Resolving Pronominal Reference to Abstract Entities
Thesis project proposal
Approved April 26, 1999

Donna K. Byron
University of Rochester Department of Computer Science

Technical Report 714

June 14, 1999

The author wishes to thank James Allen, Len Schubert, Greg Carlson, and Candace Sidner for their help and support on this project. This material is based on work supported by USAF/Rome Labs contract F30602-95-1-0025, ONR grant N00014-95-1-1088, and Columbia Univ. grant OPG:1307.

Contents

Abstract

The ability to compute an interpretation for pronominal referring expressions is an important part of any natural language understanding system. Existing algorithms for pronoun resolution typically cast the problem into a coreference task, which means they simply identify an antecedent noun phrase for each pronoun. Selection of the antecedent is usually based on a calculation of salience or focus. This simplified approach is unable to account for pronouns without noun-phrase antecedents. Examples are abstract referents such as events, propositions, and speech acts that might appear in the linguistic surface form as sentential complements, verbal constructions or entire sentences, and consequences or outcomes that don't appear in the surface form at all. Since abstract entities are never in focus, algorithms that choose referents based solely on calculations of salience cannot judge when to bind a pronoun to an abstract referent. Also, because demonstrative pronouns (this/those) often refer to abstract entities, most existing algorithms completely ignore them.

This proposal first surveys the pronoun resolution techniques reported in the computational linguistics literature and also some studies on pronoun usage patterns from theoretical linguistics. After discussing inadequacies of existing approaches, the proposed thesis project is laid out. We propose a model to resolve pronominal reference to abstract entities by exploiting semantic information from the sentence containing the pronoun. The algorithm also implements a number of linguistic observations on the pragmatics and contrast of definite and demonstrative pronouns. Preliminary experiments indicate that the approach has promise. Many aspects of the model remain to be determined, and the proposed work includes a variety of machine learning experiments to discover an optimal configuration for the pronoun resolution algorithm. By using the appropriate heuristics and semantic filters, abstract entities can be resolved and a net increase in pronoun resolution can be obtained.

The hypotheses to be tested by the proposed model are:

- A domain-independent pronoun resolution method can be built (either in its details or in the method used to learn the details).
- A model containing different strategies for definite and demonstrative pronouns performs better than one with a uniform strategy.
- Separating abstract entities into Implicit and Resultant state entities produces better pronoun resolution than having only one category.
- Pronoun resolution algorithms must support a wide range of abstract entities as referents in order to achieve high resolution accuracy without excluding many pronouns from the experiment.
- For definite noun phrases with noun-phrase antecedents, the model we propose performs at least as well as existing algorithms that do not have abstract entities represented in the discourse context.

1 Introduction

1.1 Overview and Motivation

Pronominal referring expressions abound in naturally occurring discourse, and calculating an accurate interpretation for these expressions is a critical part of natural language understanding. Successful reference resolution requires the interpreter to combine syntactic, semantic, and pragmatic clues as well as arbitrarily detailed world knowledge and inference about the speaker's beliefs [Cha72; CM81]. As a result, pronominal reference resolution is a canonical artificial intelligence problem, requiring the combination of evidence from multiple, possibly conflicting, sources and the application of heuristics to guide the search through possible solutions.

Previous computational approaches to pronominal reference resolution have handled an extremely limited range of phenomena. The accuracy of those models is sometimes reported to be very high (some authors even claim up to 100% accuracy [Mit98]). While these results look impressive on first reading, they become suspect upon a closer examination of the range of phenomena handled by the algorithm. Existing algorithms are restricted to resolving *coreference* among referring expressions. Coreference occurs when multiple surface constituents refer to the same entity, for example '*Dave* and *his* brother' in which the words 'Dave' and 'his' refer to the same entity. When an algorithm considers pronoun resolution from this perspective, it is limited to resolving only the pronouns that corefer with another noun phrase. An entire class of pronouns performing *discourse deixis* [Lak74] - referring to facts, propositions, events, etc. - is left totally untouched. We claim that dialog understanding systems have no hope of achieving highly reliable interpretation without correctly processing this class of pronouns. Although few authors explain their lack of attention to discourse deictic pronouns, we assume they fear the increased number of candidate antecedents that would need to be considered for each pronoun. Hobbs is one of the few authors who explicitly mentions the problem, observing that if sentential antecedents are allowed "the problem of avoiding spurious antecedents would be quite severe" [Hob86, pg. 343].

The need to move beyond coreference resolution has been noted by many authors. Botley [Bot96] noted that 20% of pronouns in his corpus were not coreferential with any noun phrase, and Webber [Web88] found approximately two such pronouns per page of text in her corpus. While some types of discourse have only a small amount of discourse deixis, an example is the MUC7 coreference task [LDC99] with only 6% of pronouns performing discourse deixis, in the dialog corpus used for our experiments, fully 50% of third-person pronouns fall into this category [BA98].

This paper describes proposed research to develop an improved model of pronoun resolution, one that resolves discourse deictic pronouns. Although our model still fails to resolve certain types of pronouns correctly, we believe it will result in a net increase of pronoun resolution accuracy. Preliminary studies have already been conducted to demonstrate the feasibility and power of the approach. A fortunate side effect of the model is that it can resolve both definite (eg. his/she/them) and demonstrative (eg. this/those) pronouns. Demonstrative pronouns have rarely been addressed in the computational linguistics literature, in fact in a recent survey of anaphora resolution techniques, Webber remarks that

"[She is] not aware of any recent attempts to articulate the processes involved in resolving demonstrative pronouns."[Web99]. Because of the improvements introduced in our model and our utilization of linguistic theory, we are able to resolve demonstrative pronouns (almost) for free.

The remainder of this paper is organized as follows: Section 2 provides motivation for the work by surveying some naturally-occurring examples of discourse deixis; Section 3 surveys related work on pronoun resolution models from computational linguistics and observations from theoretical and corpus linguistics; Section 4 outlines our proposed model; Section 5 describes some preliminary experiments conducted to evaluate the utility of the model; finally, Section 6 details the project milestones and schedule.

1.2 Terminology used in this paper

A wide variety of terminology is found in the literature when discussing pronouns and the objects for which they stand proxy. This section establishes the terminology that will be used in this paper.

We assume, along with Webber [Web79], that the speaker (or writer) of a discourse possesses a mental model of some situation. By virtue of his discourse, the speaker attempts to instruct the addressee (either a listener or a reader) to construct a similar model in his own mind. This model contains entities of all sorts and describes their attributes and the relationships between them. As the discourse unfolds, a linguistic record of what has been said exists as part of both the speaker and the addressee's mutual knowledge. Certain constituents in the linguistic record point to entities in the mental model. We will call the mental model being synthesized by the addressee the *discourse model* (DM) and the entities in it *discourse entities* (DE). Karttunen introduced the term *discourse referent* for these objects. His choice of term was motivated to highlight their function as things that can be referred to [Kar76]. We prefer to distinguish between the entity itself and the process of referring. The term *discourse entity* creates a nice distinction between entities that are available for reference by virtue of the discourse from other salient entities - in the physical context, surrounding social context, or the discourse participants' pre-existing knowledge, for example. Discourse entities can be individuals, sets, events, propositions, beliefs, etc.

Noun phrases and pronouns with existential force are often called *referring expressions*. Within a discourse, the initial reference to a particular entity is called the *evoking reference*, and subsequent reference to an entity already in the discourse model is called *anaphoric*. Both anaphoric and evoking referring expressions generate discourse entities. In order to build the proper conceptual representation of the discourse entity for anaphoric expressions, the addressee must determine which previously mentioned entity it relates to and determining the nature of that relationship. The discourse entity is called the *referent* of the referring expression. When a discourse participant utters a referring expression, the expression refers to a discourse entity in his own discourse model. When he uses an anaphoric expression, he expects the listener to have a corresponding entity in his own model and be able to calculate the correct entity as the referent for the anaphoric expression [Sch88]. The linguistic constituent that causes the addressee to construct a DE in his model will be called the *linguistic trigger*. This term is preferable to the traditional term *antecedent*, because it

does not imply that the pronominal reference must come after the trigger in the linguistic record. We reserve the term *antecedent* for simple noun phrase linguistic triggers that occur prior to the pronoun in the discourse. A single linguistic trigger can trigger many discourse entities. Notice that linguistic triggers are not necessarily themselves referring expressions; they can be full sentences, verbs, sentential complements, etc.

The process of referring is seen here as a relationship between linguistic objects and mental representations. Referring expressions *refer* to discourse entities. This terminology differs from that found in philosophy, where referring is described as a relationship between a symbol in a logical system and the real entity in the world. The process of relating the discourse entity of the anaphoric expression with a previous discourse entity is called *resolution*. There are many types of anaphoric reference, but in this study we limit our domain of interest to pronominal anaphors. The anaphoric referring expression itself is a complicated object. We can consider it to be a special type of pointing in which the anaphoric expression points to the linguistic trigger. However, the anaphor does *not* refer to the linguistic trigger [Sid83b]. Anaphors refer to discourse entities. Philosophers of language have acknowledged this phenomenon in other contexts: what is pointed to does not have to be the same as what is referred to [Nun79]. By defining referents to be discourse entities, we bypass the question of how the current conversation fits in with the hearer's prior beliefs and whether it fits with reality. The problem of how linguistic expressions can refer to fictional entities like the golden mountain, the cause for much concern to philosophers, does not trouble us. The link between the cognitive entities and the external world we leave undefined.

This paper uses the following terms for referring expression surface forms. Definite noun phrases include both definite descriptions, which have descriptive content (eg. *the red car*) and pronouns. Pronouns come in many flavors. In this paper we concentrate on third-person pronouns, which are broken down into definite pronouns *he/she/it/them/they* (which can also be cast into possessive pronouns *his/hers/their/theirs*) and demonstrative pronouns *this/that/these/those*. Demonstratives can also be used as determiners, for example *that train*, in which case the form is called a demonstrative noun phrase rather than a demonstrative pronoun. Pronouns in certain constructions, called pleonastic pronouns, do not actually refer to anything. Examples are common expressions such as "When it comes to picking a truck I'd go American", and clefted syntactic constructions with sentential complements, such as "It was good that he told her."

When we speak of computer implementations, we use the term *Discourse Context* (DC) as the name for the data structure the system builds specifically to support anaphoric reference resolution. The DC is not a full discourse model. Elements of the DC are organized differently and have a different lifespan than DM entities. The term *DE token* is used for the data item representing each discourse entity in the DC. DE tokens are representationally impoverished conceptual objects, small proxy tokens that are linked to, but not identical to, full cognitive representations of objects in the discourse model. Although the discourse context can be used in a variety of interpretation processes, we concentrate here on its use as a list of candidate referents in support of pronominal anaphora resolution.

2 Corpus Examples and Observations

The primary motivation for the proposed work is the phenomenon of pronominal reference to what I will call *abstract entities* in naturally-occurring discourse. These entities can be propositions, events, facts, outcomes, actions[1], etc. In many cases these entities are not explicitly present in the surface form of the utterance, but are nonetheless available for subsequent pronominal reference.

2.1 Examples from the TRAINS93 corpus

To demonstrate the phenomenon of pronominal reference to abstract entities, let me start with some examples from naturally-occurring dialog. These examples are taken from the TRAINS93 corpus of task-oriented spoken dialogs, collected at the University of Rochester [HA95][2]. These examples demonstrate the versatility of pronominal referring expressions, as well as the variety of syntactic constituents that can cause an entity to be available for subsequent pronominal reference.

The first example shows a canonical pronoun use - reference to an entity that was introduced into the discourse via a noun phrase.

(1)

Canonical Noun Phrase followed by pronoun [d93-10.4]

utt4:	u:	first we're going to take two engines um both engines
utt5:	s:	okay
utt6:	u:	from Elmira to Corning
utt7:	s:	okay
utt8:	u:	and then to Dansville
utt9:	u:	in Dansville **they** should pick up the three boxcars
		(they = both engines that were picked up in Elmira)

Example 1 shows canonical pronoun behavior, *they* in utterance 9 corefers to the engines previously mentioned in utterance 4 via the definite noun phrase *both engines*.

(2)

Reference to an event type [d92a-1.4]

utt73:	s:	oh let me just check that we don't have two trains
		uh trying to cross each other on the same track
utt74:	u:	okay
utt75:	s:	um but I don't think **that's** happening
		(that = the event type of two trains crossing each other on the same track)

[1]The definitions for these classifications will be formalized in Section 4.

[2]In the examples, numbers like d93-10.4 represent the dialog number, 'uttx' means utterance number x, and 's'/'u' indicate the dialog participants, both humans. 'S' is the 'system' and 'u' is the 'user'.

In Example 2, *that* in utterance 75 refers to the event type (not an instance of the event type) of two trains crossing each other on the same track, which was stated as a sentential complement in utterance 73.

(3)

Reference to an action type [d93-14.2]

utt37:	u:	how long does it take to convert the oranges into orange juice?
utt38:	s:	**it** takes one hour
		(it = the action of turning oranges into orange juice)

In Example 3, *it* in utterance 38 refers to the action type of converting oranges into orange juice, not to a specific instance of performing that action. Converting oranges into orange juice was previously mentioned as an infinitive phrasal complement to the verb 'take' in utterance 37.

(4)

Reference to a property of an action [d92a-1.3]

utt49:	s:	okay so that'll take two hours to get to Corning an hour
		to load the oranges and two hours to get to Bath
utt50:	s:	so **that'll** be another five hours
		(that = the time required to perform the action of going to Corning,
		loading, and going to Bath)

Example 4 demonstrates a reference to an entity that requires slightly more inference to resolve. The correct interpretation of the pronoun *that* in utterance 50 is a property of a mentioned action, specifically the amount of time required to complete the action. This example is further complicated by the fact that several actions are listed, and the list must be consolidated into one big action to get the correct referent of the pronoun. The pronominal reference *that* in utterance 50 refers to the total amount of time to complete all three actions. The syntactic constituent used to introduce the action initially is a sequence of noun phrase complements.

(5)

Reference to a fact [d93-14.2]

utt75:	u:	okay and um we need to pick up a boxcar of bananas in Avon
utt76:	s:	okay um there are boxcars that are closer to Avon if **that** helps any
utt77:	u:	um **it** doesn't really matter but okay ... (that = it = the fact that there are closer boxcars)

In Example 5, an entire proposition is the linguistic trigger for the subsequent pronouns. The entire sentence "There are boxcars that are closer to Avon" is the trigger for *that* in utterance 76, which in turn is the trigger for *it* in utterance 77. We can see in this example both a definite pronoun and a demonstrative used to refer to the same abstract entity. The first pronominalization of the abstract entity takes the form of a demonstrative pronoun and subsequent mentions take the form of definite pronouns. Several previous authors (cf.[Web90; Sch85] have observed this behavior, sometimes called *pronoun chains* [Sch85].

(6)

Reference to a proposition [d93-11.2]

utt73:	s:	we also have engine E one that's going back and forth right
utt74:	u:	oh right **that's** true (that = the proposition that engine E1 is going back and forth)

Example 6 shows another demonstrative pronoun referring to a fact or proposition. In this case, the pronoun is the subject of a copular construction, and the predicate of the sentence 'is true' provides strong semantic clues about the correct interpretation of the pronoun. As these examples demonstrate, discourse deixis is often, but not exclusively, performed by demonstrative pronouns.

2.2 Corpus observations

In order to get a feel for how pronominal reference works in task-oriented dialog, a study of third-person pronouns in the TRAINS93 corpus was conducted in 1998 [BA98]. In this study, two human annotators determined the correct referent of definite and demonstrative pronouns in a set of 20 TRAINS93 dialogs containing 368 markable pronouns. The study's goals included characterization of the contrastive behavior of demonstrative and definite pronouns in the corpus, corroboration of previous linguistics research on demonstrative pronouns in this corpus (since those studies did not analyze task-oriented language), and development of a catalog of abstract entities that are actually the subject of pronominal reference in this domain.

In the study, we first wanted to investigate what types of entities appear as the referent of pronouns in the corpus. Table 1 shows the categories found. In this chart, starred category numbers indicate the referent is considered an abstract entity. While domain objects are the most common referents, reference to abstract entities such as the evolving plan or times

Referent Semantic Type	Example Linguistic Trigger	Count	
		It	That
1) Train engines/boxcars/cargo	engine 1	118	44
2*) The plan or a plan segment	(none - global focus)	9	64
3*) The end clock-time of a plan segment	three p.m.	26	18
4*) The amount of time to execute an action	three hours	16	15
5*) An action in the task domain	loading the oranges	2	18
6*) The propositional content of an utterance	That's right	0	12
7) The task itself, the instructions	they said we're supposed to...	7	3
8*) A fact about the task domain	an engine can only carry 3 boxcars	1	5
9*) Routes in the hypothetical task world	from Avon to Bath	2	3
10) Cities in the hypothetical task world	Avon	0	3
11*) The goal, the assigned completion time	arrive in Avon by noon	0	1

Table 1: Types of referents found in the annotated dialogs

	Syntactic Form of Linguistic Trigger			
	Pronoun	Noun Phrase	Non-Noun Phrase	No Linguistic Trigger
Definite	38%	37%	7%	18%
Demonstrative	8%	17%	35%	37%
Overall	23%	27%	21%	29%
	$\overleftarrow{definite\ \ preferred}$		$\overrightarrow{demonstrative\ \ preferred}$	

Table 2: Distribution of linguistic trigger syntactic form by pronoun class

are also common, comprising around 50% of the total pronouns. This chart also reveals the fact that demonstrative pronouns are more likely than definites to be chosen for reference to abstract entities. 50% of third-person pronouns in the corpus were demonstratives, and 67% of discourse deictic pronouns were demonstratives. For cases where an abstract entity is referred to via a definite pronoun, the linguistic trigger of the definite pronoun is a demonstrative pronoun 25% of the time. This interacts with the relationship between the choice of pronoun and the syntactic form of the linguistic trigger, discussed next.

Demonstrative and definite pronouns have been found to have different preferences for certain linguistic trigger surface forms[Sch85]. Table 2 shows the distribution of syntactic forms of linguistic triggers for demonstrative and definite pronouns in the annotated dialogs. Even though triggers of each syntactic type occur with almost equal probability overall, the pronoun chosen for anaphoric mention of the entity correlates strongly with the syntactic form of the linguistic trigger. A definite pronoun is strongly preferred when the trigger is a pronoun, and the more non-noun-phrase-like the trigger, the more likely a demonstrative pronoun becomes. Demonstrative pronouns are strongly preferred for non-noun-phrase

	Pronoun and Trigger Subj.	Pronoun and/or Trigger not Subj.
It	28%	72%
That	10%	90%
Overall	19%	81%

Table 3: Distribution of it/that based on subjecthood

triggers and for entities that have no prior explicit mention[3] A definite pronoun is nearly 5 times as likely to be chosen when the trigger is a pronoun.

Linguists have shown that when both the pronoun and its trigger are the subject of their respective clauses, definite pronouns are highly preferred over demonstratives, otherwise demonstratives are preferred [Sch85]. As Table 3 shows, this pattern holds true in the TRAINS93 dialogs.

[3]In the annotation manual, *no linguistic trigger* is used in two cases 1) when the exact referent has not been spoken before in the discourse; for example after an action is discussed the next sentence can be "So that'll be 2am when we get to Bath". The referent for *that* has never been mentioned, it is inferred based on the discussion of the action. 2) when the referent has been discussed but it requires more than one turn in the dialog to represent the linguistic trigger. For example, the entire plan so far.

3 Background and Related work

Due to their central role in the discourse interpretation process, a significant body of research on how to process anaphoric referring expressions exists in the computational, psychological, and theoretical linguistics literature. The work of anaphora resolution can be divided into two subtasks: "1) identifying what a text potentially makes available for anaphoric reference and 2) constraining the candidate set of a given anaphoric expression down to one possible choice."[Web79, pg 1]. That distinction provides a good way to organize the work reported here, so this section first surveys previous work exploring what discourse entities are evoked in a text and how to organize them, then it discusses theory and implementations for pronoun resolution. We do not discuss resolution of other types of anaphoric expressions.

3.1 Discourse Entities

Ideally, the DC should contain tokens for every entity, whether abstract or concrete, that could be re-mentioned in the discourse [Web79]. But distinguishing which linguistic constituents should cause a token to be instantiated is a complex problem and an active area of research. It involves, *inter alia*, set construction [All95], detailed analysis of quantifier scope and negation [Kar76], and calculation of discourse segment boundaries [Web90]. In order for all possible anaphoric reference to be resolved, DE tokens need to be created by other constructions in addition to noun phrases. Although the anaphoric referring expression is nominal, as we saw in section 2, the linguistic trigger can have a variety of syntactic forms.

Many previous authors have acknowledged the need for a wide variety of entities to be represented in the discourse model. Schank [Sch77] includes objects, people, locations, actions, states, or times. Webber states that "discourse entities may have the properties of individuals, sets, events, actions, states, facts, beliefs, hypotheses, properties, generic classes, typical set members, stuff, specific quantities of stuff, etc."[Web81, p. 283], however her model does not describe techniques for adding all these sorts of entities to the DM. Kameyama asserts that discourse entities need to be created for all sets, states, objects, actions, events, and individuals [Kam86]. This section first describes the few studies conducted by previous authors on discourse entity creation. Implemented systems are lagging behind the theory in this area. Section 3.1.3 describes the state of the art.

3.1.1 Theoretical studies

Several detailed studies have analyzed when a text constituent makes a discourse entity available for subsequent reference. Many of these studies concentrate on noun phrases, but some more recent studies address other constituents. Prior to these studies, it was assumed that all indefinite noun phrases create discourse entities. A more insightful analysis was provided by Karttunen in his seminal work on this issue [Kar76]. He describes the impact of the construction enclosing a definite noun phrase (mainly negation and quantifier scope) on the status of a mentioned entity. Karttunen introduced the practice of judging whether a discourse entity should be instantiated or not based on legality of anaphoric reference. His criteria for discourse-referenthood is "the appearance of an indefinite noun phrase establishes

a *discourse referent* just in case it justifies the occurrence of a coreferential pronoun or a definite noun phrase later in the text."(pg. 366). The term "discourse entity" is now more commonly used in CL but means essentially the same thing as Karttunen's discourse referent. He also points out that the form of the sentence is what drives DE creation, not the ontological status of the entity in the real world, because sentences about Santa Clause or Unicorns are just as valid for setting up anaphora as sentences about houses and cars. He describes the following cases[4]:

- **Copula:** "*Pat* is a linguist." introduces an individual Pat. The phrase 'a linguist' does not create a discourse entity.
- **Generics:** "*A lion* is a mighty hunter." Would trigger a token for lions in general, but not for one particular lion.
- **Negation:** "Pat has *a car*$_i$" can be followed by "It$_i$ is brown" while "Chris does not have *a car*" cannot.
- **Positive implication verbs:** Certain verbs such as *manage* and *remember* imply the truth of the complement sentence. As long as the verb isn't negated, indefinite noun phrases in the complement structure do establish referents. "Chris managed to find *an apartment*$_i$. It$_i$ has a balcony."
- **Negative implication verbs:** Other verbs such as *fail* and *neglect* imply the negative truth value of their complements and therefore have the opposite rule. "Pat forgot to write *a term paper*. *It describes the French Revolution."
- **Factive Verbs:** Presuppose the truth condition of the complement sentence. Indefinite NPs in the complement of these verbs should create a DE whether or not the main verb is negated. For example "Pat (knew/didn't know) that Chris had *a car*$_i$. It$_i$ was in storage."
- **Nonfactive verbs:** such as *believe*, *think*, and *claim* make no presupposition about the truth of the complement sentence. Indefinite NPs in these sentences should instantiate a DE whether the verb is negated or not. If the verb is negated, reference to the entity can only be performed by other speakers. For example, contrast A: "I doubt that Pat has a car. I've seen *it." with A: "I doubt that Pat has *a car*$_i$." B: "I've seen it$_i$."
- **Modality maintenance:** Verb modality must be interpreted and DEs must continue to be mentioned in the same mode (eg. future tense). For example, you can't say "I wanted a dog. *It was a chihuahua". But you can say "I wanted a car. It would be shiny and fast." You can't say "Chris doesn't have a car. *It is brown." But you can say "Chris doesn't have a car. It was totally demolished last week." "Pat expects to have a baby. The baby's name (*is/will be) Chris."

[4]In the examples, subscripts denote coreference, an asterisk signifies illegal reference.

- **Yes/No Questions and Commands do not introduce a DE:** "Does Chris have a car?" "*It is a mustang". Commands do not introduce a DE: "Give me a hotdog. *It looks delicious." but within the same sentence a short-term entity is created "Give me a hotdog and please put mustard on it."

- **Quantification:** Short-term referents can be built within quantifiers that are only valid as long as the scope of the quantifier is prolonged with phrases like "usually" or "at every conference".

In her seminal work on discourse deixis, Webber [Web88; Web90] analyzed reference to discourse segments in text. She found that a hierarchic discourse structure representation, à la Grosz and Sidner [GS86], can be built to determine which segments are available for anaphoric reference. This theory is problematic because there is no clear definition of what constitutes a discourse segment [All95; Rei85], and as a result there are no clear computational mechanisms for demarcating discourse segment boundaries in real time, unless an extremely naive approach is used such as considering discourse segments to be equivalent to sentences. She proceeds under the assumption that an oracle exists that will decide the hierarchic structure in real time during the discourse understanding process. Her proposed algorithm adds a DE to the context for each discourse segment, where a discourse segment can be either a major clause or a sentence. Each such DE is attributed with three properties: the speech act import of the segment, the form of the segment, and its interpretation as a situation, event, object description, etc. She leaves open the question of exactly when in the process of discourse interpretation to add such entities to the DC, and the salience of these entities relative to other DE's introduced by noun phrases. She argues that even if the algorithm doesn't create DEs for all discourse segments, it should definitely create one for each segment that is the referent of a subsequent pronoun.

In [Web90], she refined her model somewhat to include a notion of salience of discourse segments. A tree representing the relationships between discourse segments is built. Segments along the *right frontier* are taken to be in focus and available for deictic reference. The set of segments comprising the right frontier is all open segments and the most recently closed segment. These segments provide a frame of reference for discourse interpretation in a similar fashion to the stack of *focus spaces* representing the attentional state in Grosz and Sidner's model [GS86]. The set of open focus spaces corresponds loosely to those discourse segments whose purpose is still open to additional support. Webber made a key observation in this work that discourse deixis refers to some aspect of the *interpretation* of a stretch of text, not to the text itself. Therefore, aspects of the interpretation of each discourse segment, for example its speech act and propositional content, must be in the DC.

3.1.2 Logical Form

Beyond heuristics for generating discourse entities from natural language, semanticists have attempted to formalize the accessibility between referents in a logical form and the pronominal elements that need to be bound to them. The most commonly used logical representation that aids anaphora resolution by explicitly encoding accessibility for anaphoric referents

is Discourse Representation Theory (DRT), developed independently by Heim and Kamp [Kam81; Hei82; KR93][5].

Previous semantic theories suffered from the fact that indefinite expressions were treated as existential quantification. The scope of the existential quantification ends at the end of the sentence, so there was no way to keep the discourse entities open for pronoun binding in subsequent sentences because the existential scope had been closed.

Discourse entities in this formalism are called *discourse markers* and the traditional formulas of first order logic are now represented as Discourse Representation Structures (DRS's) that consist of pairs $<M,C>$. M is a set of discourse markers and C is a set of conditions relating to the discourse markers. In sentence 7, there are two discourse markers x and y with three conditions applying to them.

(7)

Pedro owns a donkey.

x y
PEDRO(x)
DONKEY(y)
OWNS(x,y)

We can add more conditions and discourse markers in subsequent sentences.

(8)

Pedro owns a donkey. He rides it.

x y w z
PEDRO(x)
DONKEY(y)
OWNS(x,y)
$w = x$
$z = y$
RIDES(w,z)

DRS's can be related by implication (conditionality) and can be embedded to represent enclosing contexts.

Because of its excellent form for modeling anaphoric reference, recent attempts have been made to extend the formalism so that abstract entities are represented. Poesio and Traum extend DRT to represent speech acts and situations [PT97]. Although DRT provides a formal way of representing things that can be referred to and equality of reference among the symbols, it does not provide a way to compute that equality.

[5]This summary is a paraphrased version of the longer summary in [Poe94].

3.1.3 Implementations

Although there is ample evidence of the need to construct discourse entities for abstract referents, in practice the simplest possible model is usually implemented. Most implemented systems operate under a *coreference model* in which pronoun resolution is computed by co-indexing the pronoun with another constituent in the surface form. In such a model, no DE tokens are required. No discourse model need be built, so there is no deeper representation to link the anaphoric expression to.

Some implementations do contain a discourse context model, but only generate discourse entities for simple nominal expressions. Larger constituents such as infinitives, gerunds, sentential complements, and even conjoined nouns do not have corresponding tokens in the DC; and no Kartunnen-like analysis of indefinites is performed. Pronoun resolution against such a DC can only find referents that were previously mentioned in a noun phrase (cf.[Hob86; Mit98; Str98]. Discourse deictic pronouns have no hope of being resolved correctly because tokens for abstract entities are not in the DC to begin with. The vast majority of contemporary algorithms for pronoun resolution fall into this category.

A few extant systems do build a discourse context and resolve pronouns to discourse entities rather than surface constituents. The TRIPS system developed at the University of Rochester [FA98] has an explicit model of discourse context against which all anaphoric referring expressions are resolved. The system has an explicit representation of the task world, and tokens for all objects in the world are initially loaded into the discourse context so that explicit reference using names or definite descriptions can be resolved. As the dialog progresses, objects mentioned by the human user of the system are saved in a focused-objects list, and the salience of those objects gradually fades. Tokens are only generated for domain objects such as vehicles, cities, and cargoes. No discourse-deictic tokens are generated.

The SRI Commandtalk system implements a complex discourse context model that can resolve many types of definite anaphora and understands an impressive range of ellipsis [SDG+99]. The discourse context includes a semantic representation of both the system and the user's last utterance, and a representation of the propositional content of the system's last utterance. These objects are used to interpret subsequent anaphora and ellipsis. Objects created by the simulation agent are available for pronominal reference even if no prior linguistic reference to them exists, and their salience is calculated in decreasing order of recency. Linguistic reference to an object causes its salience to be boosted.

The Lucy system developed at the University of Texas contains a very robust discourse representation. Its anaphora resolution component contains an implementation of a discourse pegs model [Lan86a; Lan86b]. The discourse pegs are effectively DE tokens, and provide a separation between the linguistic and knowledge base layers. Luperfoy's dissertation work [Lup91] puts this representation to work in a system that handles many types of nominal anaphora. Her system handles many partial anaphora including generic to specific, count to mass, and individuals to sets. However, her work did not include pegs for abstract entities other than Kinds, and does not address demonstrative pronouns.

3.2 Discourse Entity Cleanup

Prior authors differ in their opinions about the life span of a DE token. Kartunnen discussed the lifetime of short-term referents, but did not address the lifespan for other DEs. In models based on embedded discourse structure, discourse entities have an embedded scope and are deleted from the DM at the conclusion of an embedded segment [GS86]. Other authors have used a sliding-window approach, fixing a maximum number of sentences during which entities are left in the discourse context. For example, Hobbs' naive algorithm searches only the previous four sentences for a referent. DM tokens are deleted once this limit is reached. Models based on the centering framework tend to have a sliding-window size of one. As each sentence is interpreted, a new C_f list is built to replace the previous one [BFP87; Str98; SE99]. Such models cannot resolve long-distance pronominal anaphors. Walker [Wal96] proposes that the DC can contain a maximum number of tokens, like the cache in a computer, and once this limit is reached old tokens should be removed. She says "The exact specification of this capacity must be determined by future work, but previous research suggests a limit of two or three sentences, or approximately seven propositions" (pg. 258). Suri and McCoy created focus lists similar to those created in centering, but they pop each new structure onto a stack and never delete any [SM94].

I believe that these theories are ignoring a crucial distinction. When a pronoun resolution strategy depends on choosing referents by salience, it is crucial to keep the set of candidates small so that relative focus can be established. For those models only the highly-salient in-focus DEs should be kept in the DC. But it is possible to refer to an entity whose previous mention is an arbitrary distance from the pronoun by including disambiguating information in the sentence. For example, several hours after seeing a movie and with no mention of the movie in the immediately prior discourse, one can coherently say (to someone else who went to the same movie) "That was the worst movie I've ever seen." Human interpreters can resolve such referents as long as no competitor movie entities exist in the intervening discourse. An approach, such as Suri and McCoy's, that makes no assumption about the amount of discourse between a linguistic trigger and pronoun, can handle more cases than one that deletes DEs after a certain number of sentences.

3.3 Pronoun Resolution

The second subtask within pronominal anaphora resolution is narrowing the candidate set of referents in the DC down to one correct referent. Because pronouns are so minimally marked for semantic content, evidence from the rest of the sentence must be gleaned to help isolate possible referents for the pronoun. This chapter surveys syntactic, semantic, and pragmatic factors that can be brought to bear on the pronoun resolution process.

3.3.1 Theory

Syntactic constraints For those pronouns with an antecedent noun phrase, a powerful first pass for resolution is the application of syntactic constraints to rule out candidate referents. English has some specific agreement and structural rules that can be reliably applied to determine which referring expressions can/can't be coreferential.

Some syntactic constraints are trivial, including gender and number agreement between the anaphoric expression and the referent. Syntactic structure can be used to find possible referents of reflexive and reciprocal constructions such as *itself* or *one another*, and can also be used to find impossible referents for definite pronouns. Government and Binding theory [Cho81] is one of the most influential theories demonstrating reliable syntactic constraints on intrasentential coreference. The theory exploits a structural concept called *C-command*: a node c-commands its siblings in the parse tree and all the nodes dominated by its siblings. The impact of c-commanding on reference is as follows (from [JS89]):

1. A non-pronominal NP cannot overlap in reference with any NP that c-commands it.
2. The antecedent of a bound anaphor must c-command it. (bound anaphora are of the type in example 10 below.)
3. A personal pronoun cannot overlap in reference with an NP that c-commands it.

Rule one determines that 'Pat' cannot corefer with the entity or set of entities referred to by any of the pronouns in example 9 (Examples taken from [JS89]).

(9)

(a) He_i likes Pat_j.
(b) They_i like Pat_j.
(c) He_i likes pictures of Pat_j.
(d) They_i like pictures of Pat_j.
(e) He_i told them about Pat_j.
(f) They_i told him about Pat_j.

Rule two helps identify the possible antecedents of reflexives, as in these examples where possible coreferential phrases are emphasized:

(10)

(a) Pat_i likes $himself_i$.
(b) Pat_i likes pictures of $himself_i$.
(c) Pat_i told $Chris_j$ about $himself_{i/j}$.

Rule three disallows the following subjects as coreferential with the pronouns, because they violate the rules for using the reflexive:

(11)

(a) Pat_i likes him_j.
(b) Pat_i likes pictures of him_j.
(c) Pat_i told Chris_j about him_k.

These syntactic rules can be used as a first pass to eliminate impossible and identify possible antecedents [JS89].

Semantic constraints It is incontrovertible that semantic information is required to disambiguate some referring expressions (cf.[Cha72; Hob86; CB88; PiWB98]. Techniques for using semantic information to disambiguate reference either calculate semantic distance or use semantic features. Semantic distance involves, for example, finding the *closest* previous entity as the referent of a bridged reference, where closeness is based on some representation of a semantic landscape. Semantic feature systems usually assign the entity a value along many orthogonal dimensions, for example, entities can be animate, edible, movable, etc. These features are typically used to build restrictions on the objects for the head verb. Semantic restrictions are useful to resolve ambiguous pronouns in sentences like example 12 (from [CB88]), in which the preferred object of eating is something edible (the cake) and the preferred object of washing is something solid (the table).

(12)

Pat took *the cake*$_i$ from the table and ate *it*$_i$.
Pat took the cake from *the table*$_i$ and washed *it*$_i$.

It is in theory possible to encode lexical entries for verbs detailing the features required of their arguments, but we would obviously prefer to avoid hand-coding lexical features if an automatic means for boot-strapping the knowledge from a corpus can be found. Manual coding of semantic constraints for general language understanding is infeasible, for example we would be unlikely to trap the feature that makes *the cake* in the example above dispreferred as an object for *wash* (imagine a feature +/-disolvable). An active area of research is the use of statistical analysis to cluster entities into feature sets and to find preference patterns of thematic roles for particular verbs (cf.[DI90; Res92].

Carbonell and Brown also discuss another semantic constraint: case-role parallelism. This is similar to the result reported in Kameyama (discussed below), but an important difference is that here they are talking about semantic, rather than syntactic, parallelism. For example:
(13)

(a) *Pat*$_i$ carried the box of papers from Chris to *Peter*$_j$.
(b) *He*$_i$ also sent *him*$_j$ Mary's address book.

In this example *he* and *him* in the (b) sentence each have three possible antecedents, but the semantic parallelism triggered by *also* totally disambiguates the pronoun binding. The authors report that this rule accounts for a large number of disambiguations in their test set.

Semantic restrictions carried by the verb and other adverbs in the sentence can be used to constrain the reference. Contrast, for example, the sentences in example 14 (taken from [Web88]). In B1, *where* can be used to disambiguate *that*: it must be a location. Analogously, *what* in B2 and *when* in B3 give strong clues about the correct referent. Webber compares this to spatio-temporal use of deictic gestures, "where the listener recognizes the general pointing gesture, and then tries to figure out the intended demonstratum based

on what the speaker says about it (and on general heuristics about what might be worth pointing to)" [Web88, pg. 118].
(14)

(a) In the Antarctic autumn, Emperor penguins migrate to Tasmania.
(b1) **That's where** they wait out the long Antarctic winter.
(b2) **That's what** you're likely to see there in May.
(b3) **That's when** it begins to get too cold even for a penguin.

World knowledge can also be applied, as in example 15 (from Sidner[Sid83a]). Both potential continuations have two ambiguous pronouns, but their referents can be determined unambiguously based on facts about the world. Humans understand that *him* in continuation b-1 is the vet because vets have hands. In continuation b-2, *he* can only refer to the vet because dogs cannot give vaccinations.

(15)

(a) I took my dog_i to the vet_j the other day.
(b-1) He_i bit him_j in the hand.
(b-2) He_j gave him_i his heartworm vaccine.

World knowledge also includes application of preconditions/postconditions of actions, such as in example 16, in which only Chris can eat the apple because he has it as a result of sentence (a).
(16)

(a) Pat gave $Chris_i$ an apple.
(b) He_i ate it.

Pragmatic features Kameyama discovered that continuity of sentential role effects interpretation of ambiguous pronouns [Kam86]. In example 17, informants tend to interpret *he* in (b) as Max and *him* as Fred because of their sentential roles in (a). This observation helps extend centering (discussed below) so that it correctly models situations such as (b) with more than one ambiguous pronoun. In both Japanese and English, preservation of syntactic role is a stronger preference than preservation of center.

(17)

(a) **Max** is waiting for **Fred**.
(b) **He** invited **him** to dinner.

Kameyama also noted that certain verbs, like see/hear/look/sound "anchor the speaker's perspective" and create a preference for continuation of the perspective in the next sentence [Kam86]. This preference disambiguates pronouns such as the one in sentence (c) of example 18. Since sentence (b) takes Dan's perspective, Jim is understood to be seen in sentence (c).

(18)

(a) Dan$_i$ went to a party yesterday.
(b) He$_i$ saw his high school friend *Jim*$_j$.
(c) *He*$_j$ looked awfully pale.

Sentence modifiers can also be used to signal parallelism, as in example 19.

(19)

Carl$_i$ is talking to Tom$_j$ in the lab.
Terry$_k$ wants to talk to him$_j$ *too.*

Hobbs gives a heuristic that anything that has been conjoined is preferred as the referent of a plural pronoun [Hob86]. So in example 20 the entire set *human bones and relics* would be preferred as the referent for *they.*

(20)

(a) *Human bones and relics*$_i$ were found at this site. *They*$_i$ were associated with elephant tusks.

Corpus studies have revealed the preference for certain referring expressions, and those preferences can be used to build heuristics for pronoun resolution. For example, Prince found that Information status influences the choice of entity in subject position [Pri91]. 85% of syntactic subjects in her evaluation corpus were discourse-old entities. Previous authors had claimed that information status relative to the hearer (hearer-old/hearer-new) affected subjecthood, but Prince found that effect to be subsumed by the entity's discourse status.

Salience Besides syntactic constraints on coreferentiality, the second most active area of research on pronoun resolution within the computational linguistics community is on formal models of salience [Pas89]. Calculation of salience is a two-sided task, involving the discourse state and the form of referring expressions. The set of possible referents to a pronoun is constrained by the discourse state, but the very act of using a pronominal reference immediately alters the discourse state by making the pronoun's referent highly salient [Isa74]. It is a fairly well accepted fact that the focus of attention of dialog participants does affect their referring expressions [PiWB98]. Focusing is essential to discourse, guiding the listener from one utterance to a relevant next utterance. Sidner's influential early 1980's work related focusing to referring expressions. She points out that "focusing is a discourse phenomenon rather than one of single sentences" [Sid83b, pg. 366]. Sidner says that "Focus is the building block on which topics are constructed; without focusing, the discourse ceases to be a discourse"[Sid83a, pg. 129]. Many recent pronoun resolution techniques rely entirely on salience calculations; however, it is clear that salience isn't the whole story for pronoun interpretation. Salience should be brought to bear after a first pass of filtering is applied and multiple candidate referents remain. "When all other possibilities have been examined and

there remain multiple ambiguous candidates for a pronoun, [salience calculations] provide a good preference ordering for their referents" [Kam98].

Psycholinguistic studies have shown that pronouns should be used to refer to the most salient objects, while definite descriptions should be used for reference to other discourse entities. The subset of DC entries that are available for pronominal reference is sometimes called the *local context* as opposed to the *global context*, which includes the rest of the DM. Grosz coined the term *focus* [Gro77] for the local context. It is important to have a system for language understanding track some notion of the focus, because in many cases the system can use the focus to severely restrict the candidate set of items from which to choose a referent [Gro81]. Various theories attempt to establish a psychologically-plausible size and organization for the focus set, and account for how items move to this set from main memory and back [Ari90; Gro77; Haw91; Wal96]. Recency in the discourse structure [GS86] obviously affects what items are in focus, but the effects of recency must be counter-balanced against a notion of topicality. Chafe compares the focus to a stage [Cha70]. An evoking reference causes its referent to come to center stage along with all its implicitly mentioned, related DEs. From there it slowly retreats towards the wings unless it is re-mentioned. Chafe asserts that only items on stage (in focus) are accessible via anaphoric reference. Grosz proposes that any algorithm for focus must accomplish these three things: 1) differentiate among the items in the knowledge base on the basis of relevance; 2) account for implicitly focused items; and 3) include mechanisms for shifting focus [Gro77]. Most current algorithms only perform the first of these tasks.

Guindon proposes a psycholinguistically-motivated model wherein the attentional state is divided into three levels [Gui85]. The *cache* contains a finite number of the most salient items and has the quickest access time. The cache contains a buffer for the raw, unprocessed input sentence, a set of the most recent items mentioned and another set of the most topical items. Pronouns should be used to refer to items in the cache. *Operational memory* contains all the other objects mentioned in the discourse but no longer in the cache. A topic shift results from bringing new items into the cache, and old items must be spilled to the operational memory to make room in the cache. Definite descriptions should be used for items not in the cache. Guindon's experiments showed that interpretation of anaphora is hindered when a pronoun is used for an item not in the cache or when a definite reference is used for an item in the cache. While he does not specify the details of how to choose among items in the cache for a specific pronoun resolution, the fact that he combines most-topical and most-recent entities into the candidate set for pronoun referents is interesting.

Walker proposes a similar model in which the cache is more directly modeled on computer cache memory [Wal96]. She considers the cache to be a "working set consisting of discourse entities such as entities, properties, and relations that are currently being used for some process"(pg. 258). Because of the limited size of the cache, a cache replacement policy must be implemented to spill items from cache to long-term memory and to add new tokens to the cache. Walker's model, however, has no provision for a discourse model apart from the very small local context held in the cache. It also gives no indication of how to choose among items in the cache when faced with a particular anaphor. The model also gives no indication of how to maintain information about an entity's role in the current discourse when its token is spilled from the cache back to long term memory.

In contrast, a stack model of focus was proposed by Grosz and Sidner [Sid77; Gro77; GS86]. In this model, discourse structure is accounted for by three separate but interacting structures: attentional state, intentional structure, and linguistic structure. All three affect the DC. A discourse is said to be structured as a hierarchy of embedded segments. Each time a new subordinate segment begins, a new *focus space* is created. The collection of focus spaces available at any one time is the *focusing structure*, and is organized as a stack. Changes to the intentional state cause changes in the focus structure: a push occurs when the intentional purpose of a new segment contributes to the purpose of the immediately preceding purpose. A pop occurs when the intentional state returns to a higher-level segment. Reference to an entity in a higher-level segment causes the state to pop so that segment is now current. Alternations from one surface form to another, such as using a definite description coreferential with a pronoun, helps mark the transition between discourse segments. Although this model doesn't attempt to explain contrastive use of pronouns and definite descriptions, it does support the claim that discourse structure affects which DE's to consider as candidates for anaphoric reference. However, this model doesn't work for sequences such as example 21 (imaginary sequence in the discussion of how to remove the flywheel).

(21)

(a) A: Remove the setscrews.
(b) B: Done
(c) A: Now, back to the flywheel.
(d) B: Wait, I dropped the screws.

The stack-based model proposes that after utterance (c), the discourse segment regarding the setscrews has been popped, so the setscrews are no longer in the DC when utterance (d) is spoken. Therefore, the anaphoric definite reference to the screws in (d) cannot be resolved correctly. However, it is clear that unambiguous reference can be made to them since they are the only set of screws recently discussed in the dialog. In the stack-based model, resolution of anaphoric reference at discourse segment boundaries is still an open problem.

Sidner's [Sid83a; Sid83b] *process model* of focusing is a precursor of the centering work, discussed next. Her model describes a three-step process for focus:

1. The listener chooses an initial estimate of the focus.
2. The listener uses this focus to interpret anaphoric reference.
3. The listener updates the focus if anaphoric reference has been made to an object that is not the previous focus, and the focus was not mentioned.

In her model, the default choice for focus is the semantic object, usually represented syntactically as the direct object. This contrasts sharply with the centering framework's default choice of the syntactic subject as the preferred center of attention. Sidner's model maintains an *actor* focus and a *discourse* focus, a distinction that subsequent models have

abandoned in favor of a preference-ordered list of potential foci. However, for models that must handle spoken language, it appears that the actor vs. discourse focus distinction is important [BS98]. Sidner's model combines syntactic clues as well as which entities are anaphorically expressed to determine the discourse focus of each utterance. If a full NP is used to refer to a previously focused item, this is interpreted as a shift back to the old focus.

There are many factors that affect salience. Syntactic forms have a bearing on the focus, for example in a command like "Pick up the screwdriver", the screwdriver is incontrovertibly in focus. Other syntactic devices are *there-insertions* such as "There are two trucks at Calypso" and clefts like "It's a good idea to floss your teeth". I disagree with many illustrative examples in the computational literature that profess to show how one factor, usually pronominalization, impacts the topicality of an entity. Typically the example fails to hold all the other factors constant. All exert an influence over the sentence and one cannot give an example of focus that manipulates more than one of these variables at a time and still adequately accounts for the focused item.

Although many theories of discourse structure and focus attempt to divide the discourse model into accessible and inaccessible partitions, a more flexible approach would utilize a focus structure that partitions entities into relatively more or less accessible sets [Pas95]. The centering model accomplishes this by representing focus as a ranked list.

The centering framework is one of the most influential models in computational linguistics relating a speaker's local focus of attention to the form chosen for referring expressions [GJW83; GJW86b; GJW95; WJP98]. This thread of research seeks to formalize an account of discourse coherence that matches coherence judgments made by human readers. An assumption underlying the centering framework is that speakers choose referring expressions in order to minimize the inferences needed by the listener. The primary motivation for this line of research is to develop an explanation for referring expressions that represents the interaction between syntactic, semantic, and pragmatic factors. The notion of center of a sentence cannot be explained using only one of these three superstructures of language. The center is a notion of local focus of the most extreme variety: "the entity that an individual utterance most centrally concerns."[GJW95, pg. 205]. The centering model as currently formulated only handles centers triggered by singular, definite referring expressions, not demonstratives. However the original centering paper does allow that "events and other entities that are more often directly realized by verb phrases can also be centers."[GJW95].

The centering framework makes three claims: 1) given an utterance U_n, the model predicts which discourse entity will most likely be the focus of U_{n+1}; 2) when the local focus is maintained from one utterance to the next, it will be expressed with a pronoun; and 3) when a pronoun is encountered, the model provides an ordering on possible antecedents from the prior utterance. The *forward-looking centers list (C_f)* in the centering model is basically a DC that is segmented by utterance. The centering framework itself does not provide a definition of what DE tokens should be included. The canonical ordering of entities in the C_f list for English is Subject > Direct Object > Indirect Object > Complements > Adjuncts [BFP87]. This ordering reflects the relative salience of discourse entities. The centering model does not handle long-distance reference, only entities in the previous utterance are available as candidate referents. This limitation makes adaptation of the model for spoken dialog tricky.

Because the DC is represented as a ranked list, the authors claim that the centering model can dispense with the multiple categories of focus in Sidner's model (actor focus, discourse focus). Centering also handles some cases of multiple pronouns that were problematic in Sidner's model by allowing other entities to be pronominalized as long as the C_b is [GJW83]. The centering framework leaves the question of what syntactic constituents trigger an entry in the C_f open to the designer. However, most discussions of the model follow the typical pattern of creating DEs for entities introduced by simple noun phrases. Since clause-like constituents never seem to be candidates, centering as it is described does not handle reference to abstract entities.

Subsequent studies sought to overcome the limitations in centering. Passonneau questions the assumption in the centering framework that every utterance has a C_b, and especially the notion that the center is a property of a sentence [Pas93]. She observes that "if we can't determine the center without the prior utterance, why is the center said to be a property of an utterance?"(pg. 199). Her formalism defines the local center in terms of pairs of utterances. The original paper described C_f lists being generated at the end of each sentence. As a result, intra-sentential referents could not be resolved. Many authors have published subsequent work to create C_f lists at the end of each independent clause to allow for intrasentential reference in complex sentences. Kameyama extends centering theory to incorporate other constraints on pronoun resolution, such as the effect of embedded contexts and parallelism.

Several authors have found that the information status of discourse entities impacts their salience in a variety of ways. Information status has two somewhat orthogonal components, information status in the discourse and information status for the hearer. Hearer status affects whether the evoking mention of an entity can be definite or indefinite. For anaphoric mentions, Prince found that Information status influences the choice of entity in subject position [Pri91]. Work in [SH96] suggests that information structure should dominate over sentential role in determining the ordering of C_f elements within a centering-based model. They define the $C_b(U_n)$ as representing the given information (roughly, cohesive link to previous discourse) and the theme to be the C_p (roughly, topic of current sentence). Then they define a partial ordering for the C_f:

- First pass ranking: bound elements > unbound elements
- Within bound elements: anaphora > possessive pronouns $\otimes$ elliptical antecedent > elliptical expression $\otimes$ head of anaphoric expression
- Final sort: nom head_1 > nom head_2 > ... > nom head_n

Bound elements (discourse-old) are ranked above unbound elements, within bound elements the degree of accessibility is ranked, then as a final pass elements in the same category are sorted based on surface order of the head constituent. Using an approach similar to BFP, the authors compared this ranking of C_f against the canonical syntactic-role based ordering used in the original BFP experiments, and found the functional ordering produced more coherent discourse, based on the inference load implied by pairs of center transitions.

Gundel *et.al.* developed a *Givenness Hierarchy* that they claim impacts the form of referring expressions [GHZ93]. Based on the attentional status of the entity, the hierarchy is

divided into the categories shown in Table 4. The authors claim that an entity's classification in one of these categories is both a necessary and sufficient condition for referring to it with the forms listed under that category. Therefore, the form used for reference can be taken by the addressee to be a signal indicating where in his prior knowledge to search for the referent. This hierarchy is different from others that considered statuses to be mutually exclusive, because in this hierarchy each status entails all lower statuses (statuses to the right). The most crucial point in this model for our current purposes is the distinction between focused items and activated items. Only focused items can be referenced with definite pronouns, while demonstrative pronouns can refer to activated items. Their definition of focused items is similar to other models. Focused items are topical entities that are likely to be continued as the topic in the next utterance, for example "subjects and direct objects of matrix sentences."(pg. 279). To be activated, an entity must be in short term memory due to a mention in the prior discourse or because of physical copresence. Therefore, this theory allows for pronominal mention of non-focused items via demonstrative pronouns. It explains the use of demonstrative pronouns for discourse deixis.

	in			uniquely		type
Status	focus	> activated	> familiar	> identifiable	> referential	> identifiable
Form	*it/them/they*	*this/that* *this N*	*that N*	*the N*	*this N* (indefinite use)	*a N*

Table 4: Gundel's Givenness Hierarchy

Pragmatic contrasts between definite and demonstrative pronouns Several previous authors have discussed pragmatic differences in the contrastive use of definite vs. demonstratives that can be brought to bear on the pronoun resolution process. Bare demonstratives present a special challenge to language understanding systems, because they are relatively unmarked as to agreement features [Cha80] and they are also ambiguous as to scope [Web90]. Studies in theoretical linguistics have observed that demonstratives are typically used to refer to clausal or sentential arguments or composite entities with conflicting semantic features. The more complicated the linguistic trigger or its surrounding context, the more likely it is to be replaced by a demonstrative rather than a definite pronoun and the less acceptable other choices become[Cha80; Sch85].

Use of the demonstrative pronoun indicates that its referent is some entity other than the expected focus of attention. Linde [Lin79] found that *it* was preferred for entities within the current local focus, while *that* was used for items outside the current focus of attention. Several authors [Sid83a; Pas93; Kam86] have claimed that a demonstrative reference does not pull the referent into focus. Fillmore observed that demonstrative pronouns should not be used to refer to animate entities except as the subject of a copulative sentence, for example contrast "That is my brother-in-law." with "That married my sister." [Fil82].

Passonneau's thesis research [Sch85; Pas89] gives a detailed account of the contrastive

use of demonstrative compared to definite pronouns. She discovered two major patterns in her dialog corpus:

1. If both the pronoun and the antecedent are in subject position, a definite pronoun is highly preferred, while if either the antecedent or the pronoun is not in subject position, a demonstrative pronoun is preferred.

2. If the linguistic trigger expression was more noun-like the subsequent pronoun chosen tends to be definite, whereas if the trigger was more clause-like, a demonstrative pronoun is preferred. The further the form of the trigger was from being a simple NP, the more likely it was to be pronominalized as a demonstrative. When the trigger was a pronoun (so called *pronoun chains*), the pronoun chosen was far more likely to be a definite, as long as it was in subject position.

Our analysis of the TRAINS93 corpus, summarized above in Section 2, corroborates Passonneau's observations. Webber and Passonneau both observed that after an abstract or clausal entity such as a discourse segment has been made explicit via a demonstrative reference, subsequent reference usually takes the form of a definite pronoun. Webber observed that discourse deixis is most often performed by demonstrative pronouns rather than definites (82% in her corpus, 71% in ours). Because of the contrastive use of definites and demonstratives, Passonneau concludes that discourse entities evoked by non-NP triggers should have a different status in the DC [Pas93].

Another study by Borthen *et al.* [BFG97] relates the form of referring expression and salience of the referent utilizing the Givenness Hierarchy of Gundel et al. Borthen wishes to account for the contrast between example 22 and 23, and uses this example to demonstrate the different accessibility for definite and demonstrative pronouns. The wording *the fact* in example 22 makes that fact focused. But in sentence 23, the fact is only activated.

(22)

a) What do you think about the fact that linguists usually earn less than computer scientists?
b) It's(That's) terrible!

(23)

A) I have heard that linguists earn less that computer scientists.
B) That's (*it's) terrible.

To support the statement that it's the wording and not the semantic type of the object that matters, Borthen gives this example:

(24)

1) There was *a snake*$_i$ on my desk today. It$_i$ scared me.
2) *There was a snake on my desk*$_i$ today. That$_i$ scared me, and it$_i$ scared my friend too.

Without the *that*, the *it* can't refer to the event, it refers to the snake.

Borthen corroborates the claim in Gundel that "analyses of naturally occurring discourse shows that [demonstratives] are used primarily when the referent is activated but not in focus." [BFG97].

Asher's monograph on abstract entities lists several heuristics concerning how definite and demonstrative pronouns refer to abstract entities. Asher observed that the distance between the linguistic trigger and the referent can be used to explain when a definite or a demonstrative pronoun is used. "Demonstrative pronouns are more often used to refer anaphorically to propositions and other abstract objects that are more than one sentence away or in another discourse segment, while *it* is a local pronoun, referring most often to abstract entities introduced in the previous sentence and mostly within the same discourse segment"(pg. 225). He also corroborates Gundel *et al*'s observation that "*That* prefers an antecedent that is peripheral or not part of the topic." (pg. 226).

3.3.2 Algorithms

Many pronoun resolution techniques have been published, most concentrating on singular, definite, third-person pronouns. Most implementations are restricted to coreference between the pronoun and a previous noun phrase, and some authors seem to be deluded that pronoun resolution is a simple task of choosing between a small set of candidate referents. In fact, many authors actually conflate the terms *pronoun resolution* and *coreference resolution*, when in fact coreference can only account for a subset of pronominal expressions.

Most computational algorithms start with a candidate set of referents and concentrate on attempts to eliminate candidates from the set using semantic, syntactic, and sometimes pragmatic knowledge (cf [Dor90; Hob86; LM90]. These calculations of *non-coreferentiality* draw upon structural as well as pragmatic considerations. If more than one possible DE remains, preference heuristics must be applied to choose a referent. The most naive algorithm would simply search the DC in order of recency for an entity that matches all the constraints on the anaphor [All95]. However, this approach fails to consider salience, sentence and discourse structure, and common sense knowledge.

Syntax-based Algorithms Winograd's very early system used recency plus a preference ordering on the sentential role of possible antecedents, preferring subjects over objects and both over the object of a preposition [Win72].

Hobbs' naive algorithm utilizes only the surface structure, but he also describes an alternative algorithm that exploits semantic constraints [Hob86]. His naive algorithm is very commonly cited as the baseline for other algorithms to beat. The naive algorithm relies on having a complete syntactic parse as input; so its utility for spoken language interpretation is limited, because the input can be ungrammatical or incomplete sentences. The algorithm walks the parse tree beginning from the pronoun node, checking candidate antecedents for matching gender and number. First candidates to the left of the pronoun in sibling nodes are checked, and if no match is found in the current sentence, previous sentences are checked breadth-first. This algorithm prefers intra-sentential antecedents but can also resolve inter-sentential reference. Hobbs reports 88% pronoun accuracy for this algorithm on text. After

augmenting the syntax-only naive algorithm with semantic restrictions, pronoun resolution accuracy increases to 90%.

Another algorithm that is very often cited as a baseline for evaluating pronoun resolution accuracy is Lappin and Leass's RAP (*R*esolution of *A*naphora *P*rocedure) algorithm [LL94]. The authors describe this approach as an algorithm to "select the antecedent noun phrase of a pronoun from a list of candidates"[pg. 535]. The algorithm is implemented as a series of filters:

1. Rule out intrasentential candidates on syntactic grounds (basically GB constraints).
2. Rule out candidates on morphological grounds (gender, number, person).
3. Selection of intrasentential antecedent for lexical anaphora (reflexives and reciprocals).
4. Rank the salience of candidates based on grammatical function using the order Subject > Direct Object > Indirect Object > Complements > Adjuncts and objects of prepositions.
5. Choose the highest-salience antecedent.

A novel characteristic of this technique is that coreferential nouns in the entire text so far are gathered into equivalence classes and the salience ranking for each element of the set is computed as the total salience of the class as a whole. That results in a global, discourse level of salience rather than a local sentence-by-sentence measure. This algorithm results in 85% correct pronoun resolution in their evaluation corpus. Once again, because of the detailed level of syntactic judgments needed in a system to implement this algorithm, such as adjuncts vs. complements, it is difficult to implement for spoken language interpretation.

Other algorithms depend on syntactic information alone but, unlike Hobbs' algorithm, are robust to incomplete or partial parser output. Such systems can be implemented for unconstrained domains. Breck Baldwin's CogNIAC system [Bal97] is an example. A unique property of his model is that it only chooses an antecedent when it is very confident of the selection. In other cases when the antecedent is ambiguous, it leaves the pronoun unbound. The authors claim that the approach is domain-independent and reflects human processing strategies (hence the name). The input must be POS-tagged, with simple noun phrases identified, the gender and number of those noun phrases can be calculated as well as very rough-grained semantic tagging (something like choosing between person/company/place). Once again, this system only considers pronouns with noun-phrase antecedents to be 'pronouns'. The algorithm correctly assigns 96% of antecedents for pronouns it resolves, but this represents only 60% of all pronouns. CogNIAC's rules are interesting, extending syntactic parallelism to possessives and using strict most-recent selection for reflexives:

1. If there is a single possible antecedent in the discourse, choose it.
2. For reflexives, choose the nearest candidate.
3. If there is a single possible antecedent in the prior sentence and in the read in portion of the current sentence, choose it.

4. For possessive pronouns, if the same possessive pronoun exists in the prior sentence, choose it as the antecedent.

5. If there is a single possible antecedent in the read in portion of the current sentence, choose it.

6. If the pronoun is the subject of the current sentence and the subject of the prior sentence contains a single possible antecedent, choose it.

7. Otherwise, the pronoun is ambiguous, no antecedent is selected.

Notice that the algorithm prefers intra-sentential reference over inter-sentential. The rules are evaluated in order and when an antecedent passes the test no more candidates are evaluated. This is one of the only pronoun resolution models that is well-suited for domains in which only high-confidence pronoun bindings should be processed, such as database queries.

Salience-based Algorithms Because salience is such an important factor in resolving pronominal referring expressions, most contemporary formalisms for pronoun binding are cast in terms of a theory of focus.

Although the Centering model was originally intended to explain local coherence, subsequent research by [BFP87] showed that it could also be used to bind ambiguous pronouns. This algorithm (henceforth called *BFP*) used the preference ordering of transition types described by centering rule 2 to choose between competing referents for ambiguous pronouns. The algorithm proceeds in three stages:

1. Build a list of *proposed anchors* containing each pronoun in the current sentence with each possible antecedent from the previous sentence.

2. Filter the anchors based on contraindexing and centering rule 1. If the proposed Cb of the anchor is not a pronoun, eliminate the anchor.

3. Rank the remaining anchors based on the transition type (as defined by centering). The ranking is continue > retain > shift-1 > shift-2.

4. Select the highest-ranked anchor as the new C_f list.

The BFP algorithm was the first to apply centering theory to pronoun resolution. It established centering as an important model in the pronoun resolution literature.

Subsequent evaluation of the BFP algorithm shows that it does not correctly model preferences for incremental pronoun interpretation [Keh97]. Because the C_b cannot be determined until the entire sentence has been interpreted, Kehler refutes the claim that the framework correctly models an addressee's immediate tendency to interpret a pronoun as referring to the current C_b. In fact, even minor changes to the wording of a sentence can impact the incremental assignment of pronoun interpretation. Kehler concludes that "In addition to the salience factors utilized by BFP, additional types of intersentential relationships must be taken into account."(pg. 474). In addition to intersentential coherence

relationships as explained by centering, other types of parallelism affect the default interpretation of pronouns.

Using the definition of utterance in Suri and McCoy, Strube extended centering so that it could handle inter-clausal as well as inter-sentential antecedents [Str96]. He observed that in a corpus of 15 German texts, 49% of intra-sentential anaphors have an antecedent in subject position, while 89% have an antecedent that is itself an anaphor. Because of that, he concludes that information structure should dominate grammatical role in determining preference orderings for pronoun resolution. He processes each clause in a compound sentence as a separate utterance. Complex sentences are considered a single utterance, however the C_f set for each utterance contains only elements from the matrix clause. The search order for referents is as follows:

1. For an anaphor in the first clause of U_n, propose the elements of $C_f(U_n - 1)$ in order.
2. For an anaphor in a subsequent clause of U_n:
   (a) Search context-bound elements of U_n from left to right.
   (b) Search elements of $C_f(U_n - 1)$ in order
   (c) Search all elements of U_n not yet checked from left to right.

This algorithm allows for some cataphoric reference resolution in step 2c. However, I disagree with the assumption that referents can only come from the matrix clause of previous sentences. Our annotation in the BUR corpus shows that 25% of inter-sentential anaphora are bound to elements in dependent clauses [Byr99a]. The unique contribution of Strube 96 is the change from syntactic criteria to information-status criteria in step 1, otherwise the algorithm is very close to Suri and McCoy. He claims that while his algorithm and Suri and McCoy might perform equally for fixed-word-order languages, his approach performs better for free-word-order languages. Therefore it is more generally applicable.

The LaSIE system at the University of Sheffield utilizes a simple focusing strategy: the expected focus is the object of a transitive verb or the subject of an intransitive or copula [AHG98]. In this system, agent pronouns prefer antecedents that have been the *actor focus* in previous sentences (searching from most recent of course) and non-agent pronouns prefer the current focus. It is unclear how they calculate whether a pronoun is acting as an agent since the two examples they provide are "he knew" as an agent and "it plowed" as a non-agent. This system reflects the results of pronoun resolution back to androgenous referents. For example, if *she* is co-indexed with *president*, and the president's gender wasn't known before, it should be updated.

Suri and McCoy developed an interesting pronoun resolution algorithm that balances the effect of continuation of focus versus subjecthood in calculating the salience of prospective antecedents [SM94]. Their RAFT/RAPR model is limited to inter-sentential reference for simple sentences. Candidate antecedents are pulled from previous sentences, one sentence at a time, working backwards from the pronoun. When the pronoun is the subject, the salience ordering is: subject > current focus > direct object > indirect object > other NPs in surface order. When the pronoun is not the subject, the salience ordering is: current

focus > subject > direct object > indirect object > other NPs in surface order. Their computation of current focus takes into account the surface form of the entity (pronouns are considered more focused than descriptive NPs), its referent (whether already in focus or a new DE), and discourse markers and other highly marked syntactic patterns such as there-insertion. This model overcomes one limitation in the centering model, that the local focus cannot be computed for the first sentence in a discourse. The RAFT/RAPR model falls back to syntactic preference of subject for the first sentence.

Suri and McCoy propose modifications to this model to allow it to process complex sentences. For 'X because Y' sentences, they gathered judgments from colleagues about preferences for pronoun antecedents, and conclude that the subject of the X clause should be resolved to the subject of the previous sentence, the subject of the Y clause prefers the subject of the X clause before the subject of the previous sentence, and the subject of the following sentence will prefer the subject of the X clause. My intuition is that because the RAFT/RAPR approach maintains two foci, the subject focus and the current focus, it is better suited for spoken language than centering.

Drawing on Sidner's focusing mechanism and the observation that pronouns in embedded clauses prefer intrasentential antecedents, Azzam et. al. [Azz96] develop an algorithm that achieves 95% pronoun resolution accuracy. A key observation in this research is that what they call non-PRR pronouns (everything but possessives, reflexives, and reciprocals) have intrasentential antecedents *only* if they occur in an embedded sentence. I suspect this heuristic applies well to the news reports in their evaluation corpus but may not apply to spoken language. Because a large percent of pronouns with noun-phrase antecedents have intrasentential antecedents (73% in Azzam's evaluation corpus, 71% in our annotated Treebank excerpt), the authors' main goal was to develop a focusing mechanism which could compute intra- as well as inter-sentential antecedents. In this algorithm, the focus list is updated after processing the first clause and before interpreting subsequent embedded clauses of the sentence.

More recently, so-called *low knowledge* techniques have attempted to derive rules for coreference determination that are robust to incomplete parses and lack of semantic information. Kameyama [Kam97] notes that "there seems to be a trade-off between the completeness of syntactic input and the robustness with real-world sentences." Using only reliable part-of-speech tagging, Kennedy and Boguraev implement a simplified version of the Lappin and Leass algorithm [KB96]. This algorithm isolates noun phrases in the input text, then chooses antecedents for each pronoun on the basis of salience ranking. The salience score is based on a list of syntactic properties. Since they do not have a full structural parse to use in computing the non-coreferentiality constraints, they make some assumptions using the partial phrases derived through a set of regular expressions. The pronoun resolution accuracy of this algorithm is 75% against third person pronouns with noun-phrase antecedents.

Mitkov implements a similar algorithm that contains a grab-bag of *factors* that are combined to determine antecedents [Mit97]. The factors are a combination of syntactic and semantic properties that each effect salience, such as topicalization constructions (there insertion and clefted it), subjecthood, technical terms within the domain under analysis (for example computer terms when the text under analysis is a computer user manual),

NP repetition, phrasal heads, distance from the pronoun. Each factor is implemented as a filter and is assigned a relative weight (weights can be positive or negative amounts). If a potential antecedent matches the factor criteria, its salience is incremented by the weight amount. For each pronoun, all DE tokens are passed through all the factors and the one that accumulates the most salience weight is chosen as the antecedent. An advantage of this algorithm is that it can judge salience in the first sentence of a text. However, as implemented it only resolves the pronoun *it*.

In SRI's FASTUS system [HAB+96], the search for a referent is constrainted by locality. Definite pronouns should find a reference within the preceding three sentences, but reflexives are limited to the current sentence. Candidates in the same and previous sentences are ordered left to right, then back further they are ordered right to left. This ordering is based on the observation that "fine grained syntax-based salience fades with time." (pg. 48)

The TRIPS system resolves definite pronouns using a focusing structure and very limited representation of common ground. Any demonstrative pronoun in the input is interpreted as a reference to the current plan. Definite pronouns are resolved against entities mentioned in the previous sentence or included in the previous plan action.

Corpus-Based Approaches An interesting recent trend in Computational Linguistics is the development of statistical techniques. One such algorithm, developed by Ge *et al.*, uses fully parsed Wall Street Journal Treebank texts [MSM93] and computes several statistical measures to determine pronoun bindings. The algorithm considers four factors: distance between candidate and pronoun, gender/number as computed by the training algorithm, noun phrase repetition, and collocation preferences between the noun and the phrasal head. The distance measure is computed based on the order in which Hobbs' naive algorithm visits nodes in the parse tree. After training these statistical measures on a test corpus, their algorithm correctly resolves 84.2% of pronouns correctly in a cross-validation experiment.

Using preferences gathered from manual annotation of over 3000 anaphors in naturally-occurring text, Rocha [Roc97] attempted to explicitly encode the relationship between anaphora resolution and topicality. The annotators collected a list of collocations which could be used in typical cases such as "I mean it". Using the annotation, a statistical model was built so that the best resolution strategy for each type of pronoun could be employed (i.e. lexical strategy, discourse strategy, etc.). His technique also employs semantic restrictions and takes into account the way verbs are used in each discourse. It is unclear from his publications how much of this algorithm was actually implemented.

Algorithms Specifically for Demonstrative Pronouns While some authors have claimed that demonstrative pronouns occur infrequently in natural language [Kam81; Pin86], we have seen that this is untrue for some corpora. A selection of Wall Street Journal Treebank texts annotated for coreference at Brown University contains 110 demonstrative pronouns in 1300 sentences (compared to 2026 definite pronouns in the same set). While this is a much lower concentration than in the TRAINS93 corpus, it is still a frequent enough phenomenon to merit analysis. Implemented algorithms for resolving demonstrative pronouns are rare. Luperfoy's thesis work included demonstrative pronouns but did not define

a separate resolution mechanism for them [Lup91]. Instead it used the same mechanism for definite and demonstrative pronouns.

Winograd's system [Win72, pg. 160] included a heuristic to interpret *that* after *do* as referring to the event most recently mentioned by anyone, while *it* was interpreted as referring to the event most recently mentioned by the same speaker.

Webber's model of discourse deixis handles demonstrative reference to discourse segments [Web88; Web90]. Because discourse deictic referring expressions *can* refer to arbitrarily large discourse segments, the first step in the process for dealing with a particular demonstrative reference is to determine whether it is referring to an entity or a discourse segment. To resolve the ambiguity in demonstrative reference to discourse segments, she resorts to pragmatic, contextual information for which she does not provide a computational account.

Passonneau [Pas93] extends the centering framework to distinguish between demonstrative and definite pronouns, since the original formulation treated all pronouns equally. She adds a notion of *local center*, as opposed to the C_b, which is stronger than the C_b. The local center establishment rule is: "Two utterances U_1 and U_2 that are adjacent in their segment establish an entity $\mathcal{E}$ as a local center only if U_1 contains a third person, singular, non-demonstrative pronoun N_1 referring to $\mathcal{E}$, U_2 contains a co-specifying third person, singular, non-demonstrative pronoun N_2, and N_1 and N_2 are both subjects or both non-subjects, in that order of preference."(pg. 204). This captures the cohesive effect of using *it* in two successive utterances - there is extremely high preference for the second *it* to corefer to the same entity.

As the previous discussion demonstrates, work within the computational linguistics community on pronoun resolution is currently dominated by algorithms that are cast in terms of focusing. Even algorithms that utilize grammatical function preferences or sentence structure often frame their approach in terms of calculating salience. A very important source of evidence for the correct referent is missing from these algorithms; the semantic attribution of the pronoun within its own context. While the attentional state of discourse participants certainly affects their use of pronominal referring expressions, it is not true that only focused objects can be the subject of pronominal reference. Disambiguating information in the sentence allows coherent pronominal reference to less-salient entities. For example, the first two sentences of example 25 establish Rex the dog as the center of attention. But the third sentence switches pronominal reference unambiguously to the cat. The third sentence might be judged to be slightly less coherent, and a human reader might momentarily assign Rex to *he*, but the correct interpretation is easy to derive before the end of the sentence.

(25)

1) My dog Rex was playing at the park today when the neighbor's cat ran by.
2) He doesn't usually chase cats but today he couldn't resist.
3) Luckily he's a really fast cat and he got away.

Adequate pronoun resolution algorithms must take into account the semantic information available in the construction that contains the pronoun. Because pronouns have so little semantic content of their own, speakers must provide information in the rest of the sentence to indicate what entity the pronoun refers to. Semantic information can be provided in copular constructions, such as "He was a fast cat", or in thematic role restrictions such as "I know that". This author's intuitive theory, unsupported by empirical evidence, is that the more potential a specific pronoun has for being ambiguous, the more semantic information the speaker will pack into the sentence in order to make sure that his reference is understood. This semantic content should be used in any pronoun resolution algorithm. Kameyama put it well when she said that the correct use of salience is "to identify the ... *default order that gives rise to preferred interpretations in neutral contexts.* Note that this default order alone does not *determine* interpretations of pronominal elements. Rather, its role ... is to give an ordered list of referents (centers) so that commonsense inferences can be controlled."[Kam86, pg. 201].

Approaches which rely solely on calculations of focus are particularly problematic for resolving discourse deixis, since the abstract entities are never in focus. An algorithm that seeks to resolve discourse deictic pronouns must utilize semantic information to first recognize that a pronoun is being employed for discourse deixis, then to choose the correct referent from the vast number of possible abstract referents.

4 Proposed model

We propose a new model of discourse context in which DE tokens for abstract entities are built during parsing. By incorporating knowledge of semantics into the pronoun resolution method and restricting access to abstract entities only in certain semantically-marked constructions, the problem of spurious referents for pronouns, mentioned by Hobbs, can be minimized. The overall goal of this research is to design a pronoun resolution algorithm that resolves pronouns against both mentioned and abstract entities. A secondary goal is to resolve both demonstrative and definite pronouns in a way that utilizes linguistic theory on the contrast between these two types of expressions. We know of no extant model that makes use of the contrasting pragmatics of definite and demonstrative pronouns as a basis for defining a pronoun resolution algorithm.

At this time we have designed an initial algorithm which will be used as a starting point in the research. The algorithm has been hand-simulated against a set of TRAINS93 dialogs and performs well above baseline. We have also implemented two large test components: 1) a testbed platform for agile comparison of different pronoun resolution algorithms within a plug-and-play architecture and 2) a machine learning layer that can run experiments within the testbed for comparing different pronoun resolution strategies. Once the baseline technique has been implemented, the testbed can be used in conjunction with the machine-learning driver routine to optimize our pronoun resolution algorithm. We plan to use the testbed in a variety of machine learning experiments and discover whether any particular algorithm performs consistently best across language domains or whether the algorithm must be trained separately for each domain of discourse. A secondary, but more ambitious, goal is to explore whether machine-learning techniques can be incorporated into pronoun resolution at runtime to create an adaptive algorithm that changes its pronoun resolution strategy for each discourse.

Even though the evaluation corpus proposed for this project is a spoken dialog corpus, the model we develop is not intended exclusively for spoken language or for dialog. Features such as turn-taking, acknowledgments, repairs and prosody will not necessarily be used in the pronoun resolution process. We intend the model to be applicable to either written or spoken discourse.

As in our review of related work above, the description of our proposed model in this section falls into two main sections. First, we describe what discourse entities should be created from a particular stretch of text. Before the model can be implemented, many issues of discourse entity management must be worked out. At this point we have not chosen a model of discourse entity salience, how discourse structure should be reflected in the organization of the DE tokens, and when tokens should be deleted. Second, we describe what happens when a pronoun is encountered. The model we propose defines separate resolution algorithms for definite and demonstrative pronouns, and utilizes semantic classifications provided by the TRIPS parser.

In the following sections, we describe the model of discourse-entity creation and pronoun resolution that has been defined at this point. Keep in mind that this model is only a starting point and will surely evolve and change during the course of this project. The machine learning supervisor will be described in Section 5.2.

4.1 Discourse Entity Categories

In an ideal model, a Discourse Entity token would be created for every entity evoked by the discourse that could possibly be subsequently referred to with a pronoun. In addition to entities explicitly mentioned in a noun phrase, many tokens for discourse deictic entities would be generated. It should be clear from the above discussion of related work that it is the phrasing used to evoke entities into the discourse, not the referent's ontological status in the world, that is important to their status in the discourse context. Due to this observation, we partition DE's into three categories based on the amount of interpretation of the surface form involved to uncover the entity. The first category, *Explicitly Mentioned Entities*, includes all entities that are explicitly mentioned in the discourse in simple noun phrases. The entities themselves can be abstractions or concrete entities, fictional or real world entities, they can be previously known to the discourse interpreter or not. What matters is that they are mentioned in simple noun phrases. Most extant pronoun resolution methods include only entities from this category in the discourse context representation.

The second and third categories together comprise what we have been calling *abstract entities*. *Implicitly Mentioned Entities*[6] are pulled from the surface form of the discourse but not from simple noun phrases. Larger nominal constituents, such as sentential complements, and verbal constructions such as infinitives and gerunds, trigger entities in this category. The third category, which we call *Resultant State Entities*, contains entities triggered by the perlocutionary effect of the utterance. As the result of an utterance, tasks are performed, objects move from one location to another and arrive at certain times, addressees compute the overall speech act of the previous utterance, etc. An utterance impacts the world in a variety of ways, and each of them can be the referent of a pronoun in the subsequent utterance. Note that this category of referent is only available in systems that compute deep understanding of the input. Entities in this category are completely missing in a coreference model of pronoun resolution. We know of no other model of discourse context which divides the abstract entities into two different tiers in this way, and no implementation that includes tokens for either category of abstract entities in its discourse model.

Because entity categories are defined by surface form, one of the first major tasks in the proposed project is to establish an exact formalization of the rules for turning surface form constituents into tokens. This section shows some preliminary work with examples of tokens in each category. In general, rules must be created to choose which syntactic constituents and semantic classes of objects in the parse trigger the creation of DE tokens. The TRIPS parser is especially well suited for this because it already has semantic class information included in the grammar rules.

The categories themselves were developed based on the referent categories discovered during our annotation exercise (reported in Section 2). The purpose of the current study is not to develop a better ontology of abstract entity types, but rather to handle the pronominal reference behavior that actually exists in the corpus. The only referent found in the corpus but not covered in this model is the task instructions themselves, which are available

[6]Note that this category bears no resemblance to the 'implicitly focused objects' in Grosz's model [Gro81] which included entities related in the task space to explicitly mentioned entities.

for reference by virtue of their physical copresence and applicability to the problem solving task, but which have no linguistic trigger within the discourse.

The examples below show each type of entity in a sentence and the resulting DC tokens (in order to avoid clutter, not all tokens are shown for each example, only those required to demonstrate the type of object in the category and related tokens). This list may be expanded later, but for a first pass these are the categories needed most often for pronoun resolution, so this is the set I'll start with. At this point only discourse entities triggered within declarative statements have been investigated. Discourse entities operate slightly differently within questions, and the exact specification of how grammar rules for questions should be modified to generate discourse entities remains to be defined.

1. **Explicitly mentioned entities**

   (a) **Simple noun phrases**
   All simple indefinite noun phrases, subject to Karttunen's rules regarding indefinites, create an entity in this category. Other surface forms include definite, demonstrative, reflexive, and genitive third person pronouns, bare plural nouns (trains), quantified nouns (two engines), compound nouns (information systems), proper names (Archbishop McKinney), and mass nouns (sand).
   Ex: "The problem is there are five boxcars of oranges waiting at Corning".
   DC = {problem, boxcars, oranges, Corning}

   Note that in the current model we do not create tokens for entities referred to via first and second person pronouns, so there are no tokens for the dialog participants themselves or the generic use of *you* meaning *one*. Possessive NPs create two tokens, as in:
   Ex: 'Pat's dog'
   DC = {Pat, Pat's dog}.

   (b) **Indexicals** Simple indexicals for times, days, and dates, create an entity token in this category. Surface forms include adverbials such as 'now', 'today', and 'here' as well as nominal forms such as 'at 2.a.m.'.

   (c) **Constructed sets and subsets**
   Any mentioned entities in a clause that have the same semantic type should be combined into a set, regardless of whether they are mentioned in a conjunction. Surface forms include sequences (Elmira, Bath, Delta), conjunctions (Elmira and Bath), selected sets (two of the boxcars), partitives (some of the oranges), and quantified expressions (every available boxcar).
   Ex: "Engine 1 and Engine 2 are at Corning"
   DC = {engine1, engine2, set{engine1, engine2}}.
   Ex: "and then load some of the oranges."
   DC = {oranges, some oranges}
   Ex: "Pat ran into Chris and they went to lunch"
   DC = { Pat, Chris, set{Pat, Chris}}

Ex: "Pat is Chris's roommate".
DC = { Pat, Chris, set{Pat, Chris}}

2. **Implicitly mentioned entities**

   (a) **Actions**
   This category contains any mentioned action, whether it is stated as a proposal for future action or as a report of past action doesn't matter. Actions, as opposed to events, have identifiable agents. Jackendoff's diagnostic test for actions versus events are: an action can fill in the x of 'what (agent) did was x' and an event can fill in the x of 'what happened was x' [Jac83]. Surface forms for actions include verbals (go to Corning), and *do* constructions.
   Ex: "Then go to Avon"
   DC = {Avon, action(go to Avon)}

   (b) **Action Types**
   When an action is mentioned in the abstract instead of a particular instance of the action, a token is created for the action type. At this point I'm not sure how to tell whether each mention of an action is a token or a type.
   Ex: "To go from Bath to Corning takes three hours."
   DC = { Bath, Corning, actionType(go from Bath to Corning)}

   (c) **Events**
   Events have no agent but may or may not have a time duration. For example, arriving at a destination, completing a task, having a thought are all events with no time duration, but a thunderstorm is an example of an event with time duration.
   Ex: "When we get to Elmira, we'll pick up the boxcars."
   DC = {event(arrive at Elmira), event(complete attaching boxcars)}.

   (d) **Event Types**
   In this category are generic kinds of events rather than a specific instance of an event. I'm unsure at this point what are the typical surface forms for event types.
   Ex: "For Chris to be late is unusual." DC = {eventtype(Chris is late)}

   (e) **Composite entities**
   This category includes semantically heterogeneous entities mentioned as aggregates. They are typically mentioned as prepositional phrases, such as "boxcars of oranges" but can also be conjoined sets of different semantic types such as "a boy and his dog".
   Ex: "The train loaded with oranges"
   DC = {train, oranges, composite(train with oranges)}

   (f) **Facts/Propositions**
   Propositions are denotations of sentences. They may, for example, be expressed in first-order logic. Facts are simply propositions that happen to be true, so they do not need to be in a separate category as far as discourse entities are concerned. It is unclear at this point whether I can define surface constituents to trigger entities in this category or whether the trigger comes from the logical form. At

a minimum, each full utterance which is of type Statement (in DAMSL[AC96] terminology) triggers a token for the propositional content of the utterance *in toto*.
Ex: "An engine can only carry two loaded boxcars." DC = { fact(max loaded boxcars on one engine = 3)}

(g) **Routes**
In our task-oriented domain, routes are mentioned quite often. The conceptual path created by a particular instance of traveling from one place to another creates an abstract object, the route used to get there.
Ex: "Engine e1 goes from Corning to Bath."
DC = {engine1, Corning, Bath, route(Corning to Bath)}

3. **Resultant State Entities**

(a) **Constructed Entities**
This category includes the object created by any action that results in the physical or conceptual attachment of explicitly mentioned entities. Entities in this category behave a little differently than entities formed by set construction, because in set construction the items are still conceptually separate, even though they can be referred to as a set. In this category, a new entity is formed from the individuals.
Ex: "Engine 1 goes to Corning and picks up the boxcars"
DC = {engine 1, Corning, boxcars, constructed{engine 1 and boxcars}}.
Ex: "My brother got married this weekend"
DC = {me, my brother, constructed{the wedded couple}}

(b) **Kinds**
This category includes the semantic type of mentioned objects. They are triggered by predicates in the logical form, for example Rottweiler(x), Red(x), President(x), etc.
Ex: "My dog is a Rottweiler."
DC = {me, my dog, kind{Rottweilers}}
Some proper names and compound entities trigger the creation of a KIND token.
Ex: 'Archbishop O'Malley'
DC = {Mr. O'Mally, Archbishops as a class}

(c) **Outcomes**
Outcomes are properties of mentioned entities that result from acting on the object. For example, moving a train from one city to another results in the state that the engine is at (ATLOC) the destination city.
Ex: "Engine e1 goes to Corning."
DC = {atloc(engine1, Corning)}

(d) **Speech acts of discourse segments**
For each utterance input to the parser, a determination of the speech act of the utterance is calculated. Possible speech acts are WH-QUESTION, SUGGEST, GREET, etc.

Ex: "Engine e1 goes to Corning."
DC = {engine1, Corning, speechact(suggest)}

(e) **Plan-level results and objects** This category includes resulting items at the task-level such as subplans, plans, arrival times, etc. A huge variety of result objects could potentially be created, in the initial model we only plan to create those necessary to support the evaluation corpus. For each task in the domain, its time duration and completion clock time are common referents of pronouns in the sample dialogs. Note that plans and tasks have similar accessibility for pronominal reference as discourse segments described by Webber, complete plans and open subplans along the right frontier of a tree representation of the task are available.
Ex: "Engine e1 goes to Corning."
DC = {engine1, Corning, planseg(e1 to Corning), planseg(entire plan with e1 to Corning as last action), endtime(e1 to Corning end time), timeduration(e1 to Corning time to complete)}

I want to make clear that these categories for DE tokens correspond to the surface form of the linguistic trigger. They are not meant to capture the semantics of the referent's counterpart in the world. Since the form of the linguistic trigger controls the category of the token, the same object could be mentioned in different ways resulting in different categories of tokens being created. For example in sentence 26, speaker A's contribution mentions an action, taking the oranges to Elmira, but since it is a verbal construction the token it creates would be an implicitly mentioned entity. However, when the pronoun *that* is used in speaker B's utterance to refer to the same action, an explicit mention DE token would be created.
(26)

A: "Take the oranges to Elmira."
B: "That's not gonna work."

This distinction can also apply to other token types, such as facts and events. Returning to Borthen's example, the statement "I heard that linguists get paid less than computer scientists" introduces an implicit entity, the fact of the pay differential. In the next sentence the same speaker could refer to that fact with a demonstrative but not with a definite pronoun. But if the sentence is re-phrased as "What do you think about the fact that linguists get paid less than computer scientists?", the noun phrase *the fact* triggers an explicit entity. Now the fact can be referred to with a definite or a demonstrative pronoun in the following sentence.

4.2 Discourse Entity tokens

4.2.1 When are tokens created?

Two possibilities exist for the creation of DE tokens: tokens can be added through an inference process when an anaphoric referring expression is encountered, or tokens can be produced as part of the interpretation process in preparation for subsequent anaphoric

reference. I prefer to create tokens as part of the interpretation process. This reduces pronoun resolution to a search through existing tokens. In this model, pronoun resolution is not responsible for generating additional inferred tokens.

4.2.2 What does a token look like

DE tokens do not contain full logical representations of their entities. Instead, they point to the system's representation of the entity if one exists. The DE token contains attributes that are important to the reference resolution process, some of which are not carried by the entity's logical representation. An example would be characteristics of the surface form such as surface position and the type of determiner used (if a demonstrative determiner, it's more likely to be pronominalized by a demonstrative than a definite).

In the current working model each token has the following attributes:

- **Semantic type** (in our implementation, a type from the TRIPS ontology). Explicitly mentioned entities may have multiple SEM values assigned by the TRIPS parser, for example the phrase *a problem* can have semantic types GOAL or PROBLEM, and the DE token would have the value (:SEM GOAL PROBLEM). Abstract entities have only one SEM type, for example if a RESULTANT STATE token is added for the end clock time of an action, it would be assigned (:SEM TIME-OF-DAY).
- **Category** One of the categories listed in section 4.1(Explicit/Implicit/Resultant State);
- **Pointer** to the full representation of the referent, if it exists elsewhere in the system. This might be in a data structure maintained by the planner or the world-state module. This link exists so that properties of the referent are available for reference disambiguation.
- **Form** For simple noun phrases only, whether it is Definite, Indefinite, or Demonstrative.
- **Position** of the linguistic trigger (start and end text position).

Other attributes are probably needed. For example, propositions may not have a persistent representation elsewhere in the system, so the entire propositional content should be saved on the DE. Each pronoun in the input utterance also triggers the creation of a token, but until it is resolved it has no semantic type or category. After the pronoun is resolved, the token is updated with those attributes from its referent.

4.2.3 Token organization

Previous models have used a variety of data structures to organize discourse entities. The organization of the entities should reflect the chosen representation of discourse structure and how it affects discourse entity accessibility. In our preliminary experiments, we used linear surface order to organize DE tokens on a stack. This is a very naive model and must be improved, but at this time no clear direction has been chosen. Calculation of discourse structure is an open problem.

4.2.4 Token deletion

Tokens from different categories have different lifetimes. In the current working model, the rules are as follows:

1. Implicit and Explicit Mentioned entities have unlimited lifespan, they are never removed from the DC. Since there is no delete operation for explicit entities, we make no assumption about the amount of intervening discourse between one mention of an entity and subsequent pronominal reference to it.

2. Resultant State entities are cleaned up after interpretation of the subsequent sentence, so only resultant state entities from the immediately previous sentence are available during interpretation of a particular pronoun.

Token lifetimes can be implemented either as additional filters in the search process during pronoun resolution, or the tokens can actually be cleaned up from the discourse context.

4.3 Pronoun resolution

Our model is similar to previous pronoun resolution algorithms in the fact that it creates DE tokens as part of discourse interpretation. As a result, pronoun resolution is simply a search through existing tokens for the correct referent. The task for pronoun resolution is simply to calculate the probability that each candidate in the DC is the correct referent, then output a ranked list of candidates and their associated confidence values. This technique is more flexible than choosing one referent to output. It supports an architecture in which subsequent processes might be invoked after parsing, and a different referent might ultimately be selected. For example, a process within the planner might apply world knowledge and select a different referent. After the pronoun is ultimately bound to its referent, the pronoun's discourse entity token is updated with attributes from the referent token.

The construction of our model is driven by the following observations on pronominal referring behavior:

1. Definite pronouns tend to refer to explicitly mentioned entities, demonstrative pronouns tend to refer to abstract entities.

2. For plural pronouns, demonstratives tend to be used for composite and constructed entities, constructed sets tend to be referred to via definites.

3. Explicitly mentioned entities can be the subject of long-distance pronominal reference, but abstract entities are only available for pronominal reference in the immediately following utterance. Of course, they are still available for anaphoric reference in later utterances via descriptive noun phrases but not via pronouns. For example, after A says "Pat got promoted yesterday.", B could say "It's happened before." (referring to the event), "That's a strange way to phrase it." (referring to the surface

construction), "That's a lie." (referring to the speech act), etc. But none of these anaphoric references are legal after intervening talk. Even after just a few intervening words, the accessibility of abstract entities seems to fade. For example, "I have a Harley. They are really powerful." works fine, but "I have a Harley that I rode to Nevada last year for my vacation. They are really powerful." seems less coherent.

4. Resultant state entities are never in focus, therefore they can only be referred to pronominally by explicit predication of the pronoun. In the above example, if we change the continuation to "They are really powerful bikes" adding the word *bikes* describes the referent and disambiguates the pronoun. This claim exploits the application of Grice's *maxim of quantity* to pronouns. The maxim says that contributions should be as informative as is required but not more informative than is required [Gri75]. As it applies to pronouns, the maxim means that speakers tend to elaborate the description of an entity just to the extent that will, in the speaker's judgment, allow the listener to identify it [Ols70]. For discourse deictic pronouns, this means that speakers tend to provide some semantic content in the sentence to help the addressee choose from among the many possible abstract referents.

4.3.1 Initial algorithm

We believe pronoun resolution should be modeled in two stages: 1) filter out candidate referents based on hard constraints such as binding constraints and agreement features, 2) choose among the remaining candidates based on soft constraints that are effectively calculations of salience. We are unclear about whether some factors known to influence referent selection should be implemented as soft or hard constraints. For example, numeric agreement for demonstrative pronouns seems to be looser than for definite pronouns. So should numeric agreement be implemented as a soft or hard constraint? That is why one of the main goals of this project is to implement an agile testing platform for pronoun resolution in which the exact configuration of the pronoun resolution strategy can be learned.

As we saw above in the discussion of related work, there are many factors that affect salience and many ways to filter out non-coreferential tokens. At this point we are unsure how to combine all these factors into an optimal strategy. Factors we plan to incorporate are:

- **Filters**
  - Based on observations by theoretical linguists, definite and demonstrative pronouns have different preference orders for mentioned entities versus abstract entities. Demonstratives can match either abstract or mentioned entities, but in most unmarked cases definites match explicitly mentioned entities.
  - If the pronoun is the subject of a copula, the semantic type of the proposition in the predicate of the sentence can be calculated and used as a filter. For example, in the sentence "It'll be three a.m.", the semantic type for *three a.m.* is TIME-OF-DAY, so only TIME-OF-DAY referents are valid candidates. In our initial model, we filtered out Resultant State Entities unless a semantic type could be calculated.

- If the commanding verb has a high-confidence semantic type restriction, filter out tokens that do not match that semantic type. For example the direct object of *load* must be either a COMMODITY or a CONTAINER.
- Apply number/gender/animacy filter. Definite pronouns must match the referent exactly. Demonstratives are only loosely marked for number; while *these/those* must match a plural entity *this/that* can match either a plural or singular entity.
- Plural demonstrative pronouns can be resolved to composite or constructed entities but plural definite pronouns can only be resolved to constructed sets.
- Because of the tendency to use demonstrative pronouns rather than definites for discourse deixis, filter out IMPLICIT or RESULTANT STATE tokens for definite pronouns unless the pronoun occurs in a semantically-marked construction.
- Use the discourse structure to filter out inaccessible referents.
- Apply Karttunen's criteria for short-term referents in modal and conditional contexts.
- Tense should also affect the accessibility of referents. Entities mentioned in perfective tenses cannot be referenced in other time frames. "I'm going to bake some muffins tomorrow. *They tasted great". I haven't found any other pronoun resolution algorithms that take this constraint into account. In the TRAINS93 domain, this issue comes up mainly in keeping track of different plan alternatives. Some alternatives exist in a hypothetical modality until they are accepted by the dialog participants, and referent accessibility needs to keep them separate from entities in the actual plan.
- The exact properties of items should be taken into consideration. For example, for the demonstrative pronoun in "That'll be two a.m. when we get to Corning" the referent should be the end-time of a plan segment that involves moving something from non-Corning to Corning.

- **Salience Calculation**
  Many factors have been shown to affect salience, some boost a particular discourse entity's salience and others decrease it. Some of the factors used in previous studies to calculate salience of explicit mention entities include:

  - Boost the salience of subjects of intransitive verbs and objects of intransitives.
  - Use the discourse structure to fade the salience based on the distance between the pronoun and the candidate referent. In the simplest model, this can just be based on linear text position.
  - Decrease the salience of (or maybe filter out?) referents in reported speech unless the pronoun is in the same reported speech.
  - Increase the salience of the first noun-phrase referent in each sentence.
  - Decrease the salience of referents in embedded constituents such as relative clauses and prepositional phrases.

Very little previous work exists on how to judge relative salience of the many abstract entity tokens triggered by a sentence. For task-oriented dialog, previous studies have shown that the structure of the task affects both the structure of the dialog and also the salience of potential referents for definite descriptions [Gro77]. We extend that observation by claiming that the task structure also affects the relative salience of abstract entities. However, it affects the salience of implicit entities and resultant state entities in different ways. For implicit entities, their salience ranking depends on the frequency of reference in the corpus. For example, actions are more likely referents in the trains corpus than composite entities. For resultant state entities, the effect is more subtle. A huge variety of resultant state entities related to a stretch of text could be the subject of anaphoric reference, but the specifics of the task definition make certain ones more salient than others. This salience depends on how the problem is posed. For example, task participants might be instructed to complete their task in a certain amount of time or by never visiting the same city twice. The participants' attention is thus drawn to that particular aspect of the problem, making that aspect the most salient discourse-deictic entity. After each subtask is complete, the most salient feature of that subtask is how it impacts the overall assigned goal. To illustrate, imagine that participants in the air-compressor repair domain are instructed to replace the pump in the shortest time possible. After each subtask, we might expect a reference to the amount of time it took, e.g. "That was three minutes". If instead, the goal is to use the fewest possible tools, a statement like "That was four" would be interpreted as a reference to the number of tools used. In our model, what this means is that the semantic value of predicates in copular constructions defaults to the semantic value of the most salient resultant state entity. Continuing with the above example, if the goal is to use the fewest number of tools, a statement such as "that was three" will be constrained to only match resultant state entities of semantic type TOTAL_TOOLS. In another task context, the exact same phrase would have a different semantic type filter. Phrases with the semantic type made explicit, such as "That was a total of three tools" are unchanged in the different task contexts.

Some of these factors may not be available for some language domains, for example the semantic type may not be available for unconstrained language. Therefore, the system must be built so that it is highly configurable and can disable any process not desired for a specific experiment. Each filter and salience factor will be implemented as an independent module.

After the pronoun is bound, it might be necessary to add semantic information onto the system's representation of the referent. For example, after the clause"I went to *my hair stylist* today", we cannot assign a gender to the hair stylist, but after the continuation " and *she* said..." we know the correct gender. This should block the hair stylist from being bound to any future masculine pronoun.

4.3.2 Open Issues

There are many parameters within pronoun resolution yet to be explored. Among the open questions are:

- Conflicting opinions have surfaced in the literature regarding whether demonstrative

noun phrases are used to refer to an item in the current focus of attention. How does this apply to demonstrative pronouns? Perhaps an additional filter should be implemented so that demonstrative pronouns are never allowed to refer to the most salient DE token.

- Should sets be constructed for multiple DE tokens of the same semantic value, no matter what category they are in? For example, if more than one action is implicitly mentioned in an utterance, should a set token be created for them?

- How can world knowledge from the planner or problem solver be incorporated? For example, the planner might be able to determine the thing most likely to be taken to Corning based on its knowledge of the task goal.

- If an implicitly mentioned entity is pronominalized in a subsequent utterance, are items related to the entity brought back into focus? For example, if a Kind token is pronominalized should the salience of the mentioned entity that generated it be increased?

- Should demonstrative pronouns be allowed to refer to constructed sets with homogeneous semantic values, or should they only be allowed to refer to composites and constructed entities?

These and similar questions can be answered experimentally.

4.3.3 Learning an optimal strategy

In order to build a pronoun resolution strategy that can be adapted to different language domains, the factors impacting filtering and salience will each be implemented as separate plug-and-play modules. This allows the system to be configured at run time. For example, if the current domain has no semantic categories, that module can be disable for a particular run of the system. This architecture has already been implemented in a pronoun resolution testbed system [Byr99b]. Due to its highly-configurable design, we can run the pronoun resolution system underneath a machine-learning layer. The machine learning layer can selectively enable or disable modules and can combine results from different modules into an optimal overall classification. Using machine learning, we can discover an optimal strategy for many subtasks within pronoun resolution, for instance how to combine multiple factors to calculate salience, when (if) to delete tokens of a specific category, what is the optimal model of discourse structure, etc. All the open questions about the model listed above in Section 4.3.2 will be the subject of machine learning experiments. It will be interesting to discover whether a uniform strategy emerges for pronoun resolution in different language genres and domains, or if different strategies work best in each domain. If the second is true, this project proposes to design an end-to-end method and toolset requiring no additional programming by the end user for discovering the optimal pronoun resolution strategy in his domain. If time permits, we also hope to explore whether learning processes can be embedded within the pronoun resolution component to support adaptive algorithms that improve their strategy at run time. Another advantage of this flexible architecture is that

widely-accepted baseline strategies, such as Hobbs or Lappin and Lease, can be implemented and applied to different corpora. This makes it simple to compare any pronoun resolution method we design to published baseline algorithms on a common corpus.

At this point we have performed experiments with two machine learning approaches.

Genetic Algorithms technique A genetic algorithm layer was built on top of the testbed [BA99a; BA99b] to calculate salience for candidate referents. Because syntactic, semantic, and pragmatic factors impact the salience of referents for pronouns, some method must be devised to weigh different evidence from different linguistic processes and combine it in a way that chooses the correct referent the largest portion of the time. This seems like the perfect application for a genetic algorithm.

The genetic algorithm uses a labeled training corpus to find an optimal combination of weights for different salience factors. In the initial experiments (discussed below), we attempted to improve on the combination of anaphora resolution factors reported in [Mit97]. The algorithm can be used to determine the optimal strategy for any open issue in our model. It can be configured to learn different strategies for each pronoun individually or all pronouns as a class.

Data Mining technique An association mining algorithm [LMA98] was run over our TRAINS93 annotation data described above in Section 2 and reported in [BA98]. The study found that single attributes, such as subjecthood or form of the linguistic trigger, could not be reliably used to predict the form of the pronoun; but taking attributes in combination can be highly predictive. These results can be put to use in statistical preferences as filters or as salience preferences. We intend to experiment with using these statistical patterns in combination with other factors to see if it gives us any lift in resolution accuracy.

4.4 Hypotheses of this study

The hypotheses to be tested by the proposed model are:

- A domain-independent pronoun resolution method can be built (either in its details or in the method used to learn the details).
- An optimal model contains different strategies for definite and demonstrative pronouns.
- Separating abstract entities into Implicit and Resultant state entities produces better pronoun resolution than having only one category.
- Pronoun resolution algorithms must support a wide range of abstract entities in order to get truly high accuracy (on all the pronouns).
- For definite noun phrases with noun-phrase antecedents, the model we propose performs at least as well as existing algorithms which do not have abstract entities represented in the discourse context.

4.5 Benefits of this model

By including all the filters on token accessibility discussed above, and also information from linguistic theory on the contrastive behavior of definite versus demonstrative pronouns, I believe that this model cleverly manages to add abstract entities as candidate referents for pronouns in a way that avoids increasing the search space of referents for any particular pronoun.

As Hobbs mentioned, previous authors have avoided adding abstract entity tokens to their lists of candidate referents because of the fear that pronoun resolution would be severely degraded. Unless the increase in discourse entity tokens is offset by improved filtering and selection among candidate referents, pronoun resolution accuracy would suffer. But by adding the additional filters in our model, definite pronouns with simple noun phrase antecedents (the ones that would be resolved correctly by previous models) will still be resolved correctly by our model. By only choosing abstract entities in semantically-marked constructions, our model should result in a net increase in pronoun resolution accuracy. The second major benefit of this model is a net increase in the variety of pronoun phenomena that can be handled: we can resolve demonstrative pronouns.

4.6 Out of scope

As is necessary in any thesis project, huge regions of the problem space have been excluded. It would be nice to have a model that covers the entire space of anaphoric reference, but since we have not investigated many anaphoric terms such as *back*, *much* and one-anaphora those cannot be incorporated into the model at this time. Also excluded are definite description anaphora. Although any implementation we build must include resolution of anaphoric definite descriptions, we don't have any new observations to inform this process, so the existing mechanism within the TRIPS parser will be used.

For pronominal anaphora, an entire class of objects is missing from our model. Pronouns can be used to perform several types of bridging reference, but to resolve them seems to require inference upon encountering the pronoun rather than preparation of a token in the DC. At this point we do not have enough data in our corpus studies to have an intuition for how this should be handled. An example would be "Chris had a hamburger and a coke, and I had *that* too" in which *that* refers to my hamburger and coke, not the one belonging to Chris mentioned in the sentence. Another example is "I love Russia because *they* are so friendly there" where *they* means the Russian people. Resolution of such pronouns requires real-world knowledge and complex inference upon encountering the pronoun, which this model seeks to avoid. When to trigger such inference versus binding the pronoun to a token already in the DC remains a mystery. Notice that in a coreference model of evaluation, our algorithm would be credited with resolving this example correctly as long as the pronoun is co-indexed with the correct linguistic trigger.

Also, since the model does not employ real-world or common-sense knowledge, ambiguous pronouns such as those in Sidner's vet/vaccination examples will not be resolved properly.

A complete implementation would model the relationship between the discourse model and long-term memory. As predications of DEs are communicated in a discourse, some updates from the discourse model to long-term memory should take place. Time constraints will probably not permit that process to be included in this project.

It should be noted that pure deictic use of demonstrative pronouns, such as pointing to a person and saying "That's my cousin." is not covered in this study. All pronouns in the evaluation corpus refer to an entity with a linguistic trigger in the discourse. The model does not attempt to provide a computational mechanism for handling deixis to physically-copresent entities.

Another type of anaphora not handled by the model is a quirk of spoken task-oriented dialogs. In the TRAINS93 dialogs it is common for speakers to refer to task-domain objects with first person pronouns "After *we* get to Corning, it'll pick up the boxcars" where *we* = *it* = engine1. It seems theoretically possible to calculate when a first or second person pronoun is being used in this way; however, the current project includes no plan to do so.

4.7 Spoken Dialog Phenomena

Our evaluation corpus, the TRAINS93 dialogs, is a spoken dialog corpus, therefore the current model needs to be able to work in the context of spoken dialog understanding. A variety of dialog phenomena might be modeled. We intend to use the machine learning supervisor to discover the correct interaction between these factors and other factors in the system. Because of the flexible architecture, modules which compute these factors can simply be disabled for interpretation of text, so the resulting algorithm is not limited to spoken language.

- **Turn-taking** behavior affects the ordering of discourse entities. Should the beginning of each turn constitute a new utterance, or should constituents that get split across turns be re-attached?

- **Repairs** Our algorithm accepts input from the parser, so it handles repair phenomena to the same extent that the existing parser does. It is sometimes possible to have a referent in the reparandum, so discourse entity creation will need to be built into the meta-grammar rules for repairs.

- **Acknowledgments** Previous authors have found that acknowledgments affect discourse entity creation in dialog [CM81; NS94; SE99], but we are unclear exactly how this should be implemented. Also, the published results of those authors were hand-simulated evaluations, so those authors were not able to experiment with a wide variety of models and choose the optimal one. We hope to use machine learning and experimentation to determine how this should be modeled.

- **Speaker Alternation** Our model has so far glossed over the issue of whether two speakers in a dialog maintain separate discourse contexts or whether one discourse context can be built that represents the common ground of both discourse participants. Our study last year [BS98] indicated that speaker alternation probably doesn't matter.

However, this project will explore whether the speaker of a DE token should be taken into consideration during the pronoun resolution process.

- **Overlapping speech** Overlapping speech must be incorporated into any model that calculates surface order of constituents in spoken dialog. However, a corpus study of the TRAINS93 and SWITCHBOARD dialogs (Strube personal communication) revealed that speakers do not refer anaphorically to items whose evoking reference is within overlapping speech. The constituent containing the referent is always repeated after overlap ceases before anaphoric reference is used.

- **Partial parses** The integration of discourse entity creation into the TRIPS chart parser side-steps this issue. The chart parser calculates the most probable parse, and the grammar rules used then trigger discourse entities.

- **Prosody** It is incontrovertible that prosodic stress affects pronoun binding for human interpreters. However, calculation of prosodic features by a spoken language understanding system remains an open problem. The TRIPS system currently does not calculate prosodic features, and since we have no desire to add it to the system it remains outside our model.

5 Preliminary results

5.1 Resolving Abstract entities

A preliminary version of the algorithm was hand-simulated on a corpus of 9 TRAINS93 dialogs containing 205 pronouns. The referent for each pronoun had been annotated in the corpus study described above in Section 2^7. In this test set, 68% of demonstratives and 29% of definite pronouns are discourse deictic.

The DC for this simulation included the following subset of tokens from our model:

1. Explicitly mentioned entities
2. Constructed and composite entities
3. Proposition tokens representing the entire propositional content of each utterance
4. Actions
5. Time-duration of actions
6. End-time of actions

A sample DC is shown in Table 5. This example assumes an overall task goal to perform the task in a certain amount of time, therefore tokens for the end time and time-duration of actions are included. If some other dimension of the task were more important, for

Object	Class	Semantic Type
engine1	EXPLICIT	MACHINE-COMPONENT
Corning	EXPLICIT	CITY
boxcars	EXPLICIT	CONTAINER
engine 1 and boxcars	IMPLICIT	CONTAINER MACHINE-COMPONENT
action picking up the boxcars	IMPLICIT	ACTION
end-time of action	RESULTANT STATE	TIME-OF-DAY
time-duration of action	RESULTANT STATE	TIME-DURATION
Dansville	EXPLICIT	CITY
action going to Dansville	IMPLICIT	ACTION
end-time of action	RESULTANT STATE	TIME-OF-DAY
time-duration of action	RESULTANT STATE	TIME-DURATION
action entire plan so far	RESULTANT STATE	ACTION
end-time of action	RESULTANT STATE	TIME-OF-DAY
time-duration of action	RESULTANT STATE	TIME-DURATION
oranges	EXPLICIT	SOLID-COMMODITY
engine 1, the boxcars and the oranges	IMPLICIT	SOLID-COMMODITY CONTAINER MACHINE-COMPONENT
action loading the oranges	IMPLICIT	ACTION
end-time of action	RESULTANT STATE	TIME-OF-DAY
time-duration of action	RESULTANT STATE	TIME-DURATION
action entire plan	IMPLICIT	ACTION
end-time of action	RESULTANT STATE	TIME-OF-DAY
time-duration of action	RESULTANT STATE	TIME-DURATION
propositional content of utt (top-of-stack)	RESULTANT STATE	PROP

Table 5: DC after the sentence "With engine 1 in Corning, pick up the boxcars and go to Dansville and load the oranges."

example to deliver the maximum amount of oranges to Corning, the algorithm would create a RESULTANT STATE token for the amount of oranges in Corning after each action.

Both demonstrative and definite pronouns used the same search order, simply searching backwards in linear text order for the first matching referent. However, the filtering implemented was different for demonstrative and definite pronouns:

Resolving Definite Pronouns

1. If the pronoun is the subject of a copula, calculate the semantic type of the proposition in the predicate of the sentence. The first token encountered which matches that semantic type is chosen. For example, in the sentence "It'll be three a.m.", the semantic type for 'three a.m.' is TIME-OF-DAY, so only TIME-OF-DAY referents are accessible. In some cases, the system cannot calculate a semantic type for the predicate and this filter does not apply.

2. Otherwise, only EXPLICITLY MENTIONED ENTITY tokens are available for definite pronouns.

[7]Raw inter-annotator agreement for referents was 79%, so the algorithm cannot be expected to do better.

	Number of Demonstratives Resolved Correctly	Number of Definites Resolved Correctly
Our algorithm	59	59
Baseline algorithm	9	52
Total	(need to get this off spreadsheet)	

Table 6: Summary of Pronoun Resolution for Initial Algorithm

Resolving Demonstrative Pronouns

Demonstratives are only loosely marked for number; while *these/those* must match a plural entity *this/that* can match either a plural or singular entity. We incorporated the observations of previous authors summarized in Section 3, resulting in the following algorithm for resolving demonstrative pronouns:

1. Same as filter 1 for definites.
2. Otherwise, plural demonstratives resolve to the first composite object encountered. Singular demonstratives match the first implicit or explicit token found, regardless of its number.

Notice that for both definite and demonstrative pronouns, Resultant State Entities were not accessible unless filter 1 applied. Implicit Mention Entities were available to demonstrative pronouns but not definite pronouns when filter 1 did not apply.

In hand-simulation, our algorithm correctly resolved 57.6% of the 205 pronouns in the test set (see Table 6). The baseline used for comparison is a simple 'most recent NP' technique that searches the DC and resolves the pronoun to the most-salient EXPLICIT entity with matching number. The baseline technique resolved only 29.8% of pronouns in the test set correctly, significantly lower than our algorithm. Our algorithm improved resolution accuracy for both demonstrative and definite pronouns. Only five pronouns in the test set were resolved correctly by the NP-only technique but incorrectly by our algorithm.

5.2 Genetic Algorithms experiment

Ten different factors affecting DE salience were implemented and a genetic algorithm supervisor layer was constructed. The modules implemented were inspired by a number of previous studies:

- Decrease salience of quoted speech [Kam98]
- Decrease salience of indefinite NPs [Mit98]
- Decrease if in relative clause or prepositional phrase [KB96]
- Increase salience of subjects [BFP87]
- Increase salience of the antecedent selected by Hobbs algorithm

Most-recent	Mitkov	Hobbs naive	Genetic
47%	**52.2%**	**67.8%**	**69.1%**

Table 7: Pronoun resolution accuracy on the test corpus

- Increase salience of the most recent antecedent that matches number/gender
- Increase salience of the first noun phrase in each sentence [Mit98]
- Decrease salience in proportion to the distance from the pronoun
- Decrease salience if in the predicate of the sentence

Input to the program is:

- $\vec{W} \mid 0 \leq W_i \leq 15$ is the weight assigned to module$_i$
- $\vec{C}$ is the vector of candidate referents
- The pronoun to be resolved

The experiment here was to learn the optimal way to combine all these different measures of salience into one overall ranking. For each pronoun a module might update the salience score of zero, one or all candidates by the amount of its weight. After all voters execute, the highest-scoring candidate is chosen as the antecedent. In case of a tie, the most recent candidate is chosen. $\vec{W}$ is generated by the genetic algorithm using random numbers for the first generation, then standard mutate, crossover, and replicate operations for subsequent generations. Each individual's fitness is the percent of pronouns resolved correctly. The initial population size is 15, and after each generation the five most fit individuals are allowed to reproduce, halting after twenty generations.

Our evaluation corpus was 3900 sentences from the Treebank corpus [MSM93] for which noun-phrase antecedents of definite pronouns were annotated [GHC98]. 70% of the corpus was used to train the genetic algorithm, the remaining 30% (containing 519 pronouns) was the test corpus.

Table 7 shows pronoun resolution accuracy for our four experiments against the test corpus. In order to compare our results with Mitkov's published results, we ran the following set of voters:

- Increase salience for objects of certain verbs such as *(discuss, present, illustrate)*
- Decrease salience of indefinite NPs

- Increase salience of the first noun phrase in each sentence
- Decrease salience in proportion to the distance from the pronoun
- Increase salience for items in the first clause of a complex sentence
- Decrease salience of objects of prepositions
- Increase salience for lexically repeated items

We were unable to implement three of his voter modules that were specific to his corpus. One was to increase salience of items mentioned in the section heading or chapter title of the text containing the pronoun. The second was to increase salience of domain terms, such as computer terms when the text being processed was a printer manual. This is inappropriate for our evaluation corpus since it is not domain-restricted. The third module that we did not implement was collocation patterns, since our parser does not isolate phrasal heads at this time. The absence of collocation patterns was an obvious handicap to the algorithm, but I don't think they would be as powerful in our corpus of unrestricted news articles as they were in Mitkov's evaluation corpus of computer manuals. This is because the language used in news text seeks to avoid repetition of any lexical patterns, whereas computer instruction manuals are written in a very formulaic way and strong patterns for subj/verb and verb/object pairs should emerge. Our implementation of Mitkov's factors resolved only 52% of the test pronouns correctly. It achieved exactly the same results running with the same weights used in Mitkov's experiments as with a very different set of weights learned by the genetic algorithm driver.

The 'most-recent' technique correctly resolved only 47%. Hobbs' algorithm, which uses syntactic structure, improved to 67.8%. The genetic algorithm correctly resolved 69.1%, a slight improvement over Hobbs. The genetic algorithm could certainly be improved by calculating impossible coreference, for instance by using binding constraints. This is just one example of how learning algorithms can be applied to pronoun resolution. Besides learning weights for salience factors, the algorithm might choose an optimal technique for fading old tokens, learn whether to implement a particular factor as a filter or a contributor to salience, or choose between several alternate models of discourse structure, to name just a few.

6 Project Plan

6.1 Implementation

The model described above in Section 4 will be integrated into understanding component of the TRIPS system. This involves modification in four main areas: the parser, the interface between the parser and the discourse manager, the discourse manager, and reference resolution. We propose to integrate discourse entity creation into the TRIPS parser grammar rules, which are well-suited for such an integration because they already contain semantic types where applicable. All tokens are created after the syntactic parse is completed, and pronouns will be resolved from left to right so that all DE tokens to the left of a pronoun are available when it is resolved. The TRIPS parser currently creates a structure of mentioned objects to pass to reference resolution. This interface may be modified to add additional attributes for DE tokens. The semantic categories within the TRIPS ontology are probably adequate, it is unclear at this time whether changes to the ontology will be needed to support this project.

Some simplification of the model may occur during implementation. For example, reference to speech acts is rare in the TRAINS93 dialogs, so generation of speech act DEs may not be implemented.

The flexible pronoun resolution architecture will be integrated into the reference resolution component of the TRIPS parser, but it will retain its capability to run under different driver layers, for example to test Treebank sentences instead of Trains sentences. The testbed currently contains around 20 different modules for calculating DE salience.

6.2 Evaluation

The evaluation corpus will be sentences taken from the TRAINS93 dialogs in which the pronouns have already been annotated for referents. We will use transcripts which have speech repairs removed. It may not be possible to get complete parses for all sentences in the test dialogs. In that case, I will build a corpus of small dialog fragments containing all the utterances from the linguistic trigger to the pronoun (inclusive). I may have to clean up disfluencies in these utterances and/or modify the TRIPS lexicon or grammar rules to improve parser coverage on this corpus.

It may also be possible to do some evaluation in the Monroe County domain but that will have to be determined after the system is up and running and I can see what type of language is used in that domain. The Monroe County domain would provide a good test of how extensible this method is to other task-oriented domains.

An evaluation will also be conducted on the selection of Wall Street Journal Treebank texts that we received from Charniak, however, this is just meant as a comparison and no performance objective is proposed for that corpus.

Pronoun resolution algorithms in prior studies are typically evaluated based on the percent of pronouns in a corpus for which the correct antecedent was identified by the algorithm. Dividing the percent correct by the total number of pronouns gives the percent accuracy of the algorithm. On the surface, this measure seems straightforward, but in fact

subtle details in its definition from one study to the next make comparison of pronoun resolution strategies difficult.

The most basic issue is what to consider a 'pronoun' in the discourse. Many authors offer no precise definition, so it is difficult to determine which surface forms are included. In most papers, a 'pronoun' is a singular, definite, third-person pronoun, excluding *one*.

Another very basic issue is how to count the total number of pronouns in the discourse. For most authors, the only pronouns included in the total are the ones that their algorithm is designed to handle in the first place. Pleonastic pronouns are excluded, as well as those with non-NP triggers and references to the time and weather. This makes it very difficult to determine what percent of the total pronouns in a discourse were resolved correctly.

For our evaluation, we include singular and plural definite and demonstrative third-person pronouns in the total pronoun count. We will exclude pleonastic (non-referential) and cataphoric pronouns. A referential pronoun is one for which the word *what* can be substituted and an answer can be found. For example, the sentence "It's three p.m." can be converted to the question "What's three p.m." and the answer is "The current time is three p.m.". However in the sentence "When it comes to trucks, I would probably think to go American" (switchboard 2326), the question form doesn't make any sense: "When what comes to trucks?", which means that it should be considered pleonastic. Besides non-referential pronouns, our total pronoun count excludes some referential pronouns including *one*, first and second person pronouns, and relative pronouns.

The second thing that must be defined is how to determine whether a particular pronoun was resolved correctly by the algorithm. I take issue with the manner in which this is done in most studies, although I concede that they sometimes apply criteria which are appropriate for a shallow-understanding system but which would not be appropriate in TRIPS. Most reported studies use a coreference model, linking the pronoun to its linguistic trigger in the surface form. As long as the algorithm selects the correct surface constituent as the linguistic trigger of the pronoun, it is considered correct. This model is inadequate for a system with more ambitious understanding capabilities. Such an evaluation fails to distinguish between the multiple discourse entities that can be generated by one surface constituent, for example distinctions between individual entity tokens and kinds are lost. For example, in one study [ES99] the utterance "[I don't trust them]$_i$, maybe I guess it$_i$'s because of ..." was considered to be processed correctly as long as the algorithm marked *it* as coreferential with the phrase *I don't trust them*, even though the phrase can be interpreted as a relationship between two entities, a speech act, a fact, etc. I believe evaluation performed in this way measures the wrong thing. People don't use anaphoric terms to refer to other words, they use them to refer to conceptual entities, so an evaluation procedure which ignores the existence of conceptual entities, the referents themselves, is inherently flawed. Ideally, accuracy should be judged on whether the replacement of the pronoun with the referent results in the correct interpretation of the construction containing the pronoun.

Having said that, it is also not fair to expect pronoun resolution to achieve deeper understanding than does the system as a whole. If the entire system is designed, for example, simply to determine topicality of a noun phrase by calculating mention counts and nothing beyond syntactic analysis of the text is performed in the first place, a coreference model of pronoun resolution is appropriate.

Also, pronoun resolution evaluation should not penalized by errors made by the system in interpreting the linguistic trigger. For example, if the system misunderstood a noun phrase as a reference to an individual rather than as a reference to a kind, this error will be propagated to coreferential pronouns. If a pronoun is resolved to the correct token, whether or not the system's representation for that entity is correct, the pronoun resolution should be considered accurate.

In the current project, pronouns will be judged as accurate if the correct discourse entity token is selected as the referent. Since surface constituents trigger multiple entities, measuring coreference among surface constituents is not adequate. However, a certain amount of under-specification will be allowed. For many pronouns in this domain, a scoping ambiguity exists. For example, "The engine picks up some oranges, then take *it* [ibid] to Corning." In this example it is unclear whether the speaker had in mind the engine, the cargo, or the entire train as the referent for *it*. However, for the purposes of the planning task, the choice between those three items is irrelevant because all three candidates are physically connected. They are all going to Corning together. In our evaluation, the algorithm will be judged as correct if it chooses any of the three, as long as the referent results in the correct interpretation of the utterance at the plan level.

6.3 Project milestones

- Data collection:
  1. Derive test corpus for TRAINS93 sentences.
  2. Annotate answer key for non-np pronouns in Ge's Wall Street Journal corpus. The corpus currently has only pronouns with coreferent noun-phrases marked. Joel Tetreault has written a parser for this corpus that creates input data for the testbed.
- Formalization of Model
  1. Finalize the categories of DE tokens to be included. Create new SEM values within the system if necessary.
  2. Pin down which grammar rules in the TRIPS parser create which tokens. Create new grammar rules if necessary.
  3. Update the interface between parser output and discourse manager. Not sure yet what this entails.
  4. Determine the model of discourse structure to be implemented.
- Implementation
  1. Add lexical items to the TRIPS system if necessary to handle the test corpus.
  2. Modify DM processing to manage new classes of tokens and to attribute them in the way spelled out here.
  3. Code the rules for pronoun resolution in the reference module.

4. Add Kartunnen-like analysis of explicit entities into the parser.
5. Integrate testbed code into reference resolution (in test system build only, not demo version of the system).
6. Run learning experiments to resolve open issues in the model.
7. Modify the planner or problem solving component to send resultant state entities to the discourse manager.
8. As time allows: add analysis of verb modality to the attribution of discourse entities created by the parser.

- Evaluation

1. Evaluate on the test corpus built in phase 1. The Performance Goal to reach is completely automatic resolution accuracy of 70% or more for abstract entities and 80% or more for domain entities.

6.4 Project schedule

Major project phases are scheduled to be complete on the following dates:

Semester	Project Deliverables
Summer 99	Data collection complete
	TRAINS93 Evaluation Corpus built
	Integration of initial model into the TRIPS system
Fall 99	Model Formalization complete
Spring 00	Model Formalization,
	begin implementation
Summer 00	(internship off campus)
Fall 00	Complete Implementation
Spring 01	Evaluation,
	Thesis writeup,
	look for job

Table 8: Project schedule

References

[AC96] James Allen and Mark Core. Draft of damsl: Dialog act markup in several layers. Available at http://www.cs.rochester.edu/research/trains/annotation, 1996.

[AHG98] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Extending a simple coreference algorithm with a focusing mechanism. In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pages 15–27, 1998.

[All95] James Allen. *Natural Language Understanding, 2nd edition.* Benjamin/Cummings, Menlo Park, 1995.

[Ari90] Mira Ariel, editor. *Accessing Noun-Phrase Antecedents.* Routledge, 1990.

[Azz96] Saliha Azzam. Resolving anaphors in embedded sentences. In *Proceedings of the 34^{th} Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 263–269, 1996.

[BA98] Donna K. Byron and James F. Allen. Resolving demonstrative pronouns in the trains93 corpus. In *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pages 68 – 81, 1998.

[BA99a] Donna K. Byron and James F. Allen. A genetic algorithms approach to pronoun resolution. In *submitted to AAAI-99*, 1999.

[BA99b] Donna K. Byron and James F. Allen. A genetic algorithms approach to pronoun resolution. Technical Report 713, Department of Computer Science, University of Rochester, 1999.

[Bal97] Breck Baldwin. Cogniac: High precision coreference with limited knowlege and linguistic resources. In *Proceedings of the ACL-97 workshop: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, 1997.

[BFG97] Kaja Borthen, Thorstein Fretheim, and Jeanette K. Gundel. What brings a higher-order entity into focus of attention? sentential pronouns in english and norwegian. In Ruslan Mitkov and Branimir Boguraev, editors, *Operational Factos in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 88–93. Association for Computational Linguistics, Association for Computational Linguistics, 1997.

[BFP87] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. In *Proceedings of the 25^{th} Annual Meeting of the Association for Computational Linguistics (ACL '87)*, pages 155–162, 1987.

[Bot96] Simon Phillip Botley. Comparing demonstrative features in three written english genres. In *Approaches to Discourse Anaphora: Proceedings of the Discourse*

Anaphora and Resolution Colloquium (DAARC96), pages 86–105. University of Lancaster Centre for Computer Corpus Research on Language - Technical Papers Volume 8, 1996.

[BS98] D. Byron and A. Stent. A preliminary model of centering in dialog. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98)*, 1998.

[Byr99a] Donna K. Byron. Analysis of pronominal reference in two spoken language collections: Trains93 spontaneous task-oriented dialog and Boston University radio news prepared monologue. Technical Report 703, Department of Computer Science, University of Rochester, March 1999.

[Byr99b] Donna K. Byron. A flexible architecture for reference resolution. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, to appear 1999.

[CB88] Jaime G. Carbonell and R.D. Brown. Anaphora resolution: a multy-strategy approach. In *COLING '88*, pages 96–101, 1988.

[Cha70] Wallace L. Chafe. *Meaning and the structure of language.* The University of Chicago Press, 1970.

[Cha72] Eugene Charniak. Toward a model of children's story comprehension. Technical Report AI TR-266, Massachusetts Institute of Technology, 1972.

[Cha80] Robert Channon. Anaphoric *that*: A friend in need. In Jody Kreiman and Almerindo E. Ojeda, editors, *Papers from the Parasession on Pronouns and Anaphora*, pages 98–109. Chicago Linguistic Society, The University of Chicago Classics, 1980.

[Cho81] Noam Chomsky. *Lectures on government and binding.* Foris Publications, Cinnaminson,N.J., 1981.

[CM81] Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie Lynn Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 10–62. Cambridge University Press, 1981.

[DI90] Ido Dagan and Alon Itai. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, pages 330–332, 1990.

[Dor90] Joke Dorrepaal. Discourse anaphora. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, pages 95–99, 1990.

[ES99] Miriam Eckert and Michael Strube. Resolving discourse deictic anaphora in dialogs. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, 1999.

[FA98] George Ferguson and James F. Allen. Trips: An intelligent integrated problem-solving assistant. In *Proceedings of AAAI '98*, 1998.

[Fil82] Charles J. Fillmore. Towards a descriptive framework for spatial deixis. *SPeech, Pland and Action*, 3(4):31–59, 1982.

[GHC98] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.

[GHZ93] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

[GJW83] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL '83)*, pages 44–50, 1983.

[GJW86a] Barbara J. Grosz, Karen Sparck Jones, and Bonnie Lynn Webber, editors. *Readings in Natural Language Processing.* Morgan Kaufmann Publishers, 1986.

[GJW86b] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Towards a computational theory of discourse interpretation. unpublished ms, 1986.

[GJW95] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[Gri75] H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, pages 41–58, New York, 1975. Academic Press.

[Gro77] Barbara J. Grosz. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '77)*, pages 67–76, 1977. Reprinted in [GJW86a].

[Gro81] Barbara J. Grosz. Focusing and description in natural language dialogues. In Aravind K. Joshi, Bonnie Lynn Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 84–105. Cambridge University Press, Cambridge, 1981.

[GS86] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[Gui85] R. Guindon. Anaphora resolution: short term memory and focusing. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL '85)*, pages 218–227, 1985.

[HA95] Peter A. Heeman and James F. Allen. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, 1995.

[HAB+96] Jerry R. Hobbs, Douglas E. Appelt, John Bear, David Israel, MEgumi Kameyama, Mark Sticket, and Mabry Tuson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing.* MIT Press, 1996.

[Haw91] John A. Hawkins. On (in)definite articles: implicatures and (un)grammaticality prediction. *Journal of Linquistics*, 27, 1991.

[Hei82] Irene Heim. *The Semantics of Definite and Indefinite Noun Phrases.* PhD thesis, University of Massachusetts at Amherst, 1982.

[Hob86] Jerry Hobbs. Resolving pronoun reference. In *Readings in Natural Language Processing.* Morgan Kaufmann, 1986.

[HS96] Udo Hahn and Michael Strube. Incremental centering and center ambiguity. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society (CogSci '96)*, 1996.

[HS97] Udo Hahn and Michael Strube. Centering in-the-large: Computing referential discourse segments. In *Proceedings of the 35^{th} Annual Meeting of the Association for Computational Linguistics (ACL '97)*, pages 104–111, 1997.

[Isa74] Stephen Isard. Changing the context. In Edward L. Keenan, editor, *Formal Semantics of Natural Langage.* Cambridge University Press, 1974.

[Jac83] Ray Jackendoff. *Semantics and Cognition.* Current Studies in Linguistics. MIT Press, 1983.

[JS89] Robert J.P.Ingria and David Stallard. A computational mechanism for pronominal reference. In *Proceedings of the 27^{th} Annual Meeting of the Association for Computational Linguistics (ACL '89)*, pages 262–271, 1989.

[Kam81] H. Kamp. A theory of truth and semantic representation. In Groenendijk et al., editor, *Formal Methods in the Study of Language*, 1981.

[Kam86] Megumi Kameyama. A property-sharing contraint in centering. In *Proceedings of the 24^{th} Annual Meeting of the Association for Computational Linguistics (ACL '86)*, 1986.

[Kam97] Megumi Kameyama. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL-97 workshop: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, 1997.

[Kam98] M. Kameyama. Intrasentential centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Clarendon, Oxford, 1998.

[Kar76] Lauri Karttunen. Discourse referents. In J. McKawley, editor, *Syntax and Semantics, vol. 7*, pages 361–385. Academic Press, 1976.

[KB96] C. Kennedy and B. Boguraev. Anaphora in a wider context: Tracking discourse referents. In *ECAI-96*, 1996.

[Keh97] Andrew Kehler. Current theories of centering for pronoun interpretation: A critical evauation. *Computational Linguistics*, 23(3):467–475, 1997.

[KR93] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. D. Reidel, 1993.

[Lak74] Robin Lakoff. Remarks on *this* and *that*. In *Papers from the Tenth Regional Meeting of the Chicago Linguistics Society*, pages 345–356, 1974.

[Lan86a] Fred Landman. Data semantics for attitude reports. *Linguistics and Philosophy*, pages 157–183, 1986.

[Lan86b] Fred Landman. Pegs and alecs. *Linguistics and Philosophy*, pages 97–155, 1986.

[LDC99] *MUC7 Coreference Answer Key*. Science Applications International Corporation, 1999.

[Lin79] C. Linde. Focus of attention and the choice of pronouns in discourse. In Talmy Givon, editor, *Syntax and Semantics 12: Discourse and Syntax*, New York, 1979. Academic Press.

[LL94] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, pages 535–561, 1994.

[LM90] Shalom Lappin and Michael McCord. A syntactic filter on pronominal anaphora for slot grammar. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL '90)*, pages 135–142, 1990.

[LMA98] Neal Lesh, Nat Martin, and James Allen. Improving big plans. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '98)*, 1998.

[Lup91] Susann LuperFoy. *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*. PhD thesis, University of Texas, 1991.

[Mit97] Ruslan Mitkov. Factors in anaphora resolution: they are not the only things that matter. In *Proceedings of the ACL-97 workshop: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 14–21, 1997.

[Mit98] Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of ACL '98*, pages 869–875, 1998.

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[NS94] David G. Novick and Stephen Sutton. An empirical model of acknowledgement for spoken-language systems. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '94)*, 1994.

[Nun79] G. Nunberg. The non-uniqueness of semantic solutions. *Linguistics and Philosophy*, 3(2):142–184, 1979.

[Ols70] D. Olson. Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77:257–273, 1970.

[Pas89] Rebecca J. Passonneau. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL '89)*, pages 51–59, 1989.

[Pas93] Rebecca J. Passonneau. Getting and keeping the center of attention. In Madeleine Bates and Ralph M. Weischedel, editors, *Challenges in natural language processing*, pages 179–226. Cambridge University Press, 1993.

[Pas95] Rebecca J. Passonneau. Integrating gricean and attantional constraints. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '95)*, pages 1267–1273, 1995.

[Pin86] Manfred Pinkal. Definite noun phrases and the semantics of discourse. In *Proceedings of the 11th International Conference on Computational Linguistics (COLING '86)*, pages 368–373, 1986.

[PiWB98] Massimo Poesio, Sabine Schulte im Walde, and Chris Brew. Lexical clustering and definite description interpretation. In *Papers from the 1998 AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 82–89. AAAI Press, 1998.

[Poe94] Massimo Poesio. *Discourse Interpretation and the Scope of Operators.* PhD thesis, Department of Computer Science, University of Rochester, 1994.

[Pri91] Ellen F. Prince. The zpg letter: Subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins B.V., 1991.

[PT97] Massimo Poesio and David Traum. Conversational actions and discourse situations. *Computational Intelligence*, 14, 1997. (forthcoming).

[Rei85] Rachel Reichman. *Getting Computers to Talk Like You and Me.* MIT Press, Cambridge, Mass., 1985.

[Res92] P. Resnik. A class-based approach to lexical discovery. In *Proc. of the 30th annual meeting of the Association for Computational Linguistics (ACL-92)*, pages 327–329, Newark, Delaware, 1992.

[Roc97] Marco Rocha. Supporting anaphor resolution in dialogues with a corpus-based probabilistic model. In *Proceedings of the ACL-97 workshop: Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 54–61, 1997.

[Sch77] R. C. Schank. Rules and topics in conversation. *Cognitive Science*, 1977.

[Sch85] Rebecca Schiffman. *Discourse constraints on 'it' and 'that': A study of language use in career-counseling interviews.* PhD thesis, University of Chicago, 1985.

[Sch88] Ethel Schuster. Anaphoric reference to events and actions: A representation and its advantages. In *Proceedings of the 12^{th} International Conference on Computational Linguistics (COLING '88)*, pages 602–607, 1988.

[SDG+99] Amanda Stent, John Dowding, Jean Mark Gawron, Elizabeth Owen Bratt, and Robert Moore. The commandtalk spoken dialog system. In *Proceedings of the 37^{th} Annual Meeting of the Association for Computational Linguistics (ACL '99)*, June 1999.

[SE99] Michael Strube and Miriam Eckert. Attentional state and dialog. In *review*, 1999.

[SH96] Michael Strube and Udo Hahn. Functional centering. In *Proceedings of the 34^{th} Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 270–277, 1996.

[Sid77] Candace L. Sidner. Toward a computational theory of definite anaphora comprehension in english. Technical Report AI-TR-537, Massachusetts Institute of Technology, 1977.

[Sid83a] Candace L. Sidner. Focusing and discourse. *Discourse Processes*, 6, 1983.

[Sid83b] Candace L. Sidner. Focusing in the comprehension of definite anaphora. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*, pages 363–394. MIT Press, Cambridge, 1983. Reprinted in [GJW86a].

[SM94] Linda Z. Suri and Kathleen F. McCoy. Raft/rapr and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics*, 20, 1994.

[Str96] Michael Strube. Processing complex sentences in the centering framework. In *Proceedings of the 34^{th} Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 270–277, 1996.

[Str98] Michael Strube. Never look back: An alternative to centering. In *Proceedings of ACL '98*, pages 1251–1257, 1998.

[Wal96] Marilyn A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264, June 1996.

[Web79] Bonnie Lynn Webber. *A Formal Approach to Discourse Anaphora.* Garland Publishing, 1979.

[Web81] Bonnie Lynn Webber. Discourse model synthesis: preliminaries to reference. In Aravind K. Joshi, Bonnie Lynn Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 283–299. Cambridge University Press, 1981.

[Web88] Bonnie Lynn Webber. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL '88)*, pages 113–122, 1988.

[Web90] Bonnie Lynn Webber. Structure and ostension in the interpretation of discourse deixis. Technical Report MS-CIS-90-58, Department of Computer and Information Science, University of Pennsylvania, 1990.

[Web99] Bonnie Lynn Webber. *The Handbook of Discourse Analysis*, chapter Computational Aspects of Discourse and Dialogue. Blackwell Publishers Ltd., (to appear) 1999.

[Win72] Terry Winograd. *Understanding natural language.* New York: Academic Press, 1972.

[WJP98] M. Walker, A. Joshi, and E. Prince, editors. *Centering Theory in Discourse.* Clarendon Press, Oxford, 1998.