Emily G. Brassell. Demonstrative Anaphora: Forms and Functions in Full-Text Scientific Articles. A Master's paper for the M.S. in I.S. degree. April, 2000. 63 pages. Advisor: Robert Losee.

This study examines the functions and characteristics of demonstrative anaphora (*this, these, that, those*) in a collection of full-text scientific documents, confirming that they play an important role in maintaining discourse focus and binding together cohesive sections of text. Unlike corpora in other subject domains, the *Cystic Fibrosis* database contains more demonstrative expressions than any other class of anaphora. As participants in intersentential reference, demonstratives often refer to complex propositions rather than simple noun phrases. While this tendency complicates automated resolution, our results yield some suggestions toward a resolution algorithm. Primarily, we argue for the incorporation of demonstrative form since different types of demonstratives show different patterns regarding antecedent length and composition. Although further analysis is necessary, our findings provide a groundwork for future exploration.

Headings:

Linguistics – anaphora

Natural Language Processing

Scientific Literature

DEMONSTRATIVE ANAPHORA:
FORMS AND FUNCTIONS IN FULL-TEXT SCIENTIFIC ARTICLES

by
Emily Brassell

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in Information Science

Chapel Hill, North Carolina

April, 2000

Approved by:

____

INTRODUCTION

The incorporation of natural language understanding into information processing systems has been the subject of much exploration. Accounting for natural language properties could greatly enhance the ability to represent the "aboutness" of documents, painting a more detailed picture of the hierarchy of purposes present in a text. Unfortunately, encoding the complexities of natural language has proven extremely difficult, partially due to the inclusion of many ambiguities. Among these ambiguities is *anaphora*, abbreviated reference to an entity previously introduced in a text. Generally items are described most fully when introduced (in the *referent*, *antecedent* or *correlate*), allowing readers to form a mental representation. Subsequent references (*anaphors*) may contain less detail because they need only remind readers of concepts already present in consciousness. Frequently an anaphor takes the form of a pronoun:

`Before Anne leaves,` **`she`** `will eat some tasty cheddar cheese.`

but it may also consist of a noun phrase (NP):

`Pass Anne` **`the cheese`**`.`

As in the first example, both anaphor and antecedent may occur in the same sentence (*intrasentential* or *bound* anaphora), or like the second case (where "the cheese" refers back to "some tasty cheddar cheese"), anaphora may cross sentence boundaries (*intersentential* or *discourse* anaphora).

Automatic resolution of anaphora is a complex problem which has been attacked from numerous angles. However, intersentential references, trickier to resolve than their intrasentential counterparts, have only recently come under scrutiny. Moreover, most resolution algorithms treat either personal pronouns (*he, she, it, they*) or definite descriptions (*the mayor, the cow*), ignoring other types of anaphors. However, demonstrative pronouns (*this, that, these, those*) play an important role in anaphoric reference, appearing frequently in many collections and often referring to integral content. In the sample of full-text medical articles investigated in this study, demonstrative expressions account for more than half of the instances of anaphora, and their antecedents contain keywords more often than any other type of anaphor.

Since the first step in deriving a resolution algorithm is the establishment of heuristics governing anaphoric behavior, we have attempted to form a comprehensive picture of demonstrative anaphora in a corpus of full-text documents. To capture the general trends of anaphoric distribution in the collection, we examined around 330 instances of demonstrative anaphora in nineteen different documents. About 70% of these instances are intersentential, with the majority of demonstrative anaphors referring to concepts expressed in the preceding sentence. The most common pattern places an antecedent of phrase length (i.e., three or more words) in the first sentence of a paragraph, followed by an anaphor in the second sentence. Most frequently, the anaphor consists of either *this* or *these* employed as adjectives (e.g., *this group, these side effects*). Distal demonstratives (*that, those*) are much rarer, as are antecedents composed of multiple sentences.

Although our observations have not produced concrete rules for resolving demonstrative anaphora, simply locating anaphors may yield clues about the semantic structure of a document. The presence of an anaphor indicates that a concept is important enough to bear repeating, implying that most concepts integral to a text will be expressed anaphorically at some point. A document may contain a series of anaphoric references, with each reference (or sequence of references) describing a separate subtopic. Thus a sentence which contains no intersentential anaphors – and thus has no anaphoric connections to preceding sentences -- may indicate the introduction of a new subject. While major changes in topic should coincide with paragraph and section boundaries, we hypothesize that the distribution of anaphora in a document may indicate more subtle changes in subtopics. At the very least, anaphora may supplement the guides provided by structural elements like section titles, paragraph changes, etc. Toward this end, we have extracted anaphors and index terms from a small sample of documents, creating skeletal representations of the text by displaying the location of these terms within paragraphs (see Appendix B). Although these rough representations have not yielded conclusive patterns, they provide an interesting alternate view of document content.

APPLICATIONS OF ANAPHORIC RESOLUTION

Automated anaphoric resolution has implications for a variety of text-processing tasks, including passage extraction for question-answering and automatic abstracting systems. The structure of full-length texts can be viewed as "a sequence of subtopics set against a backdrop of one or two main topics;" anaphoric references may prove useful in tracking these subtopics (Hearst & Plaunt, 1993). Generally the introduction of a new subtopic precedes a series of anaphors, implying that sentences without anaphors often introduce key information. Indeed, Johnson et al. (1997) found that these *propositional sentences* often summarize important points and are excellent candidates for inclusion in an abstract. Bonzi and Liddy confirm that around 60% of anaphors in scientific abstracts refer to concepts central to the document’s topic (although certain categories of anaphors are more likely than others to refer to integral ideas) (1989). Correspondingly, Bonzi found that 49% of keywords in these abstracts had anaphoric references, as opposed to about 22% of non-keywords (1991).

Identification of integral ideas facilitates the extraction of key passages from documents. O’Conner (1973) extracted relevant passages from documents to provide specific answers to queries. He believes that his results might be improved by locating expressions referenced anaphorically -- specifically those referred to by demonstrative anaphors. Similarly, Paice extracted key sentences from full-text documents to create abstracts but found that dangling anaphoric references left the abstracts incomprehensible (cited in Liddy, 1989). Thus Johnson et al. (1997) continued the work by developing criteria to identify propositional sentences which contain no unresolved anaphors or connectors. These sentences often introduce information integral to the text and precede a series of sentences referring to the same concept. Accordingly, tracing anaphoric references (in this case, definite noun phrases) back to their original source may yield a list of the document’s most important concepts.

In information retrieval, automatically resolving anaphora could improve query analysis and contribute to the refinement of matching algorithms. Liddy et al. (1987) point out applications for query analysis, citing work by Belkin, Oddy, and Brooks which used “superficial statistical methods to analyze and represent relationships between concepts mentioned in queries.” These queries, transcripts of oral utterances, contain numerous anaphora; undoubtedly resolution would impact the representation of concepts.

As retrieval systems become capable of handling longer, more complex queries, anaphoric resolution could become increasingly important.

In the most comprehensive study on the subject to date, researchers at Syracuse University examined whether anaphoric resolution would improve representation of the “aboutness” of a document by gathering benchmark data on anaphora in scientific abstracts and then examining the impact of resolution on retrieval performance (Bonzi & Liddy, 1989; Liddy, 1989; Bonzi & Liddy, 1988; Liddy et al., 1987). They compared retrieval results for a number of queries before and after resolution of anaphora in 600 scientific abstracts extracted from PsycINFO (behavioral science) and INSPEC (engineering and computer science). Mixed results show that resolution may improve retrieval results, have no impact, or (in rare cases) actually impair performance; outcomes differed according to anaphoric class, document collection, and term-weighting formula. Judgments provided by human experts also indicate that the tendency to reference integral concepts differs according to anaphoric class; demonstrative pronouns and pro-adverbials were most likely to refer to integral ideas (1988). However, there was not a strong correlation between a term’s centrality to the document and its increase in term weight due to anaphoric resolution. Bonzi and Liddy conclude that anaphoric resolution should not be implemented indiscriminately; only certain anaphoric classes should be resolved, and they should be applied only to certain term weighting formulas (1988). Moreover, while resolution will certainly change term weighting scores, these changes may only increase scores for terms that already occur much more frequently than less important terms (Bonzi, 1991). The study thus concludes that increased accuracy in term frequency scores does not improve retrieval sufficiently to warrant anaphoric resolution in their collection of abstracts. Instead, efforts on resolution should look beyond formulas concentrating on individual terms to address discourse-level issues. Resolution may be more useful in representing the relationships among concepts in a text than in representing the concepts themselves (Bonzi & Liddy, 1988).

Using first manufactured queries and then genuine information needs, Pirkola and Jarvelin investigated anaphoric resolution using Boolean and proximity searches in a full-text database of Finnish newspaper articles (1996). They classified anaphora according to their antecedents' linguistic class, differentiating between proper names and common names and between basic words, compound words, and phrases. Results in both studies favored resolution of anaphora referring to proper names (recall increased from 10.8% to 17.6% in the first study).   Specifically, the names of people were more influential than those belonging to organizations or events, leading to a 40% increase in recall. Resolution of other classes of anaphora, on the other hand, had little effect. The researchers attribute this result to the fact that news stories often focus on individuals, making their names central to the text. They note the necessity of exploring anaphora in different subject domains, explaining that proper names probably occur more frequently in news articles than in scientific documents.

LINGUISTIC PROPERTIES OF ANAPHORA

Essentially, anaphora involves subsequent reference to an entity mentioned previously in a discourse (where *discourse* is a coherent section of written or spoken text). Technically, the referent must precede the anaphor; the opposite case is known as cataphora, as in:

> `These` are `the best oranges.`

In the quintessential case of anaphora, a pronoun replaces a simple noun phrase:

> When I waved goodbye to **Henry**, **he** obligingly waved back.

However, numerous expressions may serve as anaphors (although no expression is inherently anaphoric). Likewise, referents may encompass verbal phrases, clauses, or even complex sections of text, in addition to nominal phrases. Verbal substitution, for example, involves the verb *do*, often followed by *so*, *it*, *this*, or *that*:

> If you don't **take the garbage out** now, you'll have to **do it** later.

Larger segments of text may also be referenced anaphorically. Hirst (1981) gives an extreme example where an entire chapter of a history textbook is summarized in one word:

> **Such** was the France to which Coucy returned in 1367 (p. 14).

While we generally equate anaphora with abbreviation, *definite descriptions* offer stylistic variation without a shortened reference. However they generally do not introduce new information about an entity (*epithets* are the exception to this rule), and the linguistic context supplies ample information to identify the referent:

> A man came up behind John and hit him on the head. John turned round to face **his assailant.** (Carter, p .42).

While *assailant* is more specific than *man*, the anaphor only reiterates knowledge gleaned from the previous sentence; we know that John is the victim and the nameless man is his attacker without having to access information outside the discourse context.

The function of an anaphoric expression may extend beyond simply replacing an antecedent. Indeed, an anaphor may reference an antecedent without invoking the exact same entity. The classic example of this phenomenon, known as *identity of sense* *anaphora* (Hirst), *descriptional anaphora* (Webber) or *surface anaphora* (Allen), is often called the "paycheck sentence:"

> The man who gave his paycheck to his wife was wiser than the man who gave **it** to his mistress.

