Essays on Language Function and Language Type

Dedicated to T. Givón

EDITED BY Joan Bybee, John Haiman and Sandra A. Thompson

OFFPRINT

Participant and Event Anaphora in Newspaper Articles*

Carol Lord
California State University, Long Beach

Kathleen Dahlgren
Intelligent Text Processing, Inc., Santa Monica

1. Introduction

The ways in which people use anaphoric reference in speech and writing have been studied by linguists, psychologists, philosophers, and computer scientists among others. If our goal is to understand how people use language for anaphoric reference, to build a theoretical model, or to design a computational algorithm for anaphora resolution, we need to have a clear picture of the range of anaphor phenomena.[1]

In an expository text the first mention of an individual entity is typically an indefinite noun phrase (preceded by the indefinite article *a*)or a proper noun (a name). The noun phrase *refers* to an entity in a model of the world; the entity is called the *referent*. A subsequent noun phrase in the text can refer to the same entity. This subsequent mention, called an *anaphor*, is typically a pronoun demonstrative pronoun, noun phrase with demonstrative adjective, or noun phrase with the definite article *the* or a possessive modifier. The previous mention of the referent is called the *antecedent* of the anaphor.[2] A text may contain several noun phrases referring to the same entity, creating a chain of reference. A comprehensive account of anaphora would require information on all anaphor forms (including pronoun, demonstrative, and full noun phrase), for

all referent categories (objects, events, states, and propositions, for example), in a wide range of languages, in all discourse modes and types, ranging from spontaneous spoken conversation to formal speeches, novels, newspaper articles, and recipes, to name only a sampling. Our concern here is with all forms of anaphoric reference in English, in all referent categories, but in a single genre.

A number of corpus-based studies have contributed to our current understanding of anaphor phenomena. Most studies of anaphora in English have been restricted to pronominal anaphors, often just personal pronouns (Hobbs 1976; Fox 1987a, for example). A few studies have looked at demonstratives (Linde 1979; Brown 1983; Webber 1988; Passonneau 1989; Gundel et al. 1989 and Ariel 1990). Full noun phrase anaphors have been largely neglected (but see DuBois 1980; Givón 1983; Guindon et al. 1986 and Ariel 1990). To date, studies have focused primarily on anaphors which refer to individuals rather than to events, states, or propositions. Few studies have investigated anaphors with antecedents in the form of clauses or multi-clausal chunks of previous discourse; exceptions include Schiffman (Passonneau) (1984), Schuster (1988), and Webber (1988).

Much work on anaphora has focused on English, but languages other than English have been examined by Clancy (1980), Li and Thompson (1979), Kirsner (1979), Kirsner and Van Heuven (1988), Pu (1995), and the many studies in Givón (1983), Grimes (1978), and Hinds (1978). Some have looked at a particular discourse mode or genre. Researchers have studied spoken narratives (DuBois 1980; Clancy 1980; Tomlin 1987 and the studies in Grimes 1978), spoken conversations (Reichman 1978; Givón 1983; Passonneau 1989), spoken descriptions (Linde 1979), task-oriented 'conversations' using the computer keyboard as medium (Grosz 1977), and newspaper articles (Hinds 1977). Data from a range of genres has been studied by Fox (1987a) and Webber (1988).

We report here on a survey of anaphora phenomena in a corpus of news articles from the *Wall Street Journal* newspaper. This study addresses the following questions: Is there a uniform constraint on the 'search space' for antecedents for all anaphor forms and all referent types? Are anaphora constraints affected by genre? For this corpus, how do discourse structure and the concept of global topic affect the form and distribution of anaphora? How are the answers to these questions relevant for computational approaches to anaphora resolution? In comparison with previous studies, our current interest is broader in that it concerns all anaphor forms and referent types; it is focused by being restricted in terms of genre to newspaper article texts.

Our method in this survey was guided in part by the findings of previous studies. Section 2 reviews the findings of previous studies. Section 3 describes the corpus and assumptions of the present study. The results are presented and discussed in sections 4–8.

2. Previous studies

In referring to a previously-mentioned entity, a speaker/writer must make a pragmatic choice between a nominal with a richer semantic content (a noun) and a nominal with a leaner one (a pronoun) (Bolinger 1979). By using a pronoun, the speaker is, in effect, telling the listener that he/she should be able to identify the referent in question without further information (Clancy 1980). A concept may be assumed by the speaker/writer to be present in the addressee's consciousness because of recent mention, presence in the immediate physical context, or by association with some other concept by means of a bridging inference; a discourse segment boundary marks a shift in the set of items in the speaker's consciousness, and accordingly affects the choice of anaphor form (Chafe 1976, 1987; Hinds 1978; Grosz 1977; Reichman 1978, 1985; Fox 1987a; Tomlin 1987; Givón 1983, 1989; Ariel 1990; Pu 1995). Effective use of language requires the speaker/writer to keep track of what is going on in the addressee/reader's mind. A concept can be assumed to remain in the addressee's consciousness from one sentence into the next, but it may be assumed by the speaker to have evaporated after a sentence without the concept has intervened; the speaker's assumptions show up in intonation and choice of anaphor form (Chafe 1974). Evidence from psychological studies suggests that a pronoun may be used when a concept remains in short-term memory, but that definite full noun phrases are used when the concept is no longer in short-term memory (Guindon 1985; van Dijk and Kintsch 1983; Givón 1995).

Discourse referents are concepts or entities which have been introduced into a discourse. *Individual* discourse referents are those whose denotations are concrete individuals (*the man*). They are distinguished from referents whose denotations are events, states, facts, propositions or properties (*the move, the statement*). Following Asher (1989), we call these references *abstract* because their individuation is problematic. Discourse events can also be sums of events; they can stand for the content of sequences of clauses in a discourse (Asher 1993).

Previous studies of anaphora beyond the sentence level have identified a number of factors which are important for anaphor form and distribution. The relevance of number, gender, and animacy, marked morphologically on anaphors in English, is uncontroversial. Other factors which have been recognized include semantic role, grammatical relation, number of intervening referents between anaphor and antecedent, parallelism, repetition, and word order. Some treatments recognize the combined effect of syntactic and semantic factors in defining a local topic or focus (Grosz 1977; Sidner 1983; Grosz, Joshi and Weinstein 1983; Brennan, Friedman and Pollard 1987). Many of these factors have been recog-

nized in computational approaches to anaphora resolution (Hirst 1981a). For intra-sentential anaphora, there has been detailed study of reflexives, disjoint reference constraints, and anaphors in the scope of quantifiers and negatives (Reinhart 1983; Bosch 1983; Hawkins 1978, for example). We also recognize the importance of plausibility, based on discourse context and naive semantics (Dahlgren 1988), including world knowledge and selectional restrictions, for resolving anaphors. As Hirst (1981a: 49) has observed, "it seems that an anaphor resolver will need just about everything it can lay its hands on..."

We are concerned here with the following factors:

- recency of previous mention of referent;
- discourse structure (segment hierarchy);
- global topic;
- anaphor form (pronoun, demonstrative, full noun phrase);
- heaviness (phonological size) of anaphor;
- referent type (individual or abstract object such as event, state, proposition);
- genre.

We consider the non-independence of these factors in relation to each other.

3. Corpus and definitions

The texts in our corpus were representative of a genre of expository written English described in this section. We found that an explanation of anaphor patterns in the texts required the concepts *global topic* and *discourse segment*. These concepts are defined and discussed in sections 3.2 and 3.3; their significance for patterns of anaphora is discussed in sections 6–8.

3.1. *The corpus*

The texts used for the study are 22 articles from the *Wall Street Journal* newspaper, with a total of about 20,000 words.[3] These articles typically report a news event and provide a discussion of its background and significance along with comments from participants, observers, and/or knowledgeable sources. In our corpus the salient event is typically a government agency action, new product announcement, agreement which has been reached, or other political/economic event. It need not be an actual past event; it can be, for example, an announcement of what an institution plans to do in the future. The articles in this genre include both news and commentary. The genre does not include editorials or strict event reportage; editorials belong to a separate genre of rhetoric or argument texts.

3.2. *Global topic*

For each article, there is an event or situation which is assumed to hold interest for readers; the reporter gathers details, background information, and comments from various sources and provides analysis. This event, which we call the global topic, is usually identified in the article's first sentence. The global topic is typically what is conveyed in a one-sentence reply to the question, "What happened?" Examples from the corpus include:

1. Brazil suspends interest payments on debt to banks.
2. Bally agrees to buy back Trump's stock.
3. Writers strike TV studios.
4. Korean factory managers don't like American imports.

The reason for the article can be a state, not an event, as in number 4 above. (A list of global topics for the articles is provided in Appendix A below.)

Global topic is a genre-relative construct; some genres may not require global topics (for example, casual conversation); for some genres the global topic is not necessarily an event. Different genres may have different patterns of anaphoric reference to the global topic entity. In an oral narrative, the hero may be the global topic. A 'thematic' strategy of reference has been found in such narratives, in which inexplicit forms (often pronouns) are used for the hero of the story (Clancy 1980; Grimes 1978). In another pattern, found in oral narrative in a Quechuan language, major participants, minor participants, and a central character with special status are each distinguished from the other by anaphor form (Levinsohn 1978). The narrative genre has its own discourse structure, and its own patterns for referring to the hero. Other genres, such as oral descriptions, task-oriented dialogs, and written news commentary, have other discourse structures and other patterns for reference to a global topic. Typically, the global topic is a person in narrative, an object in task-oriented dialog, and an event in news commentary. The narrative's hero is 'on stage' for much of the story, so frequency and recency of mention, as well as local topicality, may have correlations with the pattern of anaphora. In contrast, a task-oriented dialog may be organized around subtasks which provide local topics. In comparison with other genres such as narrative and dialogs, the news commentary genre provides

a valuable source of potentially contrasting data with respect to global topic and anaphora constraints.

For the purposes of this study, we divided the discourse events into three groups: *global topics*, *topic-related events*, and *peripheral events*. Topic-related events are those which cause or lead up to the global topic, or which are sequels to or consequences of it. The article with global topic number 3 above, for example, mentions that, following the writers' strike, and probably as a consequence of it, the networks lost viewers and lost money. Discourse events related to the global topic in these ways we have called topic-related events

It is to be expected that all the events reported in a given article will be relevant to the global topic in some way. However, when an event (or state) is not a precursor or sequel with some causal relationship, we have classified it as peripheral (an example, from the article with topic 2 above, is *Mr. Trump paid between $18 and $20.75 for his Bally shares*). Global topic, topic-related, and peripheral events can all be referred to anaphorically.

Each global topic statement includes noun phrases which name the relevant people and objects, the participants in the event. These typically are arguments of the verb used to report the event. We use the term *global topic participant* to refer to an individual or entity named by a noun phrase in a global topic statement. Similarly, the term *topic-related event participant* refers to an individual or entity named by a noun phrase in a topic-related event. All other individuals are *peripheral event participants*. For simplicity, we use the blanket term *topical* to refer to events and participants which belong to either the global topic or topic-related set; all other discourse entities are called *peripheral*.

Newspapers have editors who ensure that articles report events and comment on them. Thus, for news commentary articles, global topic structures are, in effect, institutionally maintained. This fact distinguishes this genre from others, which may define their global topics differently or may lack them altogether. Our perceptions of global topic in published articles correspond to guidelines given in news writing seminars by *Wall Street Journal* writer William E. Blundell. According to Blundell (1986: 70), the most common type of story "centers on some kind of occurrence and its consequences" — that is, a global topic event and topic-related events.

3.3. *Discourse coherence and segmentation*

Coherent text is structured. The nature of this structure, as intended by the speaker/writer and as recovered by the hearer/reader, is controversial. However, there is general agreement that speakers tend to use more informative or explicit anaphor forms at the beginning of new structural units. As a first approximation

of these structural units, we segmented each news article into chunks according to our intuitions as to where the structural breaks or changes in subtopic occurred. In identifying segment boundaries, we disagreed on only two of the 140 boundaries in the corpus. Although some texts are better-constructed than others, and individual readers vary in the world knowledge, language skills, and genre experience that they bring to the reading task, we expect that other readers would substantially agree with our segmentation. Each of the chunks we identified is characterizable in terms of propositional content: each is 'about' something; each has its own distinct topic. Furthermore, we found that each topic was related to the global topic of the text in terms of logical or rhetorical function. The resulting hierarchical structures are shallow trees. The segments correspond in general to the outer edges of a Grosz and Sidner (1986) text analysis, or to the larger spans in a Mann and Thompson (1987, 1988) analysis. The formal and theoretical bases of discourse coherence assumed in this study are elaborated in Dahlgren (1988). The theory draws upon the work of Hobbs (1985), Hirst (1981b), van Dijk and Kintsch (1983), and Mann and Thompson (1987, 1988). In general, a discourse is coherent because the reader can relate the discourse events or situations to each other in terms of plausible links or chains based on causal, intentional, part-whole or other connections. The various types of connections can be identified as a set of coherence relations (Dahlgren 1996). The set of coherence relations found in the corpus is shown in Appendix B. Appendix C shows the segmentation of one of the articles in the corpus.

Newspaper articles differ from some genres in that they are planned and edited to provide a maximum amount of information in a compact package. A reasonable criterion for a well-written news article is the degree to which it is organized into sections, or segments, which the reader perceives as fitting together so that the article makes sense. In this respect, the articles in the corpus are generally well-written. The reader may or may not consciously make note of the segment boundaries, but a recognition of these segments at some level is an implicit part of the process of understanding how the sentences, paragraphs, and larger sections 'hang together.' This view of the genre is consistent with Blundell's experience as writer and editor (Blundell 1986: 101): "For years I examined pieces that seemed to me particularly well organized. I wanted to dope out why they were, and how they differed from others that seemed jumbled, confusing. The reason: Somehow the writer had succeeded in grouping material in the body of his story — the part after the main theme statement — into blocks of copy, each of them addressing a certain facet of the story." According to Blundell, the grouping of related material helps the writer meet the reader's demand for a clear, logical presentation that is convincing; when material is scattered, both logic and force are diminished.

4. Anaphors, antecedents and referents

In this section the range of the data is described in some detail, and correlations between anaphor form, referent type and antecedent structure are noted. Section 4.1 provides examples of antecedent structures. Sections 4.2, 4.3 and 4.4 illustrate anaphors in pronoun, demonstrative, and full noun phrase form. Clausal antecedents are described in section 4.5. The significance of conclusions drawn from the data is discussed below in sections 5–8.

In this study we identified each definite anaphor which had a referent whose closest previous mention was in the previous sentence or earlier in the text. (We did not deal with anaphors whose antecedent occurred within the sentence since these have been examined in some detail by others; we considered a sentence here to be defined by orthography: bounded by capital letters and periods.) We called definite anaphors those definite noun phrases which had as referent an individual or abstract object with a previous mention in the linguistic context.[4] We identified the distance between anaphor and antecedent as the number of sentence boundaries between the anaphor and the closest previous mention of its referent. Referents in the individual category were typically people, objects, or institutions. Most referents in the abstract category were events or states, but this classification also includes propositions and facts.

There were a total of 482 anaphors in the corpus. Of these, 19% were pronouns, 7% were demonstratives (demonstrative pronoun, or demonstrative adjective plus noun), and 74% were full noun phrases, with *the* or a possessive modifier. Seventy-five percent of all anaphors had individual referents; 25% had abstract referents. (See Table 1.)

As Table 1 shows, pronoun, demonstrative, and full noun phrase anaphor forms differed from each other with respect to the likelihood that their referents were individual or abstract objects; the distinction was highly significant statistically,[5] with $p < .0001$. Distinctions among anaphor forms and between referent types correlated with differences in antecedent location and antecedent properties, as discussed below. As Table 1 shows, individual referents were only 75% of the total number of anaphors, so any account of anaphora that omitted events and other abstract anaphora would be incomplete. Also, since only 19% of the anaphors were pronouns, any study limited to pronouns neglects a significant proportion of anaphoric phenomena.

4.1. *Antecedent structures*

Individual and abstract referent types differed in the range of possible linguistic structures for their antecedents. In the corpus, individuals were frequently

Table 1. Anaphor form and referent type

Referent type	Individual	Abstract	Total
Anaphor form			
Pronouns	86 (92%)	7 (8%)	93
Demonstratives	12 (36%)	21 (64%)	33
Full noun phrases	265 (74%)	91 (26%)	356
Total	363 (75%)	119 (25%)	482

people, institutions, or products; antecedent forms for individuals were noun phrases. In contrast to individuals, events can be encoded in a number of ways in English, for example, as noun phrases with event nominals as heads, as gerundive phrases, infinitival phrases, and finite clauses. We found this variety in the range of linguistic forms for the antecedents of event anaphors. Examples from the corpus are illustrated in Table 2.

The antecedent of an event anaphor is the discourse event, not a particular linguistic construction; there are many possible linguistic forms which can encode a discourse event.

A sequence of clauses can describe a situation or sum of events (Asher 1989). Anaphors can be used to refer to such sums of events. Thus the content of a sequence of clauses, that is, the discourse referents and predications they introduce, can be the antecedent of an anaphor. In one example of this, the first three paragraphs (five sentences) of an article describe the shooting down of two Libyan airplanes and related actions by US officials. The sixth sentence begins with an anaphor referring to the situation: *All of this is being closely watched by Europeans...* In this instance the antecedent of *this* is a discourse segment.

4.2. *Pronoun anaphors*

The correlation between anaphor form and referent type was strongest for pronouns. Most pronouns (92%) had individual referents rather than abstract object referents (this correlation could be genre-related to some extent, since careful writers typically try to avoid the vagueness of reference which can result when *it* is used to indicate an event or state). Although there were a few accusative pronouns *(them),* most were nominative, reflecting the tendency for topical information to occur early in the utterance, typically as subject. More than half of the pronoun tokens were *he;* 27% were *it/its,* and 19% were plurals. There were no instances of *she, her,* or *hers.*

Table 2. Antecedent structures for event anaphors

	Antecedent	Anaphor
Structure		
nominalization	Brazil's unilateral suspension of interest payments on its commercial foreign debt	Brazil's move
gerundive phrase	(Bally) agreeing to buy back... Trump's ... stake in the company	the agreement
infinitival complement	(The Securities and Exchange Commission is seeking to compel Drexel Burnham Lambert Inc.) ... to change its senior management structure	the changes
finite clause	the three major networks have lost four million nightly viewers	the erosion

4.3. *Demonstrative anaphors*

Of the anaphor forms, demonstratives were the group most likely to have abstract referents; only 36% referred to individuals. About half the demonstratives were simply the demonstrative pronouns *this, these, that, those,* and of these only 19% had individual referents... Examples from the corpus include *this (viewers are turning away)* and *that (the growing power of the military and the great corporations).*[6] The others were demonstratives with nouns ('demonstrative adjectives'); examples included *that chip (Intel's 80386 microprocessor chip)* and *those charges (Mr. Levine also had been accused by the SEC last May of making $12.6 million in illegal profits...).* For some comparisons we grouped together demonstrative pronouns and noun phrases containing demonstrative adjectives, since there were similarities in their distributions; a demonstrative pronoun can be thought of as a definite anaphor with an empty head.[7] Demonstrative pronouns with clausal antecedents showed correlations with proximal/distal and topical/peripheral distinctions, as discussed below in sections 7 and 8.

4.4. *Full noun phrase anaphors*

This section describes the form and meaning relationships found between full noun phrase anaphors and their antecedents. Full noun phrases, with *the* or a possessive modifier, were the largest group of definite anaphors; there was a range of subtypes. Of this group, 74% had individual referents, and 26% had abstract referents.

4.4.1. *Full noun phrase individuals*

The full noun phrase individual anaphors (265 total) included:

- copies of a previous mention of the referent (65%) (but with a definite determiner or possessive modifier),[8] for example, *the affidavit (an affidavit);*
- synonyms (9%), for example, *the minority companies (minority small businesses);*
- groups or deictic terms (4%), as in *both countries (Japan... the US)* or *the latter (the generating capacity a utility sets aside for the company's needs);*
- other descriptive terms (22%), for example, *the defendants (the three men arrested last month),* or *the newspaper (the Dartmouth Review),* where the anaphor names the ontological class of the proper noun.

For individual non-copy anaphors, slightly more than half the antecedents were proper nouns. Sometimes several different full noun phrases were used to refer to a single referent (this may be the result of the writers' conscious efforts to avoid repetition); for example, in one article the words *holding, stake,* and *shares* were all used to refer to the same entity.

4.4.2. *Full noun phrase events*

There was a range of form and meaning relationships between the full NP event anaphors and their antecedents (see Table 3 for examples). The full NP event anaphors (91 total) included:

- copies of the noun (57%);
- synonyms or other characterizing nominals (35%; cognates — similar words showing historical relationship — were 28% of this group);
- superordinates or general words (8%) (Halliday and Hasan 1976).[9]

Of the 91 instances of event anaphora using full noun phrases, 73% of the antecedents were encoded as nominals; the others were encoded as clauses, phrases other than nominals, or clause groups. (Clausal antecedents are described in the next section.) The non-copy full NP anaphors are presumably more challenging for people and for computers; of this group, more than half (62%) had clausal antecedents (see Table 4).

Table 3. Full NP event anaphor/antecedent relationships

	Antecedent	Anaphor
Relationship		
copy	a judge's decision	the court's decision
cognate	most of the details of the agreement were negotiated late Thursday night by a Bally director	the negotiations
synonym	a battle began last November	his three month long bout with Bally
characterizing nominal	Individuals familiar with the negotiations were careful to avoid characterizing Bally's settlement with Mr. Trump as greenmail	the reluctance to use the word 'greenmail'
superordinate	Brazil's unilateral suspension of interest payments on its foreign debts	Brazil's move

Table 4. Event full noun phrase anaphors and antecedent structure

Antecedent Structure	Nominal	Clausal	Total
Anaphor Type			
Synonym or other noun	9 (39%)	14 (61%)	23
Cognate	2 (22%)	7 (78%)	9
Superordinate	4 (57%)	3 (43%)	7
Total	15 (38%)	24 (62%)	39

4.5. *Clausal antecedents*

Anaphors with clausal antecedents referred to events, activities, states, and propositions, described in this section. Table 5 provides examples.

Most of these anaphors were full noun phrases (66%). Some had antecedents comprising more than one clause. Anaphor forms included nominalizations and non-derived event nouns, including superordinates and other characterizing terms, for example, *the events, the move, the action, the development, the situation.*

Table 5. Clausal anaphors

Referent type	Antecedent	Anaphor
event	Just before noon in the Mediterranean yesterday, two US F-14 jets shot down two MiG-23 Libyan fighters...	the Mediterranean dogfight
activity	(...the US, Germany and Japan ... allow) their currencies to fluctuate...	the fluctuations
state	Individuals...were careful to avoid characterizing Bally's settlement with Mr. Trump as greenmail.	The reluctance to use the word "greenmail" to describe Bally's agreement

Clausal antecedents can have events, activities and states as referents, as illustrated in Table 5; they can also have propositions as referents. Some anaphors were nominalizations related to verbs taking *that*-clauses or nominalizations as objects (e.g., *Brazil's announcement, these charges, the allegations*). The verb complement (antecedent) content corresponded to the noun complement (anaphor) content. The anaphor's referent was either (1) the event/situation (the act of announcing, charging, or alleging) or (2) the propositional content of the complement (i.e., what was announced, charged, or alleged). Accordingly, for these anaphors the antecedent was either (1) a complement-taking verb or (2) its complement. From the instances we have studied, it appears that either the event/situation/activity named by the verb, or the content of the complement (a proposition), can serve as referent for a later anaphor; the choice between event or propositional reference can often be inferred by the reader from the context of the anaphor through selectional restrictions, world knowledge, or other semantic knowledge. In one article, both event and proposition referents were specified anaphorically by the noun *allegations* in separate sentences:

- event referent: *The allegations by OSHA follow a four-month investigation...*
- propositional referent: *...the company has just begun to study the OSHA allegations...*

Here the anaphor is the noun *allegations*. In the first sentence its referent is the event of alleging, inferrable since what follows an event is typically another event. In the second sentence the referent is the propositional content of the allegations, inferrable since typical objects of the verb *study* are more likely to be propositional than eventive.

In the corpus, anaphors in the form of nominalizations of complement-taking verbs could have either event or propositional referents, but there were other anaphors which had only propositional referents. These included non-derived nouns such as *the point, the line* where the antecedent was typically a quotation or a verb complement. Propositional anaphors were not limited to complement antecedents; in one instance the noun anaphor *reason* referred to the content of the preceding sentence. Demonstrative pronouns with clausal antecedents showed correlations with proximal/distal and topical/peripheral distinctions. Proximal pronouns *(this)* referred to events which were global topics or topic-related. Distal pronouns *(that)* referred to states which were peripheral rather than topical. Anaphoric reference to discourse segments was to topical segments, and typically employed proximal demonstratives. (The relevance of topicality for demonstrative choice is discussed further in sections 7 and 8.)

4.6. *Summary*

In section 4 we have surveyed the range of anaphor and antecedent forms in the corpus. We have distinguished between individual and abstract referent types and noted that their distribution is not uniform across anaphor forms. We have identified a variety of nominal and clausal antecedent structures, exemplified ways that anaphors and antecedents are related formally and ontologically, and described anaphors with propositional as well as eventive and stative referents. Sections 5, 6, and 7 discuss the study's findings with respect to distance from antecedent, discourse segmentation, and global topic.

5. Distance from Antecedent

We found that a discourse constraint such as recency has values which depend upon genre, topicality, discourse referent type, and segmentation. Similarly, choice of anaphor form is a function of genre, discourse referent type, segmentation and topicality. In the next three sections we show the interdependence of these factors. Here we show that recency is a function of anaphor form.

In our corpus, nearly all antecedents of pronouns were in the previous sentence. Full NP's could have antecedents several sentences back. Table 6 shows the distance from anaphor to previous mention in number of sentences for each anaphor form. Sixty percent of the anaphors in the corpus had antecedents in the preceding sentence, but distance between anaphor and antecedent varied

with anaphor form. For antecedents in a single sentence, there were statistically significant differences between all anaphor forms compared pairwise.

Table 6. Anaphor form and distance to antecedent (measured in number of sentence boundaries)

Distance	1 S	2 Ss	3 Ss	4 Ss	5 or more	Multi-S antecedent	Total
Anaphor form							
Pronoun	89 (96%)	2	1	1	-	-	93
Demonstrative	25 (76%)	2	-	-	2	4	33
Full NP non-copy	76 (61%)	13	10	5	16	5	125
Full NP copy	100 (43%)	33	22	19	57	-	231
Total	290 (60%)	50	33	25	75	9	482

The average distance between anaphor and antecedent was calculated for each anaphor form[10]. The number obtained represents the average number of sentence boundaries between anaphor and antecedent. The average distance for the entire corpus was 1.83 sentences. There was a hierarchy or continuum of anaphor types in terms of average distance to antecedent:

- *pronoun (1.06) < demonstrative (1.30) < full noun phrase non-copy (1.90) < full noun phrase copy (2.57).*

For 96% (all but 4) of the 93 pronouns, the antecedent was in the previous sentence. Each of the exceptions was unusual in some way. One was an idiomatic construction, and in the others the intervening sentences were direct quotations or information attributed to the pronoun's referent.[11] For this corpus, to find individual pronoun antecedents one need look back no farther than the previous sentence, except for idiomatic constructions and quoted or attributed material. This restriction was not found for other anaphor types.[12]

Most demonstrative anaphors had antecedents in the preceding sentence, as in

- *Mr. Trump paid between $18 and $20.75 for his Bally shares, including brokerage commissions. That indicates...*

Pronominal *that* typically referred to events encoded clausally. In an enlarged corpus, with over 50 additional articles,[13] there were 22 sentence-initial *that* pronoun subjects. Of these, 86% had clausal antecedents in the previous sentence, or multi-clausal antecedents where the event or situation was described in the preceding sentence sequence.[14]

Full NP copy anaphors were explicit references in which the head noun of the anaphor was a copy of the head noun of the antecedent; this group was likely to have antecedents two or more sentences back. In contrast, full NP non-copy anaphors were likely to have antecedents in the previous sentence.

Evidence from spoken data shows a correlation between the phonological size of an anaphor and the distance back to its antecedent, according to a scale of accessibility, starting with smaller size and shorter distance (Givón 1983):

- *zero anaphor < unstressed/bound pronoun < stressed/independent pronoun < full noun phrase*

The speaker/writer provides a greater amount of encoding material for anaphors which are judged to be more difficult for the hearer/reader to identify. This scale is extended (for Spanish, Bentivoglio 1983), with larger anaphors at a greater distance from the antecedent:

- *full noun phrase with definite article < full noun phrase with demonstrative/ possessive/ genitive/adjective or combination < full noun phrase with relative clause*

An extended scale for all languages, incorporating names and demonstratives, is proposed by Ariel 1990.

We found a general correlation between phonological size and distance to antecedent in the news texts. We identified 65 'heavy' anaphors in the corpus as those noun phrases three or more words long (not counting articles). The heavy anaphors ranged from three to 27 words in length, with a median length of five words. There was a general correlation between heavy anaphors and longer distances to the antecedent: within three sentences of the antecedent, only 10% of the anaphors were heavy, but at longer distances 25% of the anaphors were heavy ($p < .0004$). The correlation was stronger when we considered discourse segment boundaries rather than simply distance in number of sentence boundaries.

The correlation of heavy NP anaphors with longer distances from antecedents was complicated in the newspaper texts by the occasional use of anaphoric full noun phrases to introduce new information. In one article, a global topic event participant was introduced as *Bally Manufacturing Corp.* in the first sentence. Then, it was mentioned five more times as either *Bally* or *the company.* In the fourth sentence the sixth mention occurred; although a pronoun or a minimal definite noun phrase might have been expected at this point, the writer chose to introduce new information in the heavy noun phrase *the Chicago-based operator of casinos and hotels.* Heavy noun phrase anaphors can be used in this way to bring in new 'piggyback' information, or for classification, contrast or

emphasis (Fox 1987a). But an anaphoric structure by definition points back to familiar information, an entity previously introduced into the discourse. Using anaphoric structures to carry new information goes against the structure's established discourse function; this may explain why the phenomenon occurs in consciously crafted discourse where there are pressures to be informative and concise, but rarely in informal, unplanned discourse. If the purpose of using a heavy anaphor is simply to make it easier for the hearer to identify the antecedent, we would not expect to find a heavy anaphor within three sentences of its antecedent; yet this was often the case in the news commentary texts. However, of those heavy individual anaphors within three sentences of the antecedent, nearly half (44%) contained new information, compared with only 8 of the anaphors at greater distances ($p < .036$). The use of anaphor structures close to the antecedent for the purpose of introducing new information appears to be a genre-related phenomenon.

6. Discourse Segment Effects

In this section we address the question of a segmentation constraint on anaphora in news commentary articles. For pronouns, the antecedent is found within the same segment. This constraint applies to pronouns but not to other anaphor forms, and does not appear to restrict abstract object anaphors. We show that the choice of anaphor form is a function of segmentation. Since different genres may have different discourse structures and segmentation options (narrative, task-oriented dialog, and casual conversation, for example), this finding means that choice of anaphor form is a function of genre.

Anaphor form is correlated with discourse segmentation. The corpus shows a strong pattern of full noun phrases at the beginning of new discourse segments, even though a pronoun might otherwise be expected due to the recency of an antecedent. Most anaphors in the corpus (69%) had antecedents earlier in the same discourse segment. In general, individual pronouns were restricted to antecedents in the same segment, demonstrative pronouns with topical referents could have antecedents outside the segment, and full noun phrase anaphors were not restricted in terms of discourse structure or topicality of referent. Table 7 shows distance (in discourse segments) between anaphor and antecedent for each anaphor form.

The average distance between anaphor and antecedent for the whole corpus was .62 segments. In section 5 we reported a hierarchy of anaphor forms in terms of distance to antecedent. The same hierarchy order is preserved if we calculate average distance in number of discourse segments:

Table 7. Anaphor form and discourse structure

Distance (segments) to antecedent	Same segment	1	2	3	4	5 or more	Multi-seg antecedent	Total
Anaphor form								
Pronoun	92 (99%)	1	-	-	-	-	-	93
Demonstrative	25 (76%)	5	-	1	1	-	1	33
Full NP non-copy	79 (63%)	28	4	3	2	7	2	125
Full NP copy	135 (58%)	55	13	11	4	13	-	231
Total	331 (69%)	89	17	15	7	20	3	482

- *pronoun (.01) < demonstrative (.40) < full noun phrase non-copy (.72) < full noun phrase copy (.84)*

Considering distance to antecedent solely in terms of clause or sentence counts misses the crucial importance of discourse structure for anaphor patterns. Data from many sources show that, at the beginning of a discourse segment, a full noun phrase may be used, even though a pronoun would otherwise be expected because of the recency of the antecedent.[15] The pronoun distribution in our corpus corroborates this generalization. All personal pronouns in the study *(he, him, his, they, them),* a total of 68, had previous mentions within the same discourse segment. Of the 93 pronouns in our study, only one was in a segment-initial sentence (in this instance the pronoun *it* referred to the global topic event mentioned in the preceding sentence).

The data show that referent type is a function of segmentation. Anaphors with individual referents were more likely than those with abstract referents to have antecedents in the same discourse segment. This was true for the total collection and for each anaphor form individually. All pronouns with individual referents (88) had antecedents within the discourse segment. Anaphors with abstract referents tended to have antecedents outside the discourse segment (the distinction between individual and abstract objects was highly significant, with $p < .00005$). For individuals and abstracts, either lumped together or considered as separate groups, the likelihood of an antecedent in the same discourse segment varied according to the anaphor form hierarchy above, with pronouns most likely. Full noun phrase copy abstract anaphors typically had antecedents outside the discourse segment (67%).

Choice of demonstratives correlated with segmentation. Distal demonstratives *(that, those)* were more likely to have antecedents in the same discourse segment than were proximal demonstratives *(this, these).* See Table 8 (the distinction between distal and proximal was significant, with $p < .0152$).

In some instances it appears that an anaphor signals the beginning of a new discourse segment; the use of the anaphor may help define the previous segment. Proximal demonstratives frequently functioned in this way in the text. In one example, the article began with an anecdote segment (eleven sentences) describing the effect of a writers' strike on a new TV series. The second segment (eight sentences) presented the global topic of the article, the TV networks' loss of viewers as a result of the strike. The next paragraph began, *This is taking place while television watching in general is on the rise.* The demonstrative pronoun *this* began a new segment providing background information related to the topic.

A proximal demonstrative functioned similarly in other texts. In another example, the first segment was a brief introductory description, the second and third segments described the planned expansion of an airline company and discussed its import, and the third segment, an evaluation, began with a recapitulating anaphor: *These are heady plans...* In both of these examples the anaphor is a proximal demonstrative pronoun with a global topic referent, a frequent pattern (see the following section).

7. Global Topic Effects

The texts showed the concept of global topic to be related to the distance between anaphor and antecedent, location in discourse structure, the proximal/distal distinction for demonstratives, and multi-clausal antecedents. (Topical and peripheral referents are defined for the corpus in section 3.2 above.)

7.1. *Topicality and recency*

Peripheral anaphors had closer antecedents. For full noun phrase anaphors, only 5% of the anaphors with peripheral referents were five or more sentences away from their antecedents. In contrast, for the anaphors with topical referents, 25% were five or more sentences away from their antecedents. In a notable example of long-distance pronominalization of global topic, an article reported the TV networks' loss of viewers due to the writers' strike and referred to the loss anaphorically with the proximal demonstrative pronoun *this* at the beginning of a new segment in sentence #20 (*This is taking place...,* as noted in section 6 above). The next 29 sentences contained background information and commented on the strike's legacy. A new segment began in sentence 50 with another anaphoric reference to the loss of viewers, again taking the form of a sentence-

Table 8. Demonstrative type and discourse structure

Antecedent location	Same segment	Earlier segment	Total
Anaphor			
Distal demonstrative	19 (90%)	2 (10%)	21
Proximal demonstrative	6 (50%)	6 (50%)	12
Total	25 (76%)	8 (24%)	33

initial proximal demonstrative pronoun, and presented further elaboration on the loss of viewers; the distance between anaphor and most recent previous mention was 30 sentences.[16] The pattern is consistent with a tacit agreement between writer and reader that in this genre global topic referents are readily accessible, and accordingly, anaphors at long distances from topical antecedents are not a problem.

7.2. *Topicality and segmentation*

Although peripheral antecedents were usually inside the immediate discourse segment, topical antecedents were frequently outside it. For full noun phrase anaphors, only 19% of peripheral referents had antecedents outside the segment. In contrast, for a topical referent, the most recent mention was almost equally likely to be outside the discourse segment as inside it; see Table 9. (The topical/peripheral distinction was highly significant for full NP's, with $p < .00005$.)

For full noun phrase abstract anaphors, the topical/peripheral distinction was even greater: only 32% of topical abstract anaphors (compared to 67% of peripheral abstract anaphors) had antecedents inside the same discourse segment.

For demonstrative pronouns, although peripherals were limited to antecedents within the segment, most topicals (83%) had antecedents outside the segment, as Table 10 shows ($p < .0014$).

Table 9. Referent topicality and discourse structure (full NP anaphors)

Antecedent location	Same segment	Earlier segment	Total
Topicality of referent			
Topical	152 (54%)	127	279
Peripheral	62 (81%)	15	77
Total	214 (60%)	142	356

Table 10. Referent topicality and discourse structure (demonstrative pronouns)

Antecedent location	Same segment	Earlier segment	Total
Topicality of referent			
Topical	1 (17%)	5 (83%)	6
Peripheral	10 (100%)	0 (0%)	10
Total	11 (69%)	5 (31%)	16

Pronominal anaphoric reference outside the segment has been noted for the overall topic (the apartment) in apartment descriptions (Linde 1979), and for the main topic in task-oriented dialogs (Grosz 1977, Guindon et al. 1986) . Similarly, the patterns described here establish global topics for news commentaries as referents for anaphora with antecedent outside the discourse segment.

7.3. *Topicality and demonstrative type*

The choice of demonstrative type is a function of global topicality. Among demonstratives, topical referents were usually encoded as proximals (75%), compared to only 14% of peripheral referents (see Table 11; $p < .0009$). Seventy-five percent of proximal demonstratives had topical referents, but only 14% of distal demonstratives did.[17]

Peripheral events and participants are not salient elements in the texts. They are not closely related to the central concerns of the article. These entities, at a metaphorical distance from the writer's and reader's center of attention, are encoded with the distals *that* and *those*.[18] Global topic segments frequently were antecedents for proximal demonstrative pronouns, as exemplified above in section 6. When the antecedent of a demonstrative anaphor was a clause sequence or a discourse segment, the referent was topical in all instances.

With respect to antecedent location, a comparison of Tables 8, 10 and 11 shows that a restriction to the immediate discourse segment is correlated with distal demonstratives and peripherality. In contrast, reaching outside the segment is correlated with proximal demonstratives and topicality.

8. Implications

For researchers seeking an understanding of anaphora in natural language, this study indicates the importance of distinguishing among different anaphor forms and recognizing genre-relative influences such as discourse structure and global topic. Genres differ with respect to the degree to which discourse structure

constrains anaphoric reference, and also with respect to the availability of a topical entity as a potential anaphoric referent. There appears to be no single genre-independent account of anaphora resolution.

8.1. *Anaphor form and referent type*

Constraints on the search space for antecedents of anaphora are functions of anaphor form and referent type. Different anaphor forms have different ranges of distance to antecedent, and varying correlations with referent type, topicality of referent, and location within discourse structure.

For individual pronouns in the corpus, the antecedent was in the same discourse segment, no farther back than the previous sentence. Other anaphor forms were not restricted in this way. Demonstratives and full noun phrases could have antecedents at a distance of two or more sentences, and in previous discourse segments. Although the one-sentence limitation on pronoun distance appears to be a strong preference for this genre, it is probably not an absolute requirement. Pronouns will, however, show a stricter distance limitation than demonstratives and full noun phrases.

Different anaphor forms tended to have different referent types. Most pronouns had individual referents, most demonstrative pronouns had abstract referents, and full noun phrases had both individual and abstract referents, with individuals the majority.

For each anaphor form (pronoun, demonstrative, full noun phrase), the typical distance to the antecedent was greater for abstract referents than for individual referents. Abstract referents were more likely to have antecedents in a previous segment.

8.2. *Influence of genre*

A comparison of the results of this study with those of previous studies supports the conclusion that anaphor form and distribution are genre-dependent.

Table 11. Referent topicality and demonstrative type

Demonstrative type	Proximal	Distal	Total
Topicality of referent			
Topical	9 (75%)	3	12
Peripheral	3 (14%)	18	21
Total	12 (33%)	21	33

News commentary articles show a strict constraint on distance from pronoun to antecedent, allowing distances no greater than one sentence boundary. The strength of this constraint appears to vary according to genre: written expository texts have shorter average distances than spoken conversational texts, and written narrative texts have average distances between those for expository and conversational (Fox 1987a).

News commentary articles show a strong pattern of full noun phrase anaphors rather than pronouns at the beginning of discourse segments, consistent with the pattern observed in other studies (section 6). This pattern is apparently widespread in expository written text but not at all common in conversational text (Fox 1987a).

Anaphor patterns in our corpus were sensitive to the presence of a global topic, a genre-relative feature. The presence and nature of global topic depend on the genre structure; for example, informal conversations may not have global topics. In news commentary articles, anaphors with topical referents have greater average distances to their antecedents, and are more likely to have antecedents outside the segment. Genre differences may account for different patterns for anaphors with antecedents outside the segment. For spoken descriptions (Linde 1979) and task-oriented dialogs (Grosz 1977; Guindon et al. 1986), pronominal anaphoric reference to the overall topic is possible when the antecedent is outside the segment. In our corpus, this was very rare: there was only one instance of the pronoun *it* with antecedent outside the segment, and in this case the antecedent, with global topic event as referent, was nearby in the previous sentence. Anaphoric reference to global topic outside the segment employed a full noun phrase or the demonstrative pronoun *this*. The difference in global topic anaphor patterns in this corpus, compared to others, is probably due to genre differences — namely, written expository text with event as global topic vs. spoken description or dialog with individual as global topic (apartment or pump, for example).

8.3. *Demonstratives and topicality*

In this corpus, