-referential *it*: A machine learning approach incorporating linguistically motivated patterns. In Proceedings of the ACL workshop on feature engineering for machine learning in natural language processing (pp. 40–47). Association for Computational Linguistics. 129, 130, 137, 155

Brants, T., & Franz, A. (2006). Web 1T 5-gram version 1. DVD-ROM. (Linguistic Data Consortium catalog LDC2006T13) 69

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In Proceedings of the 25th annual meeting on association for computational linguistics (ACL87) (pp. 155–162). Morristown, NJ, USA: Association for Computational Linguistics. 10, 13

Bunescu, R. (2003). Associative anaphora resolution: A Web-based approach. In Proceedings of the EACL03 workshop on the computational treatment of anaphora (p. 47-52). 24, 69, 112

Byron, D. K. (2004). Resolving pronominal reference to abstract entities. Unpublished doctoral dissertation, University of Rochester Computer Science Department. 189, 190

Campbell, R. (2004). Using linguistic principles to recover empty categories. In Proceedings of the 42nd annual meeting on association for computational linguistics (ACL04) (pp. 645–652). Morristown, NJ, USA: Association for Computational Linguistics. 83

Carbonell, J. G., & Brown, R. D. (1988). Anaphora resolution: a multi-strategy approach. In Proceedings of the 12th conference on computational linguistics (pp. 96–101). Morristown, NJ, USA: Association for Computational Linguistics. 17

Cardie, C. (1992). Corpus-based acquisition of relative pronoun disambiguation heuristics. In Proceedings of the 30th annual meeting on association for computational linguistics (pp. 216–223). Morristown, NJ, USA: Association for Computational Linguistics. 78

Cardie, C., & Wagstaff, K. (1999). Noun phrase coreference as clustering. In Proceedings of the joint SIGDAT conference on empirical methods in NLP and very large corpora (pp. 82–89). 16

Carlson, G., & Pelletier, F. J. (Eds.). (1995). The generic book. Chicago, USA: University of Chicago Press. 38

Carlson, G. N. (1977a). Reference to kinds in English. Unpublished doctoral dissertation, University of Massachusetts Amherst. 38

Carlson, G. N. (1977b, Jan). A unified analysis of the english bare plural. Linguistics and Philosophy, 1(3), 413–457. 38

Carter, D. (1987). Interpreting anaphors in natural language texts. New York, USA: Halsted Press. 13, 26, 28

Charniak, E., & Johnson, M. (2005). Coarse-to-fine *n*-best parsing and maxent discriminative reranking. In Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL05) (pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics. 158

Chinchor, N. (1992). The statistical significance of the MUC-4 results. In Proceedings of the 4th conference on message understanding (MUC4) (pp. 30–50). San Mateo, CA: Morgan Kaufmann. 152

Chomsky, N. (1993). Lectures on government and binding: the Pisa lectures (7th ed.). Berlin, Germany: Mouton de Gruyter. 10

Christophersen, P. (1939). The articles: A study of their theory and use in English. Copenhagen: Munksgaard. 42, 43

Cimiano, P., Schmidt-Thieme, L., Pivk, A., & Staab, S. (2005). Learning taxonomic relations from heterogeneous evidence. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), Ontology learning from text: Methods, applications and evaluation (pp. 59–73). Amsterdam: IOS Press. 132

Clark, H. (1977). Bridging. In P. N. Johnson-Laird & P. C. Wason (Eds.), Thinking: Readings in cognitive science (p. 411-420). London, UK: Cambridge University Press. 190

Clark, H. H. (1975). Bridging. In Proceedings of the 1975 workshop on theoretical issues in natural language processing (TINLAP75) (pp. 169–174). Morristown, NJ, USA: Association for Computational Linguistics. 48, 56, 191

Clemente, J. C., Torisawa, K., & Satou, K. (2004). Improving the identification of non-anaphoric *it* using support vector machines. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP04). 130

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. 153

Collins, M. (1999). Head-driven statistical models for natural language parsing. Unpublished doctoral dissertation, University of Pennsylvania. 70

Dagan, I., & Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In Proceedings of the 13th conference on computational linguistics (pp. 330–332). Morristown, NJ, USA: Association for Computational Linguistics. 17, 18, 19, 97

Dagan, I., Justeson, J. S., Lappin, S., Leass, H. J., & Ribak, A. (1995, November). Syntax and lexical statistics in anaphora. Applied Artificial Intelligence, 9(6), 633-644. 17

Dahl, O. (1995). The marking of the episodic/generic distinction in tense-aspect systems. In G. Carlson & F. J. Pelletier (Eds.), The generic book (pp. 412–425). Chicago, USA: University of Chicago Press. 39

Davidse, K. (2000, December). A constructional approach to clefts. Linguistics, 38(6), 1101–1131. 137

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge, UK: Cambridge University Press. Paperback. 152

Deemter, K. van, & Kibble, R. (2000). On coreferring: coreference in muc and related annotation schemes. Computational Linguistics, 26(4), 629–637. 25, 26, 38, 39, 40, 190

Demonyms. (2009, Aug. 23). List of adjectival and demonymic forms of place names. In Wikipedia, the free encyclopedia. Wikimedia Foundation, Inc. Available from `http://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_of_place_names` (Retrieved Aug. 30, 2009) 104

Dempster, A. P. (2008, June). The Dempster–Shafer calculus for statisticians. International Journal of Approximate Reasoning, 48(2), 365–377. 86, 88

Denber, M. (1998). Automatic resolution of anaphora in English (Tech. Rep.). Eastman Kodak Co. 129

Denis, P. (2007). New learning models for robust reference resolution. Unpublished doctoral dissertation, University of Texas at Austin. 26

Denis, P., & Baldridge, J. (2007). Joint determination of anaphoricity and coreference resolution using integer programming. In Proceedings of the annual conference of the north american chapter of the association for computational linguistics (NAACL-HLT07) (pp. 236–243). Association for Computational Linguistics. 21

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: a second look. Computational Linguistics, 30(1), 95–101. 154

Doherty, M. (2001). Cleft-like sentences. Linguistics, 39(3), 607–638. 63

Donnellan, K. S. (1966, July). Reference and definite descriptions. The Philosophical Review, 75(3), 281–304. 43

Dowty, D. R., Wall, R. E., & Peters, S. (1981). Introduction to Montague semantics. Dordrecht, The Netherlands: Kluwer. 40

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. New York, USA: Chapman and Hall. 152

Evans, R. (2000). A comparison of rule-based and machine learning methods for identifying non-nominal *it*. In D. Christodoulakis (Ed.), Proceedings of the 2nd international conference on natural language processing (NLP00) (Vol. 1835, pp. 233–241). Berlin: Springer. 129

Evans, R. (2001, April). Applying machine learning toward an automatic classification of *it*. Literary and Linguistic Computing, 16(1), 45–57. 128, 129, 130

Fan, J., Barker, K., & Porter, B. (2005). Indirect anaphora resolution as semantic path search. In Proceedings of the 3rd international conference on knowledge capture (K-CAP05) (pp. 153–160). New York, NY, USA: ACM. 24

Fellbaum, C. (Ed.). (1998). Wordnet: An electronic lexical database. Cambridge, Mass., USA: The MIT Press. 129

Feynman, R., & Gilbert, D. (1960, February). There's plenty of room at the bottom. Engineering & Science, 23(5), 22–36. 76

Fintel, K. von. (2008). What is presupposition accommodation, again? Philosophical Perspectives, 22(1), 137–170. 30

Fraurud, K. (1990). Definiteness and the processing of noun phrases in natural discourse. Journal of Semantics, 7(4), 395-433. 25, 38, 43, 44, 57, 58, 100

Gabbard, R., Marcus, M., & Kulick, S. (2006). Fully parsing the penn treebank. In Proceedings of the main conference on human language technology conference of the north american chapter of the association of computational linguistics (pp. 184–191). Morristown, NJ, USA: Association for Computational Linguistics. 83

Garera, N., & Yarowsky, D. (2006). Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In Proceedings of the 10th conference on computational natural language learning (CoNLL06) (pp. 37–44). 112

Gauker, C. (1998, August). What is a context of utterance? Philosophical Studies, 91(2), 149–172. 30

Ge, N., Hale, J., & Charniak, E. (1998). A statistical approach to anaphora resolution. In Proceedings of the sixth workshop on very large corpora (pp. 161–171). 13, 72

Geurts, B., & Sandt, R. van der. (2004). Interpreting focus. Theoretical Linguistics, 30(1), 1-44. 141

González, J. L. V., & Rodríguez, A. F. (2000). Applying anaphora resolution to question answering and information retrieval systems. In Proceedings of the first international conference on web-age information management (WAIM00) (Vol. 1846, pp. 344–355). London, UK: Springer-Verlag. 1

Green, P. S., & Hecht, K. (1992, June). Implicit and explicit grammar: An empirical study. Applied Linguistics, 13(2), 168–184. 157

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), Universals of language (p. 73-113). Cambridge, USA: MIT Press. 1

Grinder, J., & Postal, P. M. (1971). Missing antecedents. Linguistic Inquiry, 2(3), 269–312. 191

Grishman, R., Hirschman, L., & Nhan, N. T. (1986). Discovery procedures for sublanguage selectional patterns: Initial experiments. Computational Linguistics, 12(3), 205-215. 17

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In Proceedings of the 21st annual meeting of the association for computational linguistics (pp. 44–50). Morristown, NJ, USA: Association for Computational Linguistics. 11

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21(2), 203–225. 10, 11

Guan, J., & Bell, D. (1993). Generalizing the Dempster-Shafer rule of combination to booleanalgebras. In Proceedings of the ieee international conference on developing and managing intelligent system projects (pp. 229–236). 90

Gundel, J., Hedberg, N., & Zacharski, R. (2005). Pronouns without NP antecedents: How do we know when a pronoun is referential? In A. Branco, T. McEnery, & R. Mitkov (Eds.), Anaphora processing: Linguistic, cognitive and computational modelling (pp. 351–364). Amsterdam, The Netherlands: John Benjamins. 126, 127, 128, 171

Gundel, J. K. (1977, September). Where do cleft sentences come from? Language, 53(3), 543–559. 63, 127, 141

Haghighi, A., & Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 848–855). Association for Computational Linguistics. 20

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in english (R. Quirk, Ed.). London, UK: Longman. 26

Hamawand, Z. (2003). *For-to* complement clauses in English: A cognitive grammar analysis. Studia Linguistica, 57(3), 171–192. 140

Hawkins, J. A. (1978). Definiteness and indefiniteness: a study in reference and grammaticality prediction. London, UK: Croom Helm. 28, 29, 30, 42, 43, 44, 45, 46, 47, 49, 50, 57, 64, 101, 190

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th international conference on computational linguistics (pp. 539–545). 132

Hedberg, N. (1990). The discourse function of cleft sentences in english. Unpublished doctoral dissertation, University of Minnesota. 136

Hedberg, N. (2000, December). The referential status of clefts. Language, 76(4), 891–920. 63, 127, 137

Heim, I. (1982). The semantics of definite and indefinite noun phrases. Unpublished doctoral dissertation, University of Massachusetts. 29, 40

Henke, C., Schmoll, C., & Zseby, T. (2008). Empirical evaluation of hash functions for multipoint measurements. SIGCOMM Comput. Commun. Rev., 38(3), 39–50. 77

Hirschman, L., & Chinchor, N. (1997). MUC-7 coreference task definition. In Proceedings of the 7th message understanding conference (MUC-7). Available from http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html 25, 39

Hirst, G. (1981). Anaphora in natural language understanding: A survey (Vol. 119). Berlin, Germany: Springer-Verlag. 1, 6, 26, 27, 28, 32, 185

Hobbs, J. R. (1976, August). Pronoun resolution (Research Report No. 76-1). Department of Computer Sciences, City College, City University of New York. 12, 93

Hobbs, J. R. (1978, April). Resolving pronoun references. Lingua, 44(4), 311–338. 5, 6, 8, 12, 13, 14, 15, 72, 92, 93, 94, 96, 121

Hurford, J. R., Heasley, B., & Smith, M. B. (2007). Semantics: A coursebook (2nd ed.). Cambridge, UK: Cambridge University Press. 54

Jäger, G. (2003, Oct.). Towards an explanation of copula effects. Linguistics and Philosophy, 26(5), 557–593. 78

Jespersen, O. (1949). A modern English grammar on historical principles (Vol. 7). Copenhagen: Munksgaard. 42

Kabadjov, M. A., Poesio, M., & Steinberger, J. (2005). Task-based evaluation of anaphora resolution: the case of summarization. In Proceedings of the conference on recent advances in natural language processing (RANLP05). 1

Kaltenböck, G. (2005). *It*-extraposition in English: A functional view. International Journal of Corpus Linguistics, 10(2), 119–159. 134, 136, 141

Kameyama, M. (1998). Intrasentential centering: a case study. In M. Walker, E. Prince, & A. Joshi (Eds.), Centering theory in discourse (p. 89-112). Oxford, UK: Oxford University Press. 11

Karttunen, L. (1969). Pronouns and variables. In Prodeedings of the fifth regional meeting of the chicago linguistic society (CLS69) (pp. 108–116). University of Chicago. 191

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. Linguistic Inquiry, 8(1), 63–99. 11

Kehler, A., Appelt, D., Taylor, L., & Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In D. M. Susan Dumais & S. Roukos (Eds.), Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics (HLT-NAACL04) (pp. 289–296). Association for Computational Linguistics. 18, 98, 120, 173

Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: pronominal anaphora resoluation without a parser. In Proceedings of the 16th conference on computational linguistics (pp. 113–118). Morristown, NJ, USA: Association for Computational Linguistics. 15

Kilgarriff, A. (2007, March). Googleology is bad science. Computational Linguistics, 33(1), 147–151. 69

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. Computational Linguistics, 29(3), 333–347. 68

Krahmer, E., & Piwek, P. (2000, Aug.). Introduction: Varieties of anaphora (Tech. Rep. No. ITRI-

00-13). University of Brighton. Available from ftp://ftp.itri.bton.ac.uk/reports/ITRI-00-13.pdf 26

Krifka, M. (2003). Bare NPs: Kind-referring, indefinites, both, or neither? In Proceedings of semantics and linguistic theory (salt) xiii. New York, USA: CLC Publications. 137

Krippendorff, K. (1980). Content analysis: An introduction to methodology. Beverly Hills, USA: Sage Publications, Inc. 154

Lambrecht, K. (2001, June). A framework for the analysis of cleft constructions. Linguistics, 39(3), 463–516. 137

Lappin, S., & Leass, H. J. (1994, December). An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4), 535–561. 14, 16, 17, 129

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions. Soviet Physics - Doklady, 10, 707–710. 80

Levin, M. (1987). "the greatest health protection in cigarette history!". APF Reporter, 10(5). Available from http://www.aliciapatterson.org/APF1005/Levin/Levin.html (Retrieved Feb. 13, 2009) 41

Liu, W., & Hong, J. (2000, Jun.). Re-investigating Dempster's idea on evidence combination. Knowledge and Information Systems, 2(2), 223–241. 90

Llombart-Huesca, A. (2002, April). Anaphoric *One* and NP-ellipsis. Studia Linguistica, 56(1), 59-89. 187

Löbner, S. (1985). Definites. Journal of Semantics, 4(4), 279–326. 29, 36, 37, 42, 43, 46, 47, 49, 50, 53, 56, 61, 62

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993, June). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19(2), 313–330. 1

Markert, K., & Nissim, M. (2005, September). Comparing knowledge sources for nominal anaphora resolution. Computational Linguistics, 31(3), 367–402. 68, 112, 113

Markert, K., Nissim, M., & Modjeska, N. N. (2003). Using the web for nominal anaphora resolution. In R. Dale, K. van Deemter, & R. Mitkov (Eds.), Proceedings of the EACL workshop on the computational treatment of anaphora (pp. 39–46). 132

Merlo, P., & Ferrer, E. E. (2006). The notion of argument in prepositional phrase attachment. Computational Linguistics, 32(3), 341–378. 13, 96

Metcalf, A., & Barnhart, D. K. (1999). America in so many words: Words that have shaped America. Boston, USA: Houghton Mifflin. 132

Meyers, A., Reeves, R., Macleod, C., Szekeley, R., Zielinska, V., & Young, B. (2004). The cross-breeding of dictionaries. In Proceedings of the 4th international conference on language resources and evaluation (LREC04). 103, 114

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics (pp. 869–875). Morristown, NJ, USA: Association for Computational Linguistics. 15

Mitkov, R. (1999). Introduction: Special issue on anaphora resolution in machine translation and multilingual nlp. Machine Translation, 14(3-4), 159-161. 1

Mitkov, R. (2001, February). Outstanding issues in anaphora resolution. In A. Gelbukh (Ed.), Proceedings of the 2nd international conference on computational linguistics and intelligent text processing (CICLing01) (Vol. 2004, pp. 110–125). Berlin: Springer. 128

Mitkov, R. (2002). Anaphora resolution. London, UK: Longman. 2, 3, 8, 26, 71, 97, 185, 188

Mitkov, R., Evans, R., & Orasan, C. (2002). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In A. F. Gelbukh (Ed.), Proceedings of the 3rd international conference on computational linguistics and intelligent text processing (CICLing02) (Vol. 2276, pp. 168–186). London, UK: Springer-Verlag. 15, 130

Mitkov, R., & Hallett, C. (2007). Comparing pronoun resolution algorithms. Computational Intelligence, 23(2), 262–297. 12, 13, 14, 92, 93

Müller, C. (2006, April). Automatic detection of nonreferential *it* in spoken multi-party dialog. In Proceedings of the 11th conference of the european chapter of the association for computational linguistics (EACL06) (pp. 49–56). 130

Munafo, R. (2006). Survey of floating-point formats. Available from `http://www.mrob.com/pub/math/floatformats.html` (Retrieved Jul. 5, 2009) 77

Nanni, D. L. (1980, September). On the surface syntax of constructions with *easy*-type adjectives. Language, 56(3), 568–581. 144

Navarro, G. (2001, March). A guided tour to approximate string matching. ACM Computing Surveys, 33(1), 31–88. 81

Ng, V. (2003, December). Machine learning for coreference resolution: Recent successes and future challenges (Tech. Rep. No. TR2003-1918). Cornell University. 8

Ng, V. (2004). Learning noun phrase anaphoricity to improve coreference resolution: issues in representation and optimization. In Proceedings of the 42nd annual meeting on association for computational linguistics (ACL04) (pp. 152–159). Morristown, NJ, USA: Association for Computational Linguistics. 22

Ng, V. (2007, June). Semantic class induction and coreference resolution. In Proceedings of the 45th annual meeting of the association of computational linguistics (ACL07) (pp. 536–543). Morristown, NJ, USA: Association for Computational Linguistics. 16

Ng, V. (2008). Unsupervised models for coreference resolution. In Proceedings of the conference on empirical methods in natural language processing (EMNLP08) (pp. 640–649). Association for Computational Linguistics. 20

Ng, V., & Cardie, C. (2002a). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In Proceedings of the 19th international conference on computational linguistics (COLING02) (pp. 1–7). Morristown, NJ, USA: Association for Computational Linguistics. 22, 100, 130

Ng, V., & Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. In Proceedings of the 40th annual meeting on association for computational linguistics (ACL02) (pp. 104–111). Morristown, NJ, USA: Association for Computational Linguistics. 16

Nishida, K. (2007). Definiteness, indefiniteness, and anaphoric relations in english. In Proceedings of the annual conference of the poetics and linguistics association (PALA07). 38

Noreen, E. W. (1989). Computer-intensive methods for testing hypotheses : An introduction. New York, USA: Wiley-Interscience. 118, 152

Paice, C. D., & Husk, G. D. (1987, June). Towards the automatic recognition of anaphoric features in english text: the impersonal pronoun "it". Computer Speech & Language, 2(2), 109–132. 129, 131, 134, 135, 138, 154, 156, 160, 167

Pantel, P., & Ravichandran, D. (2004). Automatically labeling semantic classes. In Proceedings of human language technology/north american chapter of the association for computational linguistics (HLT/NAACL-04) (pp. 321–328). Boston, MA, USA. 72

Partee, B. (1987). Noun phrase interpretation and type-shifting principles. In J. Groenendijk, D. de Jong, & M. Stokhof (Eds.), Studies in Discourse Representation Theory and the theory of generalized quantifiers (pp. 115–143). Dordrecht, The Netherlands: Foris Publications. (Reprinted in P. Portner and B. Partee (Eds.), Formal Semantics: The Essential Readings (p. 357–381). Oxford:Blackwell Publishing.) 40

Pearson, P. K. (1990, June). Fast hashing of variable-length text strings. Communications of the ACM, 33(6), 677–680. 76, 77

Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006, July). Learning accurate, compact, and interpretable tree annotation. In Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the ACL (ACL06) (pp. 433–440). Morristown, NJ, USA: Association for Computational Linguistics. 158

Poesio, M. (2003). Associative descriptions and salience: A preliminary investigation. In Proceedings of the EACL 2003 workshop on the computational treatment of anaphora (p. 31-28). 23

Poesio, M., Alexandrov-Ksbadjov, M., Vieira, R., Goulart, R., & Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In Sixth international workshop on computation semantics (IWCS6). 22, 100, 105

Poesio, M., Ishikawa, T., Walde, S. S. im, & Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In Proceedings of the third international conference on language resources and evaluation (pp. 1220–1224). 23, 132

Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J. (2004). Learning to resolve bridging references. In Proceedings of the 42nd meeting of the association for computational linguistics (ACL04) (pp. 143–150). 23

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004a). Centering: A parametric theory and its instantiations. Computational Linguistics, 30(3), 309-363. 11

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004b, May). Centering: A parametric theory and its instantiations (Tech. Rep. No. CSM-369). University of Essex. 11, 37

Poesio, M., Uryupina, O., Vieira, R., Alexandrovkabajov, M., & Goulart, R. (2004). Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In Proceedings of the ACL04 workshop on reference resolution and its applications (pp. 47–54). 100, 101, 105

Poesio, M., & Vieira, R. (1998). A corpus-based investigation of definite description use. Computational Linguistics, 24(2), 183–216. 5, 20, 21, 29, 31, 42, 43, 44, 45, 56, 57, 58, 64, 100, 122, 124

Poesio, M., Vieira, R., & Teufel, S. (1997). Resolving bridging references in unrestricted text. In R. Mitkov & B. Boguraev (Eds.), Proceedings of the ACL workshop on operational factors in practical robust anaphora resolution for unrestricted texts (pp. 1–6). 21, 23

Poesio, M., Walde, S. S. im, & Brew, C. (1998). Lexical clustering and definite description interpretation. In J. Choi & N. Green (Eds.), Proceedings of the AAAI spring symposium on learning for discourse (pp. 82–89). Stanford, USA. 23, 24

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In Proceedings of the main conference on human language technology conference of the north american chapter of the association of computational linguistics (pp. 192–199). Morristown, NJ, USA: Association for Computational Linguistics. 16

Poon, H., & Domingos, P. (2008). Joint unsupervised coreference resolution with markov logic. In Proceedings of the conference on empirical methods in natural language processing (EMNLP08) (pp. 650–659). Association for Computational Linguistics. 16, 20, 21

Preiss, J., Gasperin, C., & Briscoe, T. (2004). Can anaphoric definite descriptions be replaced by pronouns? In Proceedings of the fourth international conference on language resources and evaluation (LREC04) (pp. 1499–1502). 34

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), Radical pragmatics (Vol. 14, pp. 223–255). New York: Academic Press. 29, 30, 31, 42, 44, 46

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In W. C. Mann & S. A. Thompson (Eds.), Discourse description: Diverse linguistic analyses of a fund-raising text (pp. 295–325). Amsterdam: John Benjamins. 29, 42, 44

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A comprehensive grammar of the english language. London, UK: Longman. 52, 53, 82, 93, 103

Reinhart, T. (1981). Definite NP anaphora and c-command domains. Linguistic Inquiry, 12(4), 605–635. 10

Rich, E., & LuperFoy, S. (1988). An architecture for anaphora resolution. In Proceedings of the second applied natural language processing conference (ANLP88) (p. 18-24). 17

Richardson, M., & Domingos, P. (2006, February). Markov logic networks. Machine Learning, 62(1-2), 107–136. 16

Rijsbergen, C. J. van. (1979). Information retrieval (2nd ed.). Newton, MA, USA: Butterworth-Heinemann. 118

Riloff, E. (1996, August). An empirical study of automated dictionary construction for information extraction in three domains. Artificial Intelligence, 85(1-2), 101–134. 19

Roberts, C. (2003, June). Uniqueness in definite noun phrases. Linguistics and Philosophy, 26(3), 287–350. 8, 6, 25, 26, 29, 30, 32, 33, 34, 35, 36, 40, 41, 47, 48, 49, 62, 172

Roberts, C. (2004, Sep.). Pronouns as definites. In M. Reimer & A. Bezuidenhout (Eds.), Descriptions and beyond (pp. 503–543). Oxford, UK: Oxford University Press. 29, 32, 40

Russell, B. (1905, October). On denoting. Mind, 14(56), 479–493. 29

Sanchez-Graillet, O., Poesio, M., Kabadjov, M., & Tesar, R. (2006). What kind of problems do protein interactions raise for anaphora resolution? - a preliminary analysis. In S. Ananiadou & J. Fluck (Eds.), Proceedings of the second international symposium on semantic mining in biomedicine (SMBM06). 1

Schumacher, P. B. (2008). Dependency precedes independence: Online evidence from discourse processing. In A. Benz & P. Kühnlein (Eds.), (Vol. 172, pp. 141–158). Philadelphia, USA: John Benjamins. 55

Seltzer, M. I., & Yigit, O. (1991). A new hashing package for unix. In Proceedings of the Usenix Winter 1991 conference (p. 173-184). Usenix Association. 76

Shafer, G. (1976). A mathematical theory of evidence. Princeton, USA: Princeton University Press. 86, 89

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana, USA: The University of Illinois Press. 27

Siegel, M. E. A. (1994, October). *Such*: Binding and the pro-adjective. Linguistics and Philosophy, 17(5), 481–497. 188

Sinclair, J. (Ed.). (1995). Collins cobuild english grammar. London, U.K.: Harper Collins. 128

Sloat, C. (1969, March). Proper nouns in English. Language, 45(1), 26–30. 137

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4), 521–544. 16, 20, 82

Spinillo, M. (2003). On *such*. English Language and Linguistics, 7(02), 195-210. 188

Strube, M., Rapp, S., & Müller, C. (2002). The influence of minimum edit distance on reference resolution. In Proceedings of the acl conference on empirical methods in natural language processing (EMNLP02) (pp. 312–319). Morristown, NJ, USA: Association for Computational Linguistics. 81, 82

Tetreault, J. R. (1999). Analysis of syntax-based pronoun resolution methods. In Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics (pp. 602–605). Morristown, NJ, USA: Association for Computational Linguistics. 14

Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. Comput. Linguist., 27(4), 507–520. 14, 92

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In Proceedings of the 41st annual meeting on association for computational linguistics (ACL03) (pp. 80–86). Morristown, NJ, USA: Association for Computational Linguistics. 21, 22, 101, 105

Uryupina, O. (2007). Knowledge acquisition for coreference resolution. Unpublished doctoral dissertation, Department of Philosophy, Saarland University. 105, 173

U.S. Census Bureau. (1995, Oct.). Frequently occurring first names and surnames from the 1990 census. Available from `http://www.census.gov/genealogy/names/` (Retrieved May 1, 2009) 74

Ushie, Y. (1986, Dec.). 'corepresentation'–a textual function of the indefinite expression. Text, 6(4), 427–446. 38

Van Deemter, K. (1992). Towards a generalization of anaphora. Journal of Semantics, 9(1), 27-51. 28

Versley, Y. (2007). Using the web to resolve coreferent bridging in german newspaper text. In G. Rehm, A. Witt, & L. Lemnitzer (Eds.), Proceedings of the biannual conference of the society for computational linguistics and language technology (GLDV07) (pp. 253–261). 112, 114

Vieira, R. (1998). Definite description processing in unrestricted text. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK. 4, 5, 20, 25, 44, 46, 58, 59, 63, 64, 65, 66, 101, 103, 104, 109, 110, 111, 112, 124, 190

Vieira, R., Bick, E., Coelho, J., Muller, V., Collovini, S., Souza, J., et al. (2006). Semantic tagging for resolution of indirect anaphora. In Proceedings of the 7th SIGdial workshop on discourse and dialogue (pp. 76–79). 112

Vieira, R., & Poesio, M. (2000). An empirically based system for processing definite descriptions. Computational Linguistics, 26(4), 539-593. 5, 7, 20, 21, 23, 24, 101, 104, 109

Walker, M. A. (1989). Evaluating discourse processing algorithms. In Proceedings of the 27th annual meeting on association for computational linguistics (pp. 251–261). Morristown, NJ, USA: Association for Computational Linguistics. 11, 13

Xia, F., & Palmer, M. (2001). Converting dependency structures to phrase structures. In Proceedings of the first international conference on human language technology research (HLT01) (pp. 1–5). Morristown, NJ, USA: Association for Computational Linguistics. 70

Yang, X., & Su, J. (2007, June). Coreference resolution using semantic relatedness information from automatically discovered patterns. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 528–535). Morristown, NJ, USA: Association for Computational Linguistics. 112

Yang, X., Su, J., & Tan, C. L. (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In Proceedings of the 43rd annual meeting on association for computational linguistics (ACL05) (pp. 165–172). Morristown, NJ, USA: Association for Computational Linguistics. 18, 98

Appendix A

Taxonomy of Anaphora

Anaphora can be organized along a number of different axes, such as the grammatical form of the anaphor, the form of the antecedent, the semantic relationship between the two, and their relative position in the text. Instead of attempting to provide a comprehensive analysis of the subject as Mitkov (2002, Chapter 1) or Hirst (1981) did, this section only aims to offer a quick but well-indexed overview. While a large part of the materials offered here are not dissimilar to Mitkov's (2002) offerings, examples are selected from the same corpus (WSJ) whenever possible[1]. It is hoped that this arrangement will help provide a slightly more coherent overall picture of anaphora.

A.1 Classification of Anaphors

A wide variety of anaphoric expressions are found in day-to-day uses of the English language. The most commonly-seen ones are nominal expressions, and more specifically pronouns. Not surprisingly, they are also the most frequent targets of anaphora resolution research. However, the other two major categories of referring expressions, ellipsis and adverb anaphors, are by no means rare phenomena. Finally, there are some cases where the words *so* and *such* act as pro-adjectives[2].

Nominal Anaphor

- Pronouns
  - Personal Pronouns

    (A.1) On Tuesday, <u>the judge</u> called a news conference to say *he* was quitting effective Dec. 31 to join a San Francisco law firm. WSJ 49:32

    The third-person pronoun *he* refers to the same entity encoded by the expression *the judge*. Most of the times *he* is used in circumstances similar to the example, so are the

[1] Some examples are quoted from related research. This does not rule out the possibility that the corpus contains instances of such phenomena, but rather indicates that they cannot be located with the limited time and resource available to this study.

[2] The appropriateness of the term is under dispute. For example, Baker (2003, page 131) argues that *so* should be labeled as pro-PredP. While his argument against the labels pro-adjective and pro-verb is rather convincing, they are nevertheless used in this study because they are presumably easier to understand.

other third-person pronouns (e.g. *they*), including the possessive (e.g. *their*), objective (e.g. *them*), and reflexive forms (e.g. *themselves*).

(A.2) "*I*'m very alarmed to see these rich valuations," says Smith Barney's Mr. Porter.
WSJ 34:32

First- and second-person pronouns are generally not anaphoric, at least in the sense that the entities they refer to usually do not appear in the text. One of the notable exceptions is that in reported conversations, first-person pronouns (and their respective possessive and reflective forms) tend to be anaphoric. In (A.2), the singular *I* directly corresponds with the speaker, *Mr. Porter*. However, interpretation of the plural forms may be somewhat more complex, as shown in (A.3):

(A.3) "*We* thought it was awfully expensive," said Sterling Pratt, wine director at Schaefer's in Skokie, Ill., one of the top stores in suburban Chicago, "but there are people out there with very different opinions of value. WSJ 71:37

The generic *we* actually refers to *Sterling Pratt* and his colleagues. Since his affiliation, *Schaefer's*, is also mentioned in the text, it is reasonable to infer that the quoted speech largely represents the organization's opinion. Finally, although uncommon, anaphoric uses of the second-person pronouns are still possible:

(A.4) In a letter to the federation, Raymond Campion, Exxon's environmental coordinator, said: "Recent public actions by *you* regarding the Valdez oil spill have failed to demonstrate any sense of objectivity or fairness." WSJ 620:6

One of the phenomena that deserves special attention is the non-anaphoric use of the pronoun *it*, as illustrated in (A.5) and (A.6):

(A.5) But *it* doesn't take much to get burned†. WSJ 34:20

(A.6) And most disturbing, *it* is educators, not students†, who are blamed for much of the wrongdoing. WSJ 44:26

Such non-anaphoric uses of *it*, often referred to as the pleonastic *it*, actually represent a significant part of all *it* instances.

- Relative Pronouns

(A.7) Butch McCarty, *who* sells oil-field equipment for Davis Tool Co., is also busy.
WSJ 725:100

- Demonstrative Pronouns

(A.8) Many auto dealers now let car buyers charge part or all of their purchase on the American Express card, but few card holders realize *this*, Mr. Riese says.
WSJ 116:16

The demonstrative pronoun *this* refers to the preceding clause. Demonstrative pronouns, especially the singular *this* and *that*, can serve to indicate abstract entities represented by full clauses, such as events or statements.

- Proper Nouns

(A.9) Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. *Mr. Vinken* is chairman of Elsevier N.V., the Dutch publishing group. WSJ 1:1-2

Here both *Mr. Vinken* and *Pierre Vinken* refer to the same person.

- Common Noun Phrases

(A.10) South Korea registered a trade deficit of $101 million in October, reflecting *the country*'s economic sluggishness, according to government figures released Wednesday. WSJ 11:1

VP Pro-form Anaphora and Ellipsis

(A.11) Many small investors are facing a double whammy this year: They got hurt by investing in the highly risky junk bond market, and the pain is worse because they *did it* with borrowed money. WSJ 983:1

(A.12) The banks have 28 days to file an appeal against the ruling and are expected to *do so* shortly. WSJ 117:22

The VP pro-forms *do* [*so/it/this* etc.] can be used to refer to preceding clause antecedents.

(A.13) The government includes money spent on residential renovation;
Dodge doesn't __ . WSJ 36:42

(A.14) As a result, consumer prices for the first 10 months of 1989 surged by 5% and wholesale prices __ by 1.3%. WSJ 235:3

(A.15) He had no answers then. Now there are some __ . WSJ 108:61-62

Examples (A.13) through (A.15) illustrate respectively the VP ellipsis, gapping, and the NP ellipsis, which are the most popular types of ellipsis. Following Llombart-Huesca's (2002) analysis, instances of anaphoric *one* can also be put under the framework of ellipsis treatment:

(A.16) Northeast said it would refile its request and still hopes for an expedited review by the FERC so that it could complete the purchase by next summer if its bid is the *one* approved by the bankruptcy court. WSJ 13:17

Adverb Anaphor

(A.17) Not only can they block Wellington from raising money in Japan, bankers here say, but as the largest underwriters in the Eurobond market, they might be able to scuttle borrowings *there*, too. WSJ 210:12

(A.18) The protracted downturn reflects the intensity of Bank of Japan yen-support intervention since June, *when* the U.S. currency temporarily surged above the 150.00 yen level. WSJ 33:3

References to location and time are often made through the use of the pro-adverbs such as *there*, *then*, and their corresponding relative counterparts, *where* and *when*. Similar to the case of the first-person personal pronoun *I*, concepts represented by *here* and *now* usually do not appear in the text. Adverbials such as *this/that way* can also be considered members of this group. However, the resolution of these expressions often demand significant level of reasoning.

Pro-adjectives

(A.19) "It's important to share the risk and even more *so* when the market has already peaked." WSJ 782:8

(A.20) Not only is development of the new company's initial machine tied directly to Mr. Cray, *so* is its balance sheet. WSJ 18:2

The word *so* sometimes appears as a pro-adjective, used in place of an adjectival or prepositional predicate, or an adjectival passive.

(A.21) It also issued a final rule requiring auto makers to equip light trucks and minivans with lap-shoulder belts for rear seats beginning in the 1992 model year. *Such* belts already are required for the vehicles' front seats. WSJ 64:3-4

The usage of *such* is more liberal and its target is not limited to predicates. To date, systematic investigations into the use of *such* are rather limited and there lacks a consensus on its grammatical category. This study follows the analysis of Siegel (1994) and Spinillo (2003).

A.2 Types of Antecedents

As shown in examples (A.1) through (A.18), majority of the antecedents are nominal expressions. Sometimes a nominal anaphor resolves to the combination of multiple noun phrases. Mitkov (2002) recognizes one of such situations as 'coordinated antecedents', which is characterized by overt conjunction markers linking the phrases, such as the *and* connecting *Norman Ricken* and *Frederick Deane Jr.* in (A.22).

(A.22) Norman Ricken, 52 years old and former president and chief operating officer of Toys "R" Us Inc., and Frederick Deane Jr., 63, chairman of Signet Banking Corp., were elected directors of this consumer electronics and appliances retailing chain. *They* succeed Daniel M. Rexinger, retired Circuit City executive vice president, and Robert R. Glauber, U.S. Treasury undersecretary, on the 12-member board. WSJ 14:1-2

The conjunctions are syntactic markers that explicitly suggest possible groups formed by the connected phrases. However, an overt conjunction is not required to be present in order for such grouping to happen. Further, the elements that form a group are not required to appear in the same sentence. Examples (A.23) and (A.24) illustrate this gradual departure from (A.22).

(A.23) Besides Messrs. Cray and Barnum, other senior management at the company includes Neil Davenport, 47, president and chief executive officer; Joseph M. Blanchard, 37, vice president, engineering; Malcolm A. Hammerton, 40, vice president, software; and Douglas R. Wheeland, 45, vice president, hardware.
All __ came from Cray Research. WSJ 18:35-36

(A.24) Mrs. Hills lauded South Korea for creating an intellectual-property task force and special enforcement teams of police officers and prosecutors trained to pursue movie and book pirates. Seoul also has instituted effective search-and-seizure procedures to aid these teams, she said. Taiwan has improved its standing with the U.S. by initialing a bilateral copyright agreement, amending its trademark law and introducing legislation to protect foreign movie producers from unauthorized showings of their films. That measure could compel Taipei's growing number of small video-viewing parlors to pay movie producers for showing their films. Saudi Arabia, for its part, has vowed to enact a copyright law compatible with international standards and to apply the law to computer software as well as to literary works, Mrs. Hills said. *These three countries* aren't completely off the hook, though. WSJ 20:7-12

Anaphora to abstract entities is less frequent in written text than in casual conversations. Nevertheless, they still represent a significant minority of all antecedents. Following Byron's (2004) scheme, the abstract antecedents are classified into the following categories:

- Situation

  (A.25) She became an abortionist accidentally, and continued because *it* enabled her to buy jam, cocoa and other war-rationed goodies. WSJ 39:34

- Event

  (A.26) Sindona, the onetime Vatican financial adviser with reported links to the Mafia, died on March 22, 1986, at age 65, reportedly after drinking cyanide-laced coffee in an Italian prison. *It* happened four days after he was sentenced to life in prison for ordering a 1979 murder. WSJ 1266:13-14

- Action

  (A.27) We are willing to share the political burden of being host to America, an imperial power. We think *it* isn't such a great burden, that it carries no stigma, and we are prepared to do it." WSJ 296:34-35

- Proposition

(A.28) Everyone by now understands that Congress is utterly incapable of writing legislation to help deserving people without its becoming some billion-dollar morass. We have no doubt *this* is one reason judges in New York and justices on the Supreme Court are willing to trash the law in the DES cases. WSJ 130:31-32

While Byron's (2004) approach focuses on pronominal anaphors, it is worth noting that definite descriptions can also refer to abstract entities. Consider the following example:

(A.29) In 1975, Mr. Pamplin enticed Mr. Hahn into joining the company as executive vice president in charge of chemicals; *the move* befuddled many in Georgia-Pacific who didn't believe a university administrator could make the transition to the corporate world.

WSJ 100:19

In (A.29), the definite phrase *the move* refers to an event described in the previous clause.

A.3 Anaphoric Relationships

Currently, there is no consensus on the classification of anaphoric relationships. Most researchers agree that at the highest level, anaphoric relationships can be described with a dichotomy – whether the anaphor and the antecedent refer to the same entity. However, even at this level, there are different views on what should be considered coreferential (cf. Deemter & Kibble, 2000). Previous research (cf. Vieira, 1998 for a survey) produced a number of different schemes for the classification task. This study use categories that are largely based on H. Clark's (1977) analysis[3].

Coreference

Based on Deemter and Kibble's (2000) definition, an anaphor and its antecedents enter into a coreference relationship only when the two refer to the same entity. This equivalence relationship is reflexive, symmetric, transitive, and may not be context-sensitive. In this dissertation, however, the definition is relaxed and the resolution is constrained by the document. Most of the pronominal anaphors and a large part of the definite description anaphors are coreferential. While pronouns generally carry little or no semantic information, the head words of the definite description anaphors usually specify their type information. If the antecedent is also a definite description, the two head words can either be exactly the same or form a hypernym/synonym relationship. In the case that the antecedent is a proper name, the corresponding entity is usually an instance of the type denoted by the anaphor's head word. Example (A.30) illustrates both situations:

[3] The term 'associative anaphora' is adopted from Hawkins (1978), as it is more intuitive than the largely corresponding term of 'bridging' used by H. Clark.

(A.30) Mips Computer Systems Inc.$_1$ today will unveil a new general-purpose computer$_2$ that will compete with more expensive machines from companies such as Sun Microsystems Inc. and Digital Equipment Corp. *The closely held Sunnyvale, Calif., company*$_1$ also will announce an agreement to supply computers to Control Data Corp., which will sell Mips machines under its own label. *The new Mips machine*$_2$, called the RC6280, will cost $150,000 for a basic system. *The computer*$_2$ processes 55 million instructions per second and uses only one central processing chip, unlike many rival machines using several processors. *The machine*$_2$ employs reduced instruction-set computing, or RISC, technology. WSJ 258:1-5

Mips Computer Systems Inc. is an instance of the type *company*, the head word of *The closely held Sunnyvale, Calif., company*. The phrase *a new general-purpose computer* is later referred to as *The new Mips machine*, *The computer*, and *The machine*, using both the same head and a hypernym head alternatingly. One noteworthy fact arising from the example is that noun phrase anaphors are capable of providing additional information that is not present in the anaphor. The supplied information may be already present in the context, as in the case that the new computer is produced by the Mips company, or it may be completely new, as in the case that the company is closely held and is located in Sunnyvale, Calif. Extending further along this line, there are noun phrase anaphors whose type information do not exactly match that of the antecedent along the lines of hypernym/synonym but rather focus on specifying attribute of the antecedent. The best known example of this kind of anaphors is the epithetic 'bastard' example:

(A.31) I met a man yesterday. *The bastard* stole all my money. H. H. Clark (1975, ex. 7)

Identity-of-sense Anaphora

Grinder and Postal (1971) originally coined the notion of 'identity-of-sense' to cover a broad category of phenomena, including the *one* anaphora and many other ellipsis cases. However, in this study, it is reserved strictly for the 'paycheck' cases:

(A.32) The man who gave his paycheck† to his wife was wiser than the man who gave *it* to his mistress. Karttunen (1969, ex. 18)

Obviously, the underlying entity denoted by *it* is a paycheck, but not the same paycheck as the one denoted by *his paycheck*. Definite descriptions can exhibit the same behavior, as demonstrated in this slightly modified version:

(A.32′) The man who gave his paycheck† to his wife was wiser than the man who gave *the paycheck* to his mistress.

Elided nominal contents can still be analyzed under the same framework once they have been successfully recovered from the ellipsis site. [4]

Generic References

Example (A.3) on page 186 illustrates the generic use of *we* to represent a group of persons typified by the speaker. Similarly, the pronoun *they* can be used to refer to a class of individuals.

(A.33) "It's hard to explain to a 17-year-old why someone *they* like had to go," says Mrs. Ward. WSJ 44:124

(A.34) "There's an understanding on the part of the U.S. that Japan has to expand its functions" in Asia, says J. Michael Farren, undersecretary of commerce for trade. "If *they* approach it with a benevolent, altruistic attitude, there will be a net gain for everyone." WSJ 43:37

In example (A.33), the pronoun *they* is used in place of the traditional generic *he* to refer to the group of people introduced by the indefinite *a 17-year-old*. Since the indefinite phrase already establishes the abstract 'kind' information, the pronoun *they* may simply be interpreted as coreferring to the generic group. In (A.34), however, one has to infer the real antecedent, *the Japanese*, from the mention of the country *Japan*.

Associative Anaphora

- Set-member

  (A.35) In filing an original (not amended) return, a couple should consider whether damaged property is owned jointly or separately and whether *one spouse* has larger income; that may determine whether they should file jointly or separately. WSJ 2033:15

- Whole-part

  (A.36) Mrs. Ward took over in 1986, becoming the school's seventh principal in 15 years. Her immediate predecessor suffered a nervous breakdown. Prior to his term, a teacher bled to death in *the halls*, stabbed by a student. WSJ 44:51-53

- Event-participant

  (A.37) In a highly unusual meeting in Sen. DeConcini's office in April 1987, the five senators asked federal regulators to ease up on Lincoln. According to notes taken by one of *the participants* at the meeting, the regulators said Lincoln was gambling dangerously with depositors' federally insured money and was "a ticking time bomb." WSJ 2446:6-7

- Cause-consequence

[4] This approach effectively guarantees that a pronoun inside the elided contents get an sloppy reading, which may not always be the original utterer's intention.

(A.38) San Francisco Bay area officials said nine people remain missing in the aftermath of last week's earthquake. *The death toll* rose to 63. WSJ 1216:9-10

(A.39) Even suburban Prince George's County, Md., reported last week there have been a record 96 killings there this year, most of them drug-related. Innocent bystanders often are *the victims*. WSJ 1847:32-33

- General Association

(A.40) In Thailand, for example, *the government*'s Board of Investment approved $705.6 million of Japanese investment in 1988, 10 times the U.S. investment figure for the year. WSJ 43:6

(A.41) Soon, T-shirts appeared in the corridors that carried the school's familiar red-and-white GHS logo on *the front*. WSJ 44:125