While *it* references *his paycheck*, the pronoun clearly refers to the second man's paycheck, rather than the wiser man's; antecedent and anaphor invoke two different instances of the same type of object. As with the function of replacement, in this capacity anaphoric expressions may substitute for noun phrases, verbal phrases, or clauses. With nominal phrases, the anaphor is often *one(s)*, *the first*, *the former*, *the latter*, as in:

> The red pants look better than **the green ones.**

More difficult to resolve are *associative* (Dorrepaal) or *strained* (Hirst) anaphora. Here the discourse provides context, but the referent is not explicit:

> We drove by the house. The windows were dirty (Dorrepaal, p. 4).

Clearly *the windows* belong to *the house*, but the two noun phrases do not refer to the same instance of an entity, or even to the same type of entity, but rather to two related entities. As with the cases of substitution explored above, this entity may be more complex than a simple noun phrase.

Here we reach the border of anaphora. Strictly speaking, anaphora must refer explicitly to a segment of text; an expression that alludes to an entity implied (but not explicitly defined) by the text belongs to the phenomenon of *deixis* (sometimes called *exophora*). Identifying the referents of deictic expressions requires knowledge outside the linguistic environment, while information supplied by the linguistic context suffices for anaphora. In its simplest conversational incarnation, deixis supplements the gesture of pointing. A deictic expression identifies an entity in the (non-linguistic) environment:

> Look at ***that hideous rat***!

However, in written texts deictic expressions usually refer to events or propositions arising from the discourse. Webber and Lakoff call this *discourse deixis*. In the following example from the our corpus, the demonstrative pronoun refers to an event described by the previous sentence. Note that the antecedent could not simply replace the anaphor but would have to be transformed into a nominal expression to preserve the grammaticality of the sentence.

> As a specimen was tested successively at frequencies corresponding to Zone 1, Zone 1/2 and Zone 2, viscosity decreased and elasticity increased. **This** may be because the relative importance of viscosity and elasticity in determining the rheological behavior of the sample alters with increasing shear rate.

The relationship between anaphora and deixis is the subject of much dispute, and the broad range of terminology used to describe them complicates the situation. Indisputably the phenomena serve different purposes; from the cognitive perspective, for example, Cornish (1999) argues that deixis brings an entity to the addressee's attention, whereas the use of anaphora presupposes that the reader's attention is already focused on that entity. While theoretical linguists carefully differentiate between anaphoric and deictic reference, researchers concerned with computational linguistics often allow overlap. Since the fundamental problems of automatic resolution remain inextricably intertwined in the two cases, many computational linguists define anaphora to encompass both intratextual and exophoric reference:

> An anaphor is an incomplete expression which depends for its interpretation on some other element in the sentence or context (de Swart, 1998, p. 12).

This broader definition often proves sufficient, but in certain situations it remains useful to make distinctions. In an overview of all types of anaphora in the CF database, we chose not to distinguish between anaphoric and deictic expressions (although we did note whether a reference was intratextual or exophoric). However, demonstrative expressions prove to be a particular sort of beast, and their propensity to have complex, abstract referents makes the distinction useful.

Demonstrative Expressions

Demonstrative expressions   composed of   *this*, *that*, *these*, *those* and accompanying noun phrases and modifiers   constitute a special category of anaphora. Like anaphora, demonstratives can be classified according to many different schemes. To begin, they may be distal (that, those) or proximal (this, these). Traditional linguists make many other distinctions; the following discussion borrows Himmelmann's (1996) list of the characterizations found in linguistic literature (p. 219).

Formal criteria. On a formal level, demonstratives may be used pronominally or adnominally (as adjectives), and they may comprise simple or complex noun phrases.

Activation state. The selection of an appropriate determiner or pronoun depends partly upon discourse focus; different anaphors are appropriate for entities that are the major subject under discussion and peripheral entities. Gundel, Hedberg, & Zacharski (1992) propose that different terms "signal different cognitive statuses (information about location in memory and attention state);" the occurrence of a particular term helps the addressee limit possible referents to those in the appropriate cognitive state (p. 274). The selected term should be as informative as required, but no more informative than necessary (Gundel, 1996). When these criteria are violated and a demonstrative determiner appears where the definite article would be sufficient, the author intends some special effect or implication. Gundel, Hedberg, & Zacharski's Givenness Hierarchy shows the relationship between the reader's cognitive status and the authors' word choice, with focus diminishing from left to right:

in focus >	activated >	familiar >	uniquely identifiable >	referential >	type identifiable
{*it*}	*that* *this* *this* N	{*that* N}	{*the* N}	{indefinite *this* N}	{*a* N}

Use of an indefinite article assumes only that the addressee recognizes a type of entity, rather than a specific instance; to use Gundel et al.'s example, the addressee of the sentence *A dog next door kept me awake* references a mental representation of the entity "dog" without specifying a particular canine. However, when the demonstrative *this* is used in a referential sense, *This dog kept me awake* implies not only the existence of some dog, but indicates that the author has a particular dog in mind. Use of the definite article -- *The dog kept me awake* -- presumes that the addressee can unique identify the specific dog, either from the author's description or from previous experience. Substituting *that* for *the* informs the addressee that s/he is already familiar with the dog in question. Demonstrative expressions may also indicate an "activated" referent (one that is "readily accessible to consciousness"), informing the addressee that the referent has recently been mentioned or is immediately accessible outside the linguistic context (Gundel p.145). Finally, use of the personal pronoun *it* *It kept me awake* signals that the referent is already the focus of the addressee's attention. Thus demonstrative expressions are reserved for a certain range of focus and generally signal familiarity with a particular instance of an entity.

Ariel's (1996) Accessibility Theory functions similarly. Ariel assumes that a reader identifies a referent by searching a mental list of possibilities and selecting the entity which has the appropriate degree of cognitive accessibility. (As discussed in the subsequent section on automated resolution, this perspective lends itself to discourse-level algorithms.) The Accessibility Marking Scale ranks expressions from most to least accessible:

> zero < reflexives < agreement markers < clitcized pronouns < unstressed pronouns < stressed pronouns < stressed pronouns + gesture < proximal demonstrative (+NP) < distal demonstrative (+NP) < proximal demonstrative (+NP) + modifier < distal demonstrative (+NP) + modifier < first name < last name < short definite description < long definite description < full name < full name + modifier

Accessibility depends on several factors: the antecedent's salience (i.e., topicality or centrality), its recency of mention, and cohesion between clauses containing antecedent and anaphor. Expressions with high accessibility correspond to antecedents which are highly salient, recently mentioned, and occur in cohesive units. Demonstrative expressions once again occupy the center of the scale, reserved for referents on the fringes of discourse focus.

Referent type. Byron and Allen (1998) note that demonstrative expressions, "ambiguous as to scope," enjoy a wider range of referents than definite pronouns (p. 2). In addition to single words, demonstratives may refer to discourse segments of varying lengths, or to the propositional content of these segments. Myers (1988) proposes that the range of adnominal demonstratives surpasses that of pronominal demonstratives: "the pronoun nearly always refers to a proposition expressed or implied in the previous sentence, while the [determiner + noun] can refer to a proposition expressed or implied in any immediately preceding segment, even in the entire text up to that point (cited in Cornish, 1999, p. 59)." In fact, Ariel finds that nearly 60% of demonstrative pronouns have referents in the preceding sentence (1996). Himmelmann (1996) agrees that pronominal demonstratives are governed by more restrictions than adnominals and thus are used less frequently.

In spoken English, *that* is often used when a speaker hesitates, unable to quickly choose the correct substitute for a complicated referent. In fact, the more complex a referent or its context, the more likely the speaker is to use *that* (Byron & Allen). Sidner (1983), Kameyama (1986) and Passonneau (1993) agree that *that* may remention an entity without returning it to the center of attention (Byron and Allen, p. 3).

Discourse Function. As mentioned previously, linguists traditionally distinguish between anaphora (reference to entities present in the text or utterance) and deixis (reference to entities and concepts outside the discourse, requiring contextual information for interpretation). While demonstratives may be used in either fashion, they are the archetypical means of "pointing" to an object not explicitly mentioned in the text but recognizable to both author and reader. Webber's informal analysis of pronouns in scientific texts and newspaper articles exhibits a typical distribution; here demonstrative pronouns account for 84% of deictic expressions but only 2% of references to nominal phrases (1991).

As anaphors, demonstratives indicate a particular referent among those already present in the discourse context. As deictic expressions, they establish a referent in the discourse context by "pointing" at it for the first time. In the latter case, demonstrative expressions may mention an entity outside the linguistic context (*situational* or *exophoric* use) or reference a proposition or event occurring within the text (*discourse deictic* use). Situational use occurs frequently in oral discourse, where a demonstrative expression may indicate an object that is literally present ("Look at that crazy cat!") or describe a certain measure or distance ("The man was about this tall."). In our collection, situational use generally involves self-reference (i.e., "this article" or "this study").

While discourse deixis involves concepts within the linguistic context, it does not strictly replace a segment of text. Instead it refers to an object, event, proposition, or some other occurrence whose existence is implied by the text. For example:

> It's always been presumed that when the glaciers receded, the area got very hot. The Folsum men couldn't adapt, and they died out. **That's** what is supposed to have happened. It's the textbook dogma. (Webber, 1991, p. 107)

In this excerpt, *that* does not refer to any specific NP or segment of text; rather it refers to an occurrence -- something that could *happen* -- described by first two sentences. To emphasize the difference, Webber distinguishes between the *demonstratum* (what the deictic expression points to) and the *referent* (what the deictic expression refers to). In the above example, the demonstratum consists of the first two sentences, while the referent is the event they describe. In other cases the demonstratum and referent may be the same entity. Regardless, a *referring function* can be defined to explain their relationship; this function is applied to the demonstratum to produce the referent. Herein lies the complexity of discourse deixis -- the referent is actually created by the fact of reference (Himmelmann, 1996). As with situational use, the act of pointing draws the referent into the linguistic context. Presumably this could create enormous difficulties for automatic resolution. After all, one of the most popular strategies relies on first establishing a list of possible referents and then eliminating items from this list until one possibility remains. The situation is ameliorated, at least, by the fact that deictic pronouns always involve segments of text that are immediately adjacent to the anaphor (demonstrative expressions used for other purposes may have referents that are farther away). Webber addresses the problem by using a referring function to generate a new discourse entity in the discourse model each time a deictic expression occurs. Although the theory is useful, her solution cannot actually implemented (see discussion in the following section).

In addition to their deictic role, demonstratives may be used anaphorically to replace sections of text. In this capacity, anaphoric expressions reference important entities to help readers keep track of these entities' roles in unfolding events. While this is a crucial purpose of anaphora in general, demonstratives function this way less frequently than other types of anaphoric expressions. As explored above, the role of demonstratives in this situation may be to signal a certain level of focus or accessibility. An alternate explanation is that demonstratives imply contrast or involve a shift in focus that other expressions cannot invoke. Indeed, they may provide subtle value judgments, giving clues as to the author's intentions and revealing which entities s/he thinks most important. Thus Myers believes that the pronouns employed in demonstrative descriptions "characterize the propositions to which they refer, enabling us to gain some idea of the hierarchy of purposes in the text (cited by Cornish, p. 60)." As indications of distance, the demonstratives *that* and *those* often denote contrast. In the following example, Myers (1988) demonstrates how *that* implies subjective distance (cited by Cornish, 1999, p. 60):

> A hairpin stucture could hold the point of splicing in its stem, but **that** would necessitate ligtion from one chain across to the opposite side of the helix. . .

Here the pronoun *it* could be substituted for *that*, but use of the demonstrative emphasizes that the situation is hypothetical and somewhat undesirable. From a larger perspective, demonstrative expressions may bridge paragraph boundaries to "occur at points of transition within a discourse, signaling the start of a new discourse unit by refocusing the addressee's attention on a referent which has been the object of earlier talk but has subsequently been displaced, or has been evoked in the immediately preceding segment (Cornish, 1999, p. 60)."

AUTOMATIC RESOLUTION OF ANAPHORA

Determining how linguistic theory applies to automatic language processing is a major goal of computational linguistics. Numerous automated tasks may be affected by linguistic knowledge: machine translation, natural language interfaces, speech processing, document processing, etc. Two major approaches compete in computational linguistics. The cognitive approach takes the holistic perspective that since language is a function of the brain, we must model the brain to understand language. While this approach has the advantage of a common framework for researchers working on different aspects of the problem, it may be impossible to achieve. In contrast, the probabilistic view is reductionist, arguing that we should model individual phenomena of the brain, rather than the entire system. Attempts at anaphoric resolution follow similar rationale. Certain algorithms approach anaphora from the discourse level, modeling the entire text. Others concentrate on morphology or syntax, treating individual components of discourse in order to build a coherent picture of the whole.

Most resolution algorithms rely on knowledge about language processing in the human brain. Humans resolve anaphoric references almost effortlessly. An initial detailed description of an entity allows construction of a mental image; subsequent mentions may be abbreviated because they need merely remind readers and listeners of concepts already present in their mental state (Liddy et al., 1987). Resolution occurs most quickly when the concept is still active in memory, and more distant antecedents require more specific anaphoric references. Cognitive psychologists have found that the ability to resolve anaphora may be affected by the current focus of the discourse, the anaphor's linguistic characteristics, and real-world inferences (Garrod et al., 1994). References to recently mentioned antecedents are resolved more quickly than those to distant correlates; moreover, resolution proceeds more quickly if the antecedent is a primary focus of the text. Entities introduced by proper name seem easier to remember than those introduced by definite description; correspondingly, anaphors in the form of definite descriptions or names are resolved more quickly than pronouns. (Of course names, much more explicit than pronouns, also apply to fewer possible antecedents). It is often assumed that readers and listeners apply real-world knowledge only after narrowing the list of potential antecedents by applying linguistic and discourse constraints.

Attempts at anaphoric resolution reflect these findings. Research in theoretical linguistics and natural language processing has produced a host of techniques to locate and resolve co-referring expressions; a survey of the basic considerations demonstrates the complexity of the problem. Linguists have developed a variety of approaches to the problem with varying reliance on syntax, semantics, discourse structure, and real-world knowledge. Proponents of *shallow processing* argue that linguistic knowledge should suffice, reserving real-world knowledge as a last resort (Carter, 1987), while others insist that common sense knowledge should (and can) be encoded (Hobbs, 1999). Algorithms relying on several different strategies may rank a candidate according to the criteria for each technique, compiling a final score from these results (Lappan & Leas, 1994). While there are no universal benchmarks to measure the efficacy of resolution algorithms, Mitkov argues that *recall* and *precision* may be useful evaluation measures (1998).

Traditional resolution methods concentrate on word-level (morphological) or sentence-level (syntactic) phenomena, assessing candidates for agreement in gender, number, and person and applying basic semantic constraints (i.e., personal pronouns cannot refer to inanimate objects) (Charniak, 1972; Hobbs, 1977). These techniques concern only intrasentential pronominal anaphora. Among the more influential techniques is Hobbs' algorithm, which maps texts onto "surface parse trees." The algorithm then identifies antecedents by navigating through the trees in a specified order (the antecedent is the first noun phrase reached on the tree that satisfies gender and number constraints). Despite the simplicity of his approach, Hobbs found his algorithm to be successful around 88% of the time (1977). (Hirst points out that this success rate is somewhat inflated because many of his examples involved only one possible antecedent). Subsequent efforts by other researchers to refine Hobbs' work have led to minor improvements. Lappin and Leass (1994) developed an algorithm based on syntactic measures of salience, recency, and frequency of mention; in an explicit comparison, their algorithm proved 4% more successful than Hobbs'. Interestingly, incorporation of semantic and real-world knowledge only slightly improved the algorithm's results, leading Lappin and Leass to conclude that such knowledge should only be applied to the output of the syntactic algorithm when syntactic constraints proved insufficient. Kennedy and Bourgarev (1996) argue that Lappin and Leass' parsing techniques are too sophisticated for current parsing technology and offer a modification using less sophisticated linguistic processing. Their adaptation, while less accurate than Lappin and Leass' formula, applies to more real-world text processing situations.

However, there are strong arguments against basing resolution solely upon syntactic measures. First of all, texts (especially oral ones) do not always follow rules of gender and number. (Consider, for example, nontraditional pronoun use in gender-inclusive language). Furthermore, lexical information can be crucial; as Webber (1991) points out , changing one word may alter the correct interpretation of an anaphor. Contrast the following examples:

> Segal, however, had his own problems with women. He had been trying to keep his marriage of seven years from falling apart. When **that** became **impossible**. . .
>
> Segal, however, had his own problems with women. He had been trying to keep his marriage of seven years from falling apart. When **that** became **inevitable**. . . (Webber, p. 113).

In the first version, *that* describes Segal's efforts to hold his marriage intact, but in the second version the pronoun refers to the dissolution of the marriage. Syntactic information alone could not account for the different interpretations; semantic knowledge is necessary. Most methods grounded in syntax do make a nod toward semantics, but their creators argue that the simplicity of the syntactic approach compensates for the increased accuracy of including semantics.

Traditional resolution methods based primarily on syntax generally account only for intrasentential references. To include anaphora that cross sentence boundaries, many theories approach the problem from the discourse level. Rather than analyze the grammatical properties of individual sentences, these theories attempt to make the discourse itself the basic unit of analysis. The assumption is that readers construct a mental representation of a text as they progress through it. The overall meaning of the discourse is interpreted incrementally; the big picture shifts and changes with updates of information from new sentences. These alterations correspond to changes in the discourse *focus*. Presumably, discourse is organized around a series of discourse foci, alternately known as *themes* (Halliday, 1967), *centers* (Allen, 1995) or *topics* (Reinhart, 1982). Any discourse entity   an object, person, event, proposition, or other sort of concept described in the text   may come into focus (i.e., become salient in the reader's consciousness). Generally several sentences share the same focus before attention shifts to a new object. Entities in focus are almost always the subject of anaphoric references (otherwise paragraphs would become terribly repetitive). To complicate matters, focus can be defined along a continuum; according to Kantor's (1977) idea of the *activatedness of a concept*; the more activated a entity, the easier it is to resolve an anaphor referencing it (cited in Hirst, 1981). According to this view, the choice of anaphor may determine the degree to which an entity is activated. Consider the following text:

(a) The mother picked up the baby. She had been ironing all afternoon. She was very tired.

(b) ***It*** had been crying all day.

(c) ***The baby*** had been crying all day (de Swart, p. 149).

Both mother and baby are introduced in the first sentence, but as the text goes on to describe the mother, she becomes the topic of subsequent sentences while the baby retreats to the background. To return the baby to the foreground, the definite NP of (c) is more appropriate than the indefinite pronoun of (b). The vague indefinite pronoun prevents the baby from becoming fully activated and thus makes resolution more difficult.

Alternate theories incorporating discourse theme use different representation schemes but share the same general ideas. Discourse entities are produced from the text and added to a hierarchic *discourse model* that represents the reader's mental construction of the text. Subdivided into regions corresponding to coherent sections of text, the model evolves as the reader progresses through the text and different entities come into focus. Allen (1995) describes a relatively simple version of this approach. His method judges potential antecedents on their likelihood to be a discourse center. All nominal expressions that are potential antecedents for subsequent sentences are complied in a history list ordered by recency. According to the recency constraint, a pronoun refers to the most recently mentioned noun phrase that satisfies all relevant constraints; thus the system moves down the list, applying additional constraints to each discourse entity until it locates a suitable antecedent. Constraints are based on the role an entity plays in the changing discourse focus. Webber (1979) was the first to begin the process of automated resolution by identifying entities with the potential to become referents. She adopts the formal logic used by many classical linguists, representing sentences with predicate calculus. A set of rules is applied to these logical representations to derive entities that are likely to serve as referents for anaphors in subsequent sentences. One of this method's advantages is the inclusion of anaphora that violate constraints concerning number (where, for instance, the antecedent is singular and the anaphor is plural). Another strength of the method is the introduction of formalism into automatic resolution. However, Webber's method accounts only for certain categories of anaphora and ignores referents which are not explicit in the text. Grosz (1977) accounts for one of Webber's weaknesses by considering the role of discourse structure in the identification of focus. Sidner (1978) builds upon Grosz's work but uses frames to represent world knowledge; her work has proved particularly influential.

Many other researchers address the problem of automated resolution, building resolution algorithms with varying reliance on syntactic, semantic, and discourse knowledge. Most algorithms consider only certain kinds of anaphora, and none are universally heralded as successful. The problem will most likely continue to play a major role in computational linguistics research for years to come.

Resolution of Demonstrative Anaphora

Perhaps due to the range and complexity of their referents, demonstratives have largely been ignored in automatic resolution. Although Webber (1991) attempts to include demonstratives in her model for discourse deixis, most resolution algorithms focus instead on definite pronouns. This has prompted a few comparative examinations of demonstrative and definite pronouns in the hope of adapting algorithms to satisfy either category. These comparative studies draw mainly from corpora of spoken English and limit their scope to the pronouns *that* and *it* (Passoneau, 1993; Byron and Allen, 1998). Myers' examination of written scientific documents has a somewhat broader focus (1988).

Webber's algorithm is governed by her assumption that only regions that are currently in focus may yield referents for deictic expressions. (She uses a tree structure representing hierarchal relationships among discourse entities to demonstrate the discourse model's evolution more formally; nodes located on its right frontier are in focus and may yield referents.) Among the discourse entities in the focused region is the demonstratum (a segment of text); the referent may be a new entity representing the propositional content of the demonstratum. Webber suggests first determining whether a demonstrative expression points to an entity or a “discourse segment.” If the demonstratum is a discourse segment, a new discourse entity must be created for each segment that is a potential demonstratum.

Unfortunately, the theory stumbles on the inability to define discourse segments. While it is generally agreed that discourse can be divided into segments of related sentences or clauses, there is no consensus on how to accomplish this division. Webber freely admits this flaw but assumes that it will eventually be remedied. In the meantime, she adopts the naive approach of limiting discourse segments to sentences and clauses.

In her comparison of *it* and *that* in a corpus of oral interviews, Passoneau observes that the definite pronoun occurs when both referent and pronoun are subjects of a clause or sentence, while the demonstrative is used when either the pronoun or the antecedent is not the subject. She finds that *it* and *that* have contrasting functions in most contexts. Like Webber, she believes that deictic demonstratives require the creation of new discourse entities, but she proposes that the referring function governing this creation cannot be generated automatically. Rather, it depends on a reasoning process which future research must codify.

In an examination of *it* and *that* in a corpus of task-oriented spoken dialogue, Byron and Allen find that *it* is much more likely to reference concrete objects in the discourse context, while *that* more often refers to abstract entities and propositions. While many of their findings are specific to the corpus, they do propose some syntactic criteria for determining whether a pronoun refers to an abstract entity (in their corpus, an abstract entity consists of a plan, action, task, fact, or propositional content). They plan to incorporate these syntactic patterns into an algorithm based on Webber’s method.

METHODOLOGY AND RESULTS

Initial efforts to examine anaphora were largely exploratory; we hoped to uncover trends in the collection that might lend themselves to heuristics for anaphoric resolution. Moreover, we hoped that the distribution of anaphors might coincide with the introduction and dismissal of subtopics, or that they would provide some other clues for best representing the "aboutness" of a document. Toward these ends, we analyzed anaphora in a sample of queries, abstracts, and full-text documents.

Our document collection, the CF database, contains all documents with the heading "Cystic Fibrosis" entered between 1974 and 1979 in the U.S. Government's National Library of Medicine Medlars database (Shaw, Wood, Wood, & Tibbo, 1991). Also included are 100 queries and accompanying relevance judgments from medical personnel specializing in Cystic Fibrosis. Supplementing the original database is the full text of about one third of the documents (Moon, 1993). Documents tend to have a fairly rigid structure, adhering to the standard subsections used in scientific articles: Introduction, Materials and Methods, Results, Discussion. A typical article contains around 15-25 paragraphs; while these paragraphs vary in length (some are as short as one sentence, while others may contain a couple dozen sentences), the majority are relatively short (four or five sentences). A few articles contain detailed subheadings which emphasize their structure.

Overview of Anaphora in the CF Database

In our sample of abstracts and full-text documents, more than 250 instances of anaphora were found in about a dozen different documents. However, no anaphoric references were found in the queries. Brevity may be partially responsible (no query was more than one sentence long); also, queries were written by subject matter experts rather than real users and thus might be somewhat more formal and less likely to rely on abbreviated references. Informal explanations of information needs – especially when expressed verbally – would be much more likely to contain anaphora.

Our classification scheme for anaphors draws from previous studies in linguistics and information science (Denber, 1998; Allen, 1995; Liddy, Bonzi, Katzer, & Oddy, 1987; Hirst, 1981). Table 1 presents the categories used in the first phase of the study and the corresponding number of anaphors in the sample.

Table 1: Categories of Anaphora in CF database

Category	Examples	No.	%
Pronouns			
Personal	he, she, it, they, his, hers, them, their	**32**	**12%**
Demonstrative	this, that, these, those	**142**	**51%**
Reflexive	himself, herself, itself, themselves	**0**	**0%**
Indefinite	all, any, both, each, many, one, some	**38**	**14%**
Relative	who, what, which, where, when	**1**	**.4%**
Nominal Substitutes	the first, the second, the former, the latter	**4**	**1%**
Pro-adjectives	another, identical, other, same, similar, such	**10**	**4%**
Pro-adverbials	so, similarly	**0**	**0%**
Definite descriptions (definite noun phrases, subject references)	“the dog” referring to “the furry	**34**	**12%**
TOTAL		**261**	**100%**

Nearly half the anaphors in the sample are demonstrative, a proportion much greater than any other type. Although demonstrative pronouns were among the most common classes found by Bonzi and Liddy, they did not occur in nearly as high a proportion as in the CF database (1988). Bonzi and Liddy note that sublanguages used in different domains show different linguistic properties, hypothesizing that anaphora may be among sublanguages’ distinguishing characteristics. Accordingly, demonstrative anaphora may occur particularly frequently in medical articles. However, demonstrative anaphora were important in Bonzi and Liddy’s dataset; about three fourths referred to integral concepts (a proportion greater than most other classes) (1988). Thus the difference is more likely due to document length; Bonzi and Liddy worked only with abstracts, and we have examined full-texts. Since demonstrative anaphora often summarize complex events described in lengthy phrases, they are probably less likely to occur in abstracts than full-text, where the expansion of crucial concepts requires repeated references to an entity.

Although demonstrative expressions are by far the most common category, indefinite pronouns, personal pronouns, and definite descriptions also occur in significant numbers. This seems typical of most English texts. Personal pronouns are generally considered to be the most common type of anaphor, and for the most part their presence and use in the CF collection is unremarkable. Since many articles chronicle research conducted on CF patients, a typical use is reference to groups of patients. However, personal pronouns are also used exophorically to indicate a document’s authors, as in: **We** evaluated suppressibility for each patient studied...” In fact, personal pronouns were used more often than any other category of anaphora to reference entities outside the discourse context.   The use of indefinite pronouns is fairly nondescript; *both*, *some*, and *each* refer to various subsets and combinations of previously mentioned entities. Likewise, definite descriptions tend to serve the standard anaphoric purpose of abbreviating full descriptions. Often a series of definite descriptions and indefinite pronouns occur in close proximity, offering alternate references to the same entity.

In addition to categorizing anaphors, we classified antecedents according to length and content. Following the scheme of Pirkola and Jarvelin (1996), we categorized antecedents as a simple noun (one word), compound noun (two words), or phrase (three or more words). Phrases, capable of carrying greater complexity, probably indicate content better than single terms alone; they may discuss concepts of greater complexity or specificity than can be expressed in one or two words. Indeed, it turns out that about two-thirds of the antecedents in our sample are phrases three or more words in length (Table 2). As our study progressed, it also became clear that the category "phrases" was too broad; many anaphors referred to complete sentences, and some described the contents of entire paragraphs and sections. Hence we also noted whether the antecedents of demonstrative expressions comprise sentences or longer segments of text.

Table 2: Length of antecedents in CF database sample

Length	No.	%
simple noun	19	8%
compound noun (2 words)	63	26%
phrase (3+ words)	162	66%

Although the data presented in Table 2 cannot be directly compared to Pirkola and Jarvelin's results (rather than classify all the antecedents in documents, they examined only those corresponding to pre-selected queries), their study also reports an abundance of phrase-length antecedents. However, Pirkola and Jarvelin also found that these phrases usually contain proper nouns. The CF database, in contrast, contains so few proper names that we abandoned our original plan of characterizing antecedents as proper or common nouns. As Pirkola and Jarvelin examined a collection of full-text newspaper articles, these differences again emphasize the impact of subject domain on anaphora.

The locations of antecedents and anaphors within sentences and paragraphs were also recorded, revealing that 56% of anaphora are intrasentential, while 42% cross sentence boundaries. In addition, approximately 2% of anaphora in the original sample are exophoric (i.e., they have no explicit antecedent within the text). Unsurprisingly, anaphora with antecedents of phrase length are more often intersentential, while antecedents consisting of one or two words are more likely to occur in intrasentential anaphora. The majority of intersentential references employ demonstrative anaphora, with indefinite pronouns taking a distant second place (Table 3). Intrasentential references are distributed slightly more evenly; although demonstrative expressions still comprise the majority, both personal and indefinite pronouns account for a large number of co-references within sentence borders. The importance of demonstrative pronouns in intersentential reference implies that these expressions play a crucial role in binding together sentences into a cohesive text.

Table 3: Range of Reference according to Anaphoric Category

	Personal Pronouns		Demonstrative Pronouns		Indefinite Pronouns		Relative Pronouns		Nominal Substitutes		Pro-adjectives		Definite Descriptions	
Intrasentential	15	22%	31	46%	12	18%	0	0%	1	1%	2	3%	6	9%
Intersentential	6	6%	65	68%	15	16%	3	3%	0	0%	5	5%	2	2%
Exophoric	4	80%	1	20%	0	0%	0	0%	0	0%	0	0%	0	0%

Table 4 shows the tendency of each anaphoric class to be intrasentential, intersentential, or exophoric. Unfortunately there is too little data on relative pronouns, nominal substitutes, pro-adjectives, and definite descriptions to make generalizations. However, our data does confirm that intrasentential reference accounts for a large proportion of the use of personal pronouns. Although the majority are used for intersentential reference, indefinite pronouns tend to be more evenly distributed than personal pronouns. Again, demonstrative pronouns show the most dramatic trend; nearly two-thirds are intersentential. Clearly demonstrative expressions enjoy an impressive range, often referring to antecedents beyond sentence borders. However, their role remains flexible; they serve intrasentential anaphora a healthy proportion of the time.

Table 4: Anaphoric Category according to Range of Reference

	Personal Pronouns		Demonstrative Pronouns		Indefinite Pronouns		Relative Pronouns		Nominal Substitutes		Pro-adjectives		Definite Descriptions	
Intrasentential	15	60%	31	32%	12	44%	1	100%	0	0%	2	29%	6	75%
Intersentential	6	24%	65	67%	15	56%	0	0%	3	100%	5	71%	2	25%
Exophoric	4	16%	1	1%	0	0%	0	0%	0	0%	0	0%	0	0%
Total	25	100%	97	100%	27	100%	1	100%	0	100%	7	100%	8	100%

To judge whether anaphoric expressions described concepts central to the text, we determined whether words appearing in antecedents were listed among the index terms for the document. Since only keywords themselves were counted – while synonyms and related terms were ignored – the measure is a rough gauge. Overall, about 42% of antecedents contain keywords. When considered individually, most anaphoric classes are more likely to have antecedents that do not contain index terms (Table 5). Pro-adjectives are an exception, but our data does not include a sufficient number of examples to give a fair count. Although the proportion is still the minority, definite descriptions also have a fair number of antecedents containing keywords. Many definite descriptions refer to the main subject of a paragraph, and any topic important enough to command focus during an entire paragraph quite likely contains index terms. Demonstrative antecedents are evenly split. Although we might expect a slightly higher proportion to contain keywords, comparison with other categories shows that fifty percent is a relatively high figure. While index terms occurred nearly twice as often in intersentential antecedents (65%) than intrasentential (34%), the comparison is misleading since intersentential referents tend to contain more words. Still, the presence of keywords reinforces trends apparent in the data explored above; anaphoric classes likely to be involved in intersentential reference are also more likely to have complex antecedents containing keywords. Thus in addition to making important contributions to document structure, these classes represent integral content.

Table 5: Presence of Keywords in Antecedents

	Personal Pronouns		Demonstrative Pronouns		Indefinite Pronouns		Relative Pronouns		Nominal Substitutes		Pro-adjectives		Definite Descriptions	
Keywords	7	25%	70	50%	13	34%	0	0	1	25%	6	60%	15	44%
No Keywords	21	75%	71	50%	25	66%	1	100%	3	75%	4	40%	19	56%
Total	28	100%	142	100%	38	100%	1	100%	4	100%	10	100%	34	100%

Note: Counts in Table 5 exclude exophoric references.

Demonstrative Anaphora in the CF Collection

Data Collection

The results outlined above show that demonstrative anaphora not only appear in the CF database more frequently than any other category, but also tend to serve complex and interesting functions. Thus the patterns of their occurrence were examined in more detail. Most of the analysis was done by hand, but a perl script was used to tokenize documents into sentences and locate keywords and demonstrative pronouns (see Appendix A for sample output). For certain documents, we added paragraph boundaries to better visualize the distribution of anaphors and antecedents (see Appendix B). Highlighting the position of anaphoric expressions in the underlying document structure, these representations suggest possible trends that may be more difficult to recognize in the complete text. For example, they emphasize clusters of demonstrative expressions and likewise highlight areas where no expressions are present. These views also illustrate the dispersal of index terms throughout the text, revealing locations where certain sequences of keywords may coincide with specific subtopics. While these pictures currently offer a rough and somewhat distorted perspective, they could be refined to play a more useful role in displaying the interaction between anaphora and document structure. For example, eliminating intrasentential references from the pictures would present a clearer picture of the intersentential anaphora that bind together sentences and paragraphs.

For each instance of anaphora, we recorded the terms comprising the anaphor and antecedent and noted their exact location (i.e., word number within the sentence and sentence number within the paragraph). We also noted whether any of the terms are keywords. Furthermore, we determined an anaphor's position within the larger context by noting whether it belongs to a "chain" of anaphoric references. A chain consists of a sequence of anaphors which may refer to each other but share the same ultimate antecedent. Finally, we determined whether a reference is exophoric or whether it refers to an entity within the text. A summary of the descriptive data categories is presented in Table 6. Although we collected data for about 330 different instances, we did not collect data from every category for each example.

Table 6: Data Collected to Describe Demonstrative Anaphora

Anaphor	Proximal or distal
	Singular or plural
	Pronoun or adjective
	Word number in sentence
	Sentence number in paragraph
	Position in anaphoric chain
Antecedent	Keywords contained in antecedent
	Length of antecedent
	Word number in sentence
	Sentence number in paragraph
Reference	Exophoric or endophoric
	Number of sentences between anaphor and antecedent

Antecedent Length

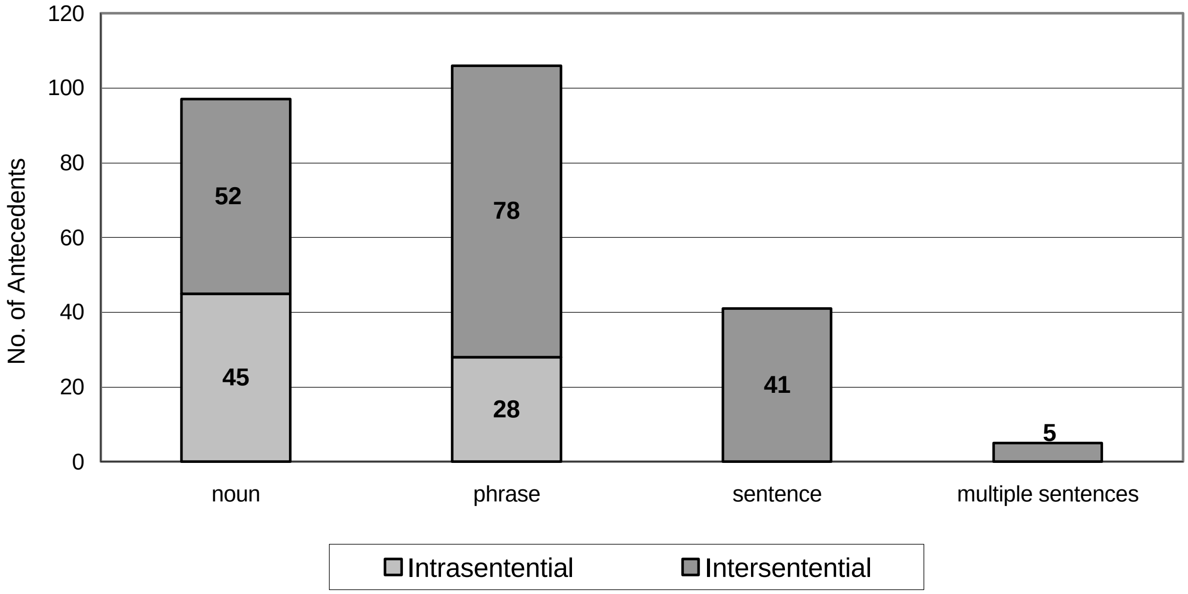
Antecedents come in varying levels of length and complexity; in our sample, each one has been classified as a noun, phrase, sentence, or sequence of multiple sentences. The most basic category, nouns consisting of one or two words, comprise nearly 40% of the antecedents within the sample (Figure 1). Forty-six percent of these noun-length antecedents are referred to by anaphors within the same sentence, while the remaining 54% involve intersentential references. Phrases, accounting for 43% of antecedents within the sample, are the most common length. However, they are also the most broadly defined category, including all antecedents between three words and one sentence in length. In the future, it may be useful to distinguish clauses from shorter phrases since clauses actually have more in common with sentences. As is, about three-fourths of the phrase-length antecedents occur in intersentential references.

Antecedents exceeding phrase-length are substantially less common. Sixteen percent of the antecedents comprise a complete sentence. Only 2% (n = 5) of the antecedents are longer than a sentence; of these, three comprise multiple sentences within a paragraph, and one antecedent spans an entire paragraph. The remaining example consists of an entire section (multiple paragraphs); here the “Discussion” section begins by referencing the findings explored in the preceding “Results” section:

> **These results** show that patients with Cystic Fibrosis have an immediate Type 1 hypersensitivity to a wide variety of allergens...

Of course, the fact that nouns and phrases may be used in both intrasentential and intersentential reference accounts partially for the frequency of their use (as opposed to antecedents composed of sentences, which obviously can only be used in intersentential anaphora). However, phrase-length antecedents remain the most common category even when intrasentential references are excluded; they account for forty-five percent of intersentential references.

Figure 1: Length of Demonstrative Antecedents



Forms of Demonstrative Anaphors

The vast majority (85%) of anaphora in the CF database are proximal (*this*, *these*), and the most common form overall is the adjective *this* (Table 7). The second most common form, the adjective *these*, accounts for nearly one-third of demonstrative anaphors. Next common is the pronoun *this*, which appears in 11% of anaphors. Comparably, *those* used pronominally occurs in 8% of anaphors. The remaining categories each account for less than 5% of anaphors in the sample. The least common form, occurring only twice, is the adjective *that*.

Table 7: Demonstrative Anaphora by Type

Expression	Number	Percentage
that	13	4%
that + NP	2	1%
those	28	8%
those + NP	5	2%
total distal	***48***	***15%***
this	37	11%
this + NP	131	40%
these	11	3%
these + NP	103	31%
total proximal	***282***	***85%***

The distribution in our collection resembles that of the SUSANNE-corpus, a database of written English documents from the press, belles lettres, learned writing, and fiction (Himmelmann, 1996). According to Himmelmann, about 72% of demonstratives in the SUSANNE collection (n = 1139) are proximal demonstratives, while the remaining 28% are distal. Proximal demonstratives are more likely to occur as adjectives, while distal demonstratives tend to function as pronouns; the single most common form is the adjective *this*. The CF collection mirrors these patterns but shows them to a greater degree; possibly the presence of several different genres flattens the trends in the SUSANNE corpus.

The form of a demonstrative anaphor has some bearing on the length of its antecedent (Table 8). While Myers asserts that the antecedents of adnominal demonstratives enjoy a greater range than those of pronominal demonstratives, our examination shows that the pronoun *this* may refer to complete sentences. However, adnominal demonstratives do account for two-thirds of sentence-length antecedents and all multiple-sentence antecedents our sample.

Table 8: Antecedent Lengths for Types of Demonstrative Anaphors

	this		this + NP		these		these + NP		that		that + NP		Those		those + NP	
noun	3	10%	41	47%	3	33%	24	28%	9	69%	1	50%	18	69%	2	50%
Phrase	12	41%	28	32%	6	67%	51	59%	4	31%	0	0%	8	31%	1	25%
sentence	14	48%	18	21%	0	0%	7	8%	0	0%	1	50%	0	0%	1	25%
multiple sentences	0	0%	0	0%	0	0%	5	6%	0	0%	0	0%	0	0%	0	0%
total	29	100%	87	100%	9	100%	87	100%	13	100%	2	100%	26	100%	4	100%

By far the most common types of demonstrative expressions in our sample are proximal demonstratives used as adjectives. These serve as anaphors for almost two-thirds of sentence-length antecedents. Antecedents of greater length appear only five times in the sample, but all of these references employ an adnominal demonstrative. Interestingly, all twenty-seven exophoric references in the sample use adnominal demonstratives; one uses the adjective *these*, and the remaining twenty-six employ the adjective *this*. In nearly half the cases when *this* is employed as an adjective, it refers to a simple noun phrase; however, it also refers to phrases and sentences in large proportions. The adjective *these*, in contrast, most often refers to a phrase and rarely refers to an entire sentence.

In almost half of its occurrences, the pronoun *this* refers to an entire sentence. In all fourteen of these instances, the pronoun begins the sentence immediately following its antecedent. A typical example:

> When intravenous arginine was used as the stimulus to insulin secretion, none of the CF patients in either group had a significant response. **This** resembles the findings of Kalk and associates. . .

The pronoun *this* refers to phrases almost as frequently as it does sentences but in this situation does not necessarily occur at the beginning of a sentence. When referring to phrases, the pronoun *this* begins a sentence about half the time.

The demonstrative *these* seldom occurs as a pronoun, but its appearances confirm Himmelmann's assertion that it is the least flexible form of demonstrative expression. All of its antecedents are nouns or phrases; it does not seem capable of referring to the more complex content contained in longer antecedents.

Distal demonstratives occur much less frequently than proximal demonstratives. The most common form of distal demonstratives is the pronoun *those*, most often referring to a noun. Similarly, the pronoun *that* most frequently refers to a noun. Distal demonstratives occur adnominally on too few occasions (n = 6) to make generalizations, but apparently they refer to antecedents of noun, phrase, and sentence length.

The fact that various types of demonstratives exhibit such different tendencies suggests that a resolution algorithm should incorporate the form of a demonstrative expression in calculating the likelihood to reference a particular length of antecedent. Upon encountering the pronouns *that* or *those*, for example, the algorithm would weight noun-length antecedents most heavily, while the pronoun *this* would cause

Complex Antecedents and Discourse Deixis

In many cases, anaphors do not function as simple substitutes for antecedents. Indeed, it is common for anaphors in the CF database to be used in discourse deixis, representing complex entities or events explained elsewhere in the text. Often these instances of anaphora refer to entities that are important indicators of the "aboutness" of a data which may help answer the question of whether intracellular mucus is or is not abnormal" – the very topic under discussion in this excerpt. The presence of several keywords (underlined in the following examples) provide further evidence of the inclusion of integral content.

> The increased viscosity of bronchial secretions in patients having cystic fibrosis is well known. The protein and enzyme concentrations have been reported to be elevated in CF salivas - and in bronchial secretions. However, viscosity measurements recently have been reported to be normal. **This apparent paradox** may be understandable in terms of calcium concentration. . .

A second example covers a smaller range but is noteworthy because it combines propositions from two previous sentences into one entity. Here the anaphor incorporates two groups of children, each described in a separate sentence. Whereas Himmelmann finds that only singular demonstratives are used for discourse deixis, the following example illustrates that we have not found this to be the case.

> In the pancreatic insufficiency group, 3 children had zero values of TPA (and of trypsin) in the fasting condition and after the test meal. One child with a low but measurable TPA had also a low trypsin content in duodenal juice (47 ug/ml). **These 4 patients** had clinical signs of malabsorption and had steatorrhoea.

Sometimes the antecedent is complex enough to warrant a lengthy anaphor, as in the following example:

> Especially the concentrations of many "acute phase proteins" are significantly changed (concordantly increased: antitrypsin, antichymotrypsin, sin, haptoglobin, ceruloplasmin and hemopexin; concordantly decreased: HS-glycoprotein and albumin). **This type of correlated alterations in the "acute phase proteins"** are generally found under circumstances where tissue damage takes place.

Anaphors used deictically can be somewhat ambiguous; it is often difficult to specify their antecedents. The anaphor in the following excerpt, lifted from the first paragraph in one article's introduction, most likely refers to the preceding sentence, but could also be interpreted as encompassing the entire previous paragraph:

> Circulating serum autoantibodies to human pancreas in children with cystic fibrosis (C.F.) have been reported by Murray and Thai (1960), and local autoantibodies to lungs from C.F. patients at necropsy have been shown in their sputum by Stein et at. (1964). In addition, a variety of serum precipitations have been detected in a high percentage of C.F. patients (Burns and May, 967; McCarthy et al., 1969). In our previous study not only were a wide variety of precipitating antibodies detected in the serum of C.F. patients but also they were found in much higher concentrations and numbers in the corresponding sputum (Wallwork et al., 1974). **These observations** prompted us to investigate the occurrence of immune complexes in C.F. patients.

Since cases like this one can confuse human readers, obviously they would present enormous difficulties for automatic resolution systems.

A sufficient number of anaphora participate in discourse deixis to demonstrate that a resolution algorithm cannot avoid addressing this issue. Unfortunately, the patterns underlying discourse deixis are not readily apparent; while deictic demonstratives often refer to the preceding sentence, the examples presented above illustrate that there are a wealth of exceptions. The variety and complexity of these exceptions — particularly the fact that their resolution can perplex human readers — necessitates further examination of the phenomenon.

Anaphoric Chains

Another interesting use of anaphora is repeated reference to one entity through the use of a pronoun chain. A typical pattern for anaphoric chains is a series of consecutive sentences, each containing an anaphor:

> Category 1. In **five patients**, **all with severe lung disease**, high AP levels developed only after the onset of cor pulmonale. In **all five**, AP determinations had been normal during the year preceding. **All five patients** had less than 3.0 gm/100 ml of albumin in their serum. **Three of these five patients** had an SGOT level between 40 and 95 units/ml; the SGOT of the **other two** was less than 40 units/ml. **None of the five** was hypoprothrombinemic or hyperbilirubinemic. Postmortem examination, subsequent performed on **two of these patients**, (E.W., E .J .) demonstrated in each case both chronic passive hepatic congestion and focal biliary cirrhosis.

Here the first sentence establishes background for the patients belonging to “Catego Subsequent sentences may take advantage of this background, providing only abbreviated references to subsets of this group. The six anaphors linked in this chain make it one of the longest in the sample; most contain only two or three anaphors. In at least five cases, including the following example, a chain consists solely of demonstrative anaphors. This example includes an antecedent and two subsequent anaphors occurring in three consecutive sentences. In one sense, the first anaphor also serves as the antecedent for the second anaphor, but ultimately the reference for both anaphors can be traced back to the original antecedent (in this example, “CF patients.”)

> In fact one could speculate whether the high number of precipitins and the persistent infection by means of a type III hypersensitivity reaction (2) could possibly contribute to the

> tissue damage in the lungs of **CF patients**. On the other hand, these antibodies possibly play a role in localizing the infection to the respiratory tract, as **these patients** rarely, if ever get generalized infection caused by Ps. aeruginosa. Hoiby & Axelsen (7) have recently suggested that the defective protection of the lung tissue offered by the many Ps. aeruginosa precipitins might-at least partly-be explained by properties of the Ps. aeruginosa strains found in **these patients**, i.e. production of great amounts of mucoid substance

We identified more than thirty anaphoric chains in the sample but believe that this underestimates the actual number.

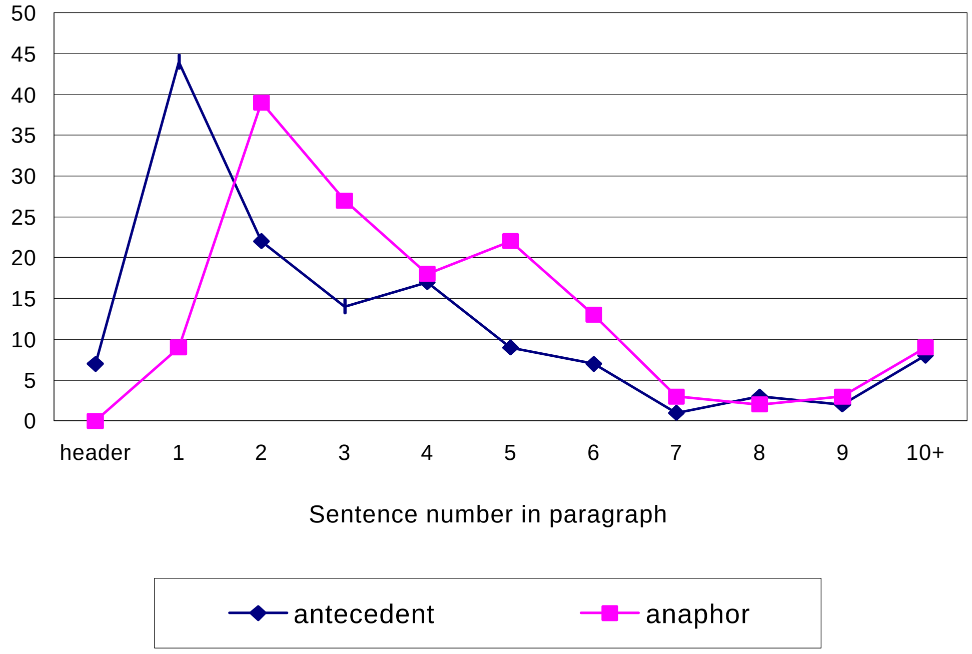
Concepts embodied in anaphoric chains seem to comprise crucial content. However, it is unclear whether the pronoun chains may indicate a certain type of content; perhaps they represent major concepts which provide the backdrop for lesser subtopics, or maybe they often represent the subtopics themselves. Again, further investigation might provide helpful insight.

Intersentential Reference: Position of Anaphors and Antecedents in Paragraphs

About 70% of anaphora in the sample cross sentence borders. When used for intersentential reference, anaphor and antecedent typically occur in consecutive sentences, and the antecedent comprises either an entire sentence or a large portion thereof. In fact, 85% of intersentential references find the antecedent occurring in the sentence preceding the anaphor. In 10% of intersentential anaphora, the antecedent is located two sentences before the anaphor, and the remaining 5% of intersentential references have antecedents that are three or more sentences away from corresponding anaphors.

Since the vast majority are located in adjacent sentences, antecedents and anaphors show comparable trends in their distribution within paragraphs (with peaks in anaphors lagging one sentence behind peaks in antecedents). Figure 2 displays these trends, demonstrating that most antecedents occur in the first (34%) or second (17%) sentences of paragraphs. Correspondingly, anaphors are most likely to occur in the second (27%) or third (19%) sentences. There is a small rise in the number of antecedents at the fourth sentence, mirrored by an increase in anaphors in the fifth sentence. It could be that at this point in the paragraph, authors are ready to present a new entity worthy of pronominalizing. The apparent rise in the number of antecedents and anaphors at the end of the paragraph only indicates the presence of an umbrella category encompassing all sentences beyond position ten; actually no more than three anaphors or antecedents occur in any given sentence position past the ninth sentence.

Figure 2: Position of Anaphors & Antecedents in Paragraphs



The existence of an anaphor in the first sentence of a paragraph raises the possibility that this paragraph continues discussing a subtopic already introduced in a previous paragraph. Of course, the distribution of subtopics among paragraphs depends on both the genre and the author's individual writing style. Obviously, longer paragraphs are likely to discuss multiple subtopics, whereas the factual, expository writing style typical of scientific works lends itself to short, sharply-focused paragraphs. In this situation, where the focus tends to shift with each new paragraph, opening sentences should contain many antecedents and few anaphors. In fact, only 6% of intersentential anaphors in our sample occur in the first sentence of a paragraph. Six of the nine cases do seem to maintain focus on a subject discussed in the preceding paragraph. Following the typical pattern, three have antecedents in the preceding sentence (i.e., the last sentence of the previous paragraph). The remaining three also draw from the previous paragraph, in less expected positions – one antecedent comes from the first sentence in the previous paragraph, one from the second sentence, and one comprises the entire paragraph. In contrast, three anaphors located in the first sentence of the paragraph actually refer to subtopics that have not yet been discussed. In these instances, the antecedent comes from the title of the subsection (which immediately precedes the sentence containing the anaphor).

TOWARD AN ALGORITHM FOR AUTOMATIC RESOLUTION

The behavior of demonstrative anaphora in our sample confirms the patterns set forth in the literature. Like Himmelmann, Ariel, and Myers, we found that:

- Demonstratives occur more frequently as adjectives than pronouns.
- Proximal demonstratives occur much more frequently than distal demonstratives.
- The antecedent of a demonstrative expression usually occurs in the sentence preceding its anaphor.

- Both intersentential and intrasentential antecedents are most often comprised of phrases.
- Demonstrative expressions often participate in discourse deixis, referring to propositions and events rather than replacing segments of text.

This knowledge could be useful in deriving automatic resolution methods. Although most algorithms prepare for anaphors by first compiling a list of all potential discourse entities, this may not be the most efficient method. When discourse deixis comes into play, the range of potential entities is vast; propositions may arise from discourse segments of any length. Therefore, we suggest that it may make more sense to start with the demonstrative anaphor, examine its characteristics, and then proceed backwards to compile a list of possible antecedents. Since certain types of demonstrative expressions are most likely to replace simple noun phrases, perhaps we should not exert effort including complex propositions in our list of potential referents unless they have a high probability of satisfying the anaphor in question. Thus we propose that an algorithm to resolve intersentential anaphora should proceed roughly according to these general steps:

1. Locate demonstrative expression.

2. Determine form of demonstrative.
    a. If the expression is the pronoun *that, these,* or *those,* consider nouns and phrases from the preceding sentence before compiling a list of more complex entities.
    b. If the expression is the pronoun *this* or any type of demonstrative employed as an adjective, assume that the referent could be complex and compile a complete list of both concrete entities and abstract propositions.

3. Compile list of entities serving as possible referents.
    a. If the expression is the adjective *this* or the adjective *these,* consider entities composed of multiple sentences.

4. Assign weights to entities according to:

a. Proximity to anaphor (Generally, entities in the preceding sentence should be weighted most heavily)
b. Probability of demonstrative type to have antecedent of given length

Of course, Step 3 presents enormous difficulties

in some fantastic hand-waving when it comes to identifying discrete discourse segments. In addition to phrases of various lengths, the referents could comprise multiple sentences or paragraphs; we would have to develop criteria to designate which sentence combinations are logical candidates for discourse segments. At this point, we can only hope that the majority of complex referents will arise from entire sentences or from easily extracted phrases. Moreover, our weighting function is vastly oversimplified. The position of anaphors within sentences and paragraphs may have some impact on the length of their antecedents and their tendency to be deictic – although possibly this impact cannot be separated from other factors. In addition, the role of discourse focus should not be discounted; when other criteria fail to identify a referent, salience could be used to make a final choice.

Our algorithm differs from others in two major respects. First, it focuses specifically on demonstrative expressions, incorporating their individual characteristics into its evaluation of possible antecedents. We have not found any other resolution techniques that were developed exclusively for demonstrative anaphora, much less algorithms that consider the specific tendencies of *this*, *these*, *those*, and *these*. Ideally our findings could be combined with other techniques to create an algorithm that carefully considers demonstrative form yet exhibits a broad scope. A second difference in our technique is its initial step; we begin by examining the anaphor, whereas most algorithms proceed linearly through a text, compiling a list of potential antecedents before encountering any anaphors. Since the referents of deictic expressions may comprise discourse segments of any length, it seems difficult to assemble all possible referents without first establishing some limitations. Otherwise every preceding paragraph and combination of consecutive sentences could conceivably be under consideration! If we begin by assessing the probability that a particular type of anaphor in a specific location participates in discourse deixis, we can narrow our selection drastically. If the anaphor is likely to be deictic, we can evaluate its likelihood to reference a certain length of antecedent to rule out specific types of discourse segments.

More in-depth analysis should be performed on our data to make this algorithm concrete enough to be useful. First, the probability of each demonstrative type to serve in intrasentential or intersentential reference should be determined, as the algorithm would have to be modified to account for both cases. Likewise, careful scrutiny of discourse deixis – in particular, determining which types of demonstrative expressions are more likely to be employed deictically – will help determine when the algorithm should apply greater weights to propositions. Furthermore, additional data should be collected on the behavior of distal demonstratives; a much larger sample is necessary to glean an accurate idea of their function in full-text articles. Finally, the positions of anaphors within sentences and paragraphs should be compared to the length and location of antecedents. Although is not immediately apparent in our overview, an anaphor's position within a paragraph may impact the length and composition of its antecedent.

ADDITIONAL APPLICATIONS

Further analysis of our data could also yield a more complete picture of the functions of demonstrative expressions in discourse. One major question is whether, as Myers proposes, demonstratives can be used to show an author’s “hierarchy of purposes” in a text. Examining the behavior of anaphora in individual documents, as opposed to the collection in general, might help us determine how to establish such a hierarchy. Analyzing the characteristics of anaphors in conjunction with our pictures showing their locations within documents might help us rank the concepts they represent.

Additional examination of anaphoric chains could also be useful. Identifying chains of anaphoric references may clarify what it means to be a “discourse segment;” it is widely agreed that cohesive sections of text exist, but there is no consensus on the definition for these segments. Reaching the end of a series of anaphoric references might be one criterion for ending a discourse segment. Of course, this definition would not solve our problems in excerpting segments to identify discourse entities for automatic resolution, but it is another possible approach to the dilemma that could prove useful in other applications.

CONCLUSION

Our cursory examination of demonstrative expressions in the CF database suggests that they play an important role in maintaining discourse focus and binding together cohesive sections of text. Our analysis could prove useful both in the development of automatic resolution techniques and in deriving a more complete idea of the roles played by demonstrative anaphora in written texts. As discussed previously, the benefits of automatic resolution are numerous; automatically locating antecedents could enhance natural language interfaces, improve passage retrieval and question-and-answer systems, and possibly benefit information retrieval matching algorithms. Similarly, discovering the functions of demonstrative expressions in full-length texts could prove useful for both theoretical and computational linguists.

Unlike corpora in other subject domains, our collection of scientific articles contains more demonstrative expressions than any other class of anaphora. In fact, more than two-thirds of intersentential anaphora employ demonstratives, with these anaphors most often referring to phrases contained in preceding sentences. Since intersentential anaphors most frequently occur in the second sentence of a paragraph, their corresponding antecedents are most commonly located in the opening sentence. The most common demonstratives are the adjectives *this* and *these*; the former most often refer to nouns, while the latter is more likely to refer to phrases. Both types, however, also reference propositions expressed in longer discourse segments. They may also participate in anaphoric chains, extending reference to an entity throughout a paragraph.

Our overview of demonstrative anaphora in CF database allows us to make some generalizations applicable to automatic resolution techniques. We have provided some suggestions toward a resolution algorithm, indicating that it may be more appropriate to start the resolution process by characterizing anaphors than by collecting potential antecedents. Additional examination of our data should produce a more concrete algorithm.

In short, our results show that demonstrative anaphora play a complex and interesting function in scientific articles and that their unique characteristics warrant specific consideration in a resolution algorithm. Although further analysis is necessary, these findings provide a groundwork for future exploration.

References

Al-Kofahi, K., Grom, B., & Jackson, P. (1999) Anaphora resolution in the extraction of treatment history language from court opinions by partial parsing. *Proceedings of the seventh international conference on Artificial intelligence and law*: 138-146.

Allen, J (1995). *Natural Language Understanding*. Redwood City, Ca.: Benjamin/Cummings Publishing.

Ariel, M. (1990). Referring expressions and the +/- coreference distinction. In T. Fretheim and J. Gundel (Eds.), *Reference and Referent Accessibility*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. ASK for information retrieval: Part I background and theory. *Journal of Documentation* 38 (2): 61-71.

Bonzi, S. (1991). Representation of concepts in text: a comparison of within-document frequency, anaphora, and synonymy. *Canadian Journal of Information Science*. 16 (3): 21-31.

Bonzi, S. & Liddy, E. (1988) Testing the assumption underlying use of anaphora in natural language tests. *Proceedings of the 51st annual meeting of the American Society for Information Science* (25): 23-30.

Bonzi, S.& Liddy, E. (1989). The use of anaphoric resolution for document description in information retrieval. *Information Processing and Management*. 25 (4): 429-441.

Byron, D. & Tetreault, J. A flexible architecture for reference resolution. [Online]. Available: http://www.cs.rochester.edu:80/u/dbyron/papers.html [2000, March 15]

Byron, D.& Allen, J. (1998). Resolving demonstrative anaphora in the TRAINS93 corpus. In *Proceedings of DAARRC2 - Discourse, Anaphora and Reference Resolution Colloquium*, Lancaster University, August, 1998.

Callan, James P. (1994). Passage-Level Evidence in Document Retrieval. In *Proceedings of the Seventeenth International Conference on Research and Development in Information Retrieval* (SIGIR'94): 302-310.

Carbonell, J. G. & Brown, R.D. (1988). Anaphora resolution: a multi-strategy approach. *Proceedings COLING '88:12th International Conference on Computational Linguistics*: 96-101.

Carter, D. (1987). *Interpreting anaphors in natural language texts*. Chichester, England: Eliis Horwood Limited.

Charniak, E. & Hobbs, J. (1999) Two techniques of natural language processing: Statistical and knowledge-based. [Presentation] Knowledge: Creation, Organization, and Use: ASIS 1999 Annual Conference. Oct. 31 - Nov. 4, Washington, DC.

Cohen, P.R. (1992). The role of natural language in a multimodal interface. *Proceedings of the ACM Symposium on User Interface Software and Technology*:143-149.

Craven, T. C. (1988). Sentence dependency structures in abstracts. *Library and Information Science Research* 10 (4):401-410.

Denber, M. (1998). Automatic resolution of anaphora in English. Technical report, Eastman Kodak Co. [Online] http://www.wlv.ac.uk/~le1825/download.htm [1999, September 30].

J. Dorrepaal (1990). Discourse anaphora. In *Proceedings COLING 1990: 13th International Conference on Computational Linguistics* (2): 95-99.

Garrod, S., Freudenthal, D. & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language* 33: 39-68.

Gundel, K. (1996). Relevance theory meets the giveness hierarchy: an account of inferrables. In T. Fretheim and J. Gundel (Eds.), *Reference and Referent Accessibility*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

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

Hearst, M. & Plaunt, C. (1993). Subtopic Structuring for Full-Length Document Access. In *Proceedings of the Sixteenth International Conference on Research and Development in Information Retrieval* (SIGIR'93): 59-68

Hirst, G (1981). Anaphora in natural language understanding: a survey. In G. Goos & J. Hartmanis (Eds.), *Lecture Notes in Computer Science*. Berlin, Heidelerg, New York: Springer-Verlag.

Hobbs, J. (1978) Resolving pronoun references. In B. Grosz, K. Sparck-Jones & B. Webber (Eds.), *Readings in Natural Language Processing*, Los Altos, Ca.: Morgan Kaufmann, 1986.

Johnson, F.C., Paice, C.D., Black, W.J. & Neal, A.P. (1997) The application of linguistic processing to automatic abstract generation. In K. Sparck-Jones and P. Willet (Eds.), *Readings in Information Retrieval*, San Francisco: Morgan Kaufman.

Johnson, M. & Kay, M. (1990) Semantic Abstraction and Anaphora. *Proceedings COLING 1990: 13th International Conference on Computational Linguistics* (1): 17-27.

Kennedy, C. & Boguraev, B. Anaphora for everyone: pronominal anaphora resolution without a parser. *Proceedings COLING '96: 16th International Confernce on Computational Linguistics*: 113-118.

Lappin, S. & Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Association for Computational Linguistics* 20 (4): 535 -561.

Liddy, E. (1998). Enhanced text retrieval using natural language processing. *American Society for Information Science Bulletin* 24 (7). [Online] Available: http://www.asis.org/Bulletin/Apr-98/liddy.html [2000, February 29].

Liddy, E. (1989) Anaphora in natural language processing and information retrieval. *Information Processing and Management* 26 (1): 39-52.

Liddy, E., Bonzi, S., Katzer, J. & Oddy, E. (1987) A study of discourse anaphora in scienctific abstracts. *Journal of the American Society for Information Science* 38 (4): 255-261.

Losee, R. M. (1999) Natural language processing in support of decision-making: Phrases and part-of-speech tagging. (in press)

Marx, M.& Schamandt. C. (1994) Putting people first: specifying proper names in speech interfaces. *Proceedings of the ACM Symposium on User Interface Software and Technology*: 29-37

Mitkov, R.& Schmidt, P. (1998) On the complexity of pronominal anaphora resolution in machine translation. [Online] Available: http://www.wlv.ac.uk/~le1825/pubs2.html#aj. [1998, November 25].

Mitkov, R. (1998) Tutorial - anaphora resolution: recent developments, future directions. [Online] Available: http://www.rali.iro.umontreal.ca/COLINGACL98/tutorials/mitkov/long.html [1998, November 25].

Mitkov, R. (1997) *Recent advances in natural language processing.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Mitkov, R. (1995) Two engines are better than one: Generating more power and confidence in the search for the antecedent. In R. Mitkov and N. Nicolov, (Eds.). *Recent Advances in Natural Language Processing*. Amsterdam, Philadelphia: John Benjamins Publishing Co.

Moon, S.B. (1993). *Enhancing Retrieval Performance of Full-Text Retrieval Systems Using Relevance Feedback*. Ph.D. thesis, U. of North Carolina, Chapel Hill, NC.

O'Conner, J. (1973). Text searching retrieval of answer-sentences and other answer-passages. *Journal of the American Society for Information Science* 24 (6): 445-60.

Passonneau, R. (1993) Getting and keeping the center of attention. In M. Bates & R. Weischedel (Eds.), *Challenges in Natural Language Processing*. Cambridge: Cambridge University Press: 179-227.

Pirkola, A. & Jarvelin, K. (1996) The effect of anaphor and ellipsis resolution on proximity searching in a text database. *Information Processing and Management* 32 (2): 199-216.

Rinck, M.; Bower, G.H. (1995) Anaphora resolution and the focus of attention in situation models. *Journal of Memory and Language* 34: 110-131.

Shaw, Jr., W. M., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13, 347-366.

Sidner, C. (1986) Focusing in the comprehension of definite anaphora. In B. Grosz, K. Sparck-Jones, B. Webber (Eds.), *Readings in Natural Language Processing*, Los Altos, Ca.: Morgan Kaufmann, 1986.

Sparck Jones, K. Assumptions and issues in text-based retrieval. 157-177.

Van Hoek, K. (1997) *Anaphora and conceptual structure*. Chicago and London: The University of Chicago Press.

Webber, B. (1991) Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes* 6 (2): 107-135.

Webber, B. (1986) So what can we talk about now? In B. Grosz, K. Sparck-Jones, B. Webber (Eds.), *Readings in Natural Language Processing*, Los Altos, Ca.: Morgan Kaufmann, 1986.

Webber, B. (1979) *A formal approach to discourse anaphora*. New York & London: Garland Publishing.

Wolters, M.; Byron, K. Functions of prosody for pronominal anaphora. Under review for COLING 2000. [Online]. Available: http://www.cs.rochester.edu:80/u/dbyron/papers.html. [2000, March 30].

Appendix A – Sample of Raw Output from Perl Script

```
-------------------------------------------------------------------------------
Results for .I00021 from cf392.ful:
-------------------------------------------------------------------------------
SEN   DEM   WORD NO.          A TERMS     WORD NO     B TERMS     WORD NO
-------------------------------------------------------------------------------
[1]   xxx
                              cystic      15          cystic      15
[2]   These preparations 1
[3]   these tabletsor   7
                                                      child 2
                                                      food  23
[4]   This procedure    1
[5]   xxx
[6]   xxx
                              asthma      15
                              Cystic      4           Cystic      4
                                                      adult 8
[7]   xxx
                              cystic      7           cystic      7
                                                      child 5
[8]   this material.    38
                              asthma      9
                                                      child 6
[9]   xxx
                                                      food  14
[10]  xxx
[11]  xxx
[12]  these required    19
[13]  xxx
                              asthma      25
                                                      food  23
[14]  xxx
[15]  xxx
                                                      powders     17
[16]  xxx
                                                      food  12
[17]  those from  7
[18]  these measures    4
[19]  these       6
[20]  these materials.  25
[21]  xxx
[22]  xxx
[23]  xxx
[24]  xxx
[25]  xxx
                              cystic      9           cystic      9
                                                      child 7
[26]  xxx
[27]  xxx
                              cystic      4           cystic      4
[28]  xxx
                              asthma      5
[29]  xxx
                              cystic      9           cystic      9
```

Appendix B: Sample of Perl Script Output with Paragraph Boundaries

```
-------------------------------------------------------------------------------

Results for .I00059 from cf392.ful:

-------------------------------------------------------------------------------
SEN   DEM         WORD NO.    A TERMS     WORD NO     B TERMS     WORD NO
-------------------------------------------------------------------------------
[1]   this serum  19
                        PHOSPHATASE 3
                        ALKALINE    2
                                    age   24
                                    and   22
                                    bone  28
                                    liver 9
                                    liver 9
                                    enzyme      21
[2]   xxx
                        CF    18    CF    18
                        cystic      18    cystic      18
                                    and   14
[3]   These patients    1
                                    liver 5
                                    liver 5
                                    abnormalities     7
--P2----
[4]   that the    7
                        CF    21    CF    21
                                    isoenzymes  25
                                    liver 9
                                    liver 9
                                    enzyme      4
                                    tests 28
[5]   this serum  8
[5]   these patients.   30
                  CF    5     CF    5
                                    gel   19
                                    electrophoresis   20
                                    polyacrylamide    19
                                    and   21
                                    enzyme      13
--HEADER---
[6]   xxx
---P3----
[7]   xxx
[8]   xxx
                                    age   20
                                    and   17
                                    sex   22
[9]   these investigators     6
                                    age   30
[10]  These limits      1
                                    age   13
[11]  this series.      22
                                    and   5
```

```
[12]  xxx
                                    and   9
                                    tests 3
----P4-----
[13]  xxx
                                    gel   15
                                    electrophoresis   8
                                    polyacrylamide    14
                                    enzyme      2
[14]  xxx
                        CF    5     CF    5
                                    gel   12
                                    and   6
[15]  xxx
                                    age   38
                                    electrophoresis   2
                                    and   4
[16]  xxx
                                    and   4
[17]  this staining     4
[17]  that normal 8
                                    age   22
                                    electrophoresis   33
                                    and   15
                                    bone  13
[18]  xxx
                                    and   2
                                    bone  11
                                    enzyme      8
------P5------
[19]  xxx
                        CF    6     CF    6
                                    and   10
[20]  these 146   2
[21]  xxx
                                    and   9
                                    diagnosis   14
------P6-----
[22]  xxx
[23]  this paper. 9
---HEADER---
[24]  xxx
-----P7----
[25]  xxx
                        CF    6     CF    6
                                    male  8
                                    female      8
[26]  xxx
[27]  xxx
                                    age   2
                                    and   10
[28]  xxx
                                    age   9
[29]  xxx
--------P8-----
[30]  xxx
                        CF    4     CF    4
```

```
                                    and   43
                                    toxic 34
                                    cirrhosis   42
---HEADER---
[31]  xxx
---p9---
[32]  xxx
[33]  xxx
[34]  these five  3
                                    and   11
[35]  xxx
[36]  these patients,   7
                                    and   21
                                    cirrhosis   23
---P10----
[37]  this studypopulation    6
                                    failure     16
                                    heart 16
                                    diagnosis   12
[38]  these seven 7
                                    and   15
                                    liver 3
                                    liver 3
                                    cirrhosis   18
----HEADER---
[39]  xxx
-----P11-----
[40]  xxx
[41]  xxx
[42]  xxx
                                    age   20
                                    and   25
                                    toxic 19
---FIGURE/TABLE---
[43]  xxx
                        CYSTIC      3     CYSTIC      3
                        PHOSPHATASE 6
                        ALKALINE    5
                                    AGE   82
                                    CIRRHOSIS   35
                                    FAILURE     44
                                    HEART 43
[44]  xxx
                        cystic      18    cystic      18
                        phosphatase 3
                        alkaline    2
                                    male  7
                                    female      12
                                    and   10
[45]  xxx
                                    age   16
                                    and   20
[46]  xxx
                                    age   10
[47]  xxx
---P12----
[48]  this therapy      16
```

```
[49]  xxx
                                    male  2
                                    liver 8
                                    liver 8
[50]  xxx
                                    and   19
                                    toxic 16
[51]  xxx
                                    and   4
[52]  xxx
                                    and   14
                                    bone  21
---HEADER---
[53]  xxx
----P13----
[54]  xxx
                                    and   14
                                    cirrhosis   18
[55]  xxx
                                    and   8
                                    Liver 1
                                    Liver 1
                                    cirrhosis   4
[56]  xxx
                                    age   13
[57]  xxx
                                    male  7
                                    age   18
[58]  xxx
                                    and   2
                                    liver 6
                                    liver 6
                                    tests 8
---HEADER---
[59]  xxx
---P14---
[60]  xxx
[61]  xxx
                                    male  3
                                    female      3
[62]  xxx
                                    and   10
[63]  This is     1
                                    age   12
                                    failure     27
                                    heart 26
[64]  this were   3
[65]  xxx
                                    male  2
                                    female      2
                                    age   4
                                    and   8
---P15----
[66]  these patients.   12
[67]  xxx
[68]  xxx
                                    and   2
                                    liver 8
                                    liver 8
[69]  xxx
[70]  xxx
[71]  xxx
[72]  this elevation    12
[73]  these 22    19
                                    and   39
                                    liver 16
                                    liver 16
                                    failure     46
                                    heart 45
[74]  this group  4
[75]  these patients    3
[76]  xxx
[77]  xxx
                                    liver 2
                                    liver 2
                                    cirrhosis   9
--HEADER--
[78]  xxx
---P16----
[79]  xxx
                                    and   3
[80]  xxx
                                    and   10
                                    bone  17
[81]  xxx
                                    age   8
[82]  this agent  30
                                    age   31
                                    and   32
                                    enzyme      34
                                    abnormalities     34
[83]  these cases 14
                                    age   25
--HEADER---
[84]  xxx
---P17---
                                    enzyme      1
[85]  xxx
                        CF    7     CF    7
                                    gel   18
                                    electrophoresis   19
                                    polyacrylamide    17
                                    and   8
                                    enzyme      15
[86]  xxx
---TABLE--
[87]  xxx
                        CF    6     CF    6
                        Phosphatase 22
                        Alkaline    15
                                    Age   36
                                    and   7
                                    Liver 45
                                    Liver 45
                                    Sex   37
[88]  xxx
[89]  xxx
                                    liver 9
                                    liver 9
[90]  xxx
                        CF    54    CF    54
                        Phosphatase 2
                        Alkaline    1
                                    Age   18
                                    electrophoresis   110
                                    Bone  38
                                    Liver 35
                                    Liver 35
                                    enzyme      3
---P17(con)---
[91]  These control     1
                        CF    23    CF    23
                                    age   16
                                    and   44
---P18----
[92]  this study  11
                                    and   7
[93]  that cor    20
                        CF    37    CF    37
                                    liver 28
                                    liver 28
                                    failure     12
                                    heart 11
[94]  xxx
                                    and   3
                                    liver 7
                                    liver 7
                                    cirrhosis   11
[95]  This is     1
[95]  this which  9
[95]  that AP     12
                        CF    42    CF    42
                                    and   27
                                    toxic 49
                                    cirrhosis   30
                                    failure     46
                                    heart 45
                                    diagnosis   54
--P19----
[96]  xxx
                                    toxic 11
[97]  This therapy      1
                        CF    9     CF    9
[98]  these examples    1
[98]  that CF     8
                        CF    9     CF    9
                                    toxic 19
[99]  this group  3
                                    and   9
[100  that or     4
                                    age   8
                                    and   19
                                    liver 17
                                    liver 17
---HEADER---
[101  xxx
---P20--
[102  xxx
                        CF    17    CF    17
                                    and   3
                                    toxic 5
[103  xxx
[104  xxx
                                    cirrhosis   11
--FIGURE--
[105  xxx
[106  xxx
                        CF    14    CF    14
                                    gel   2
                                    Polyacrylamide    1
                                    and   13
                                    bone  24
                                    enzyme      4
[107  xxx
                                    gel   8
                                    and   9
--P21----
[108  that most   6
[108  this group  21
                        CF    8     CF    8
                                    cirrhosis   16
[109  xxx
                        CF    11    CF    11
                                    cirrhosis   8
[110  xxx
                        CF    15    CF    15
                                    toxic 22
                                    cirrhosis   29
                                    failure     18
                                    heart 18
                                    diagnosis   26
---P22---
[111  this progression? 45
                                    factors     43
                                    and   19
                                    cirrhosis   18
                                    factors     43
[112  these questions   1
[112  that chemicalevidence   8
                                    age   20
                                    cirrhosis   13
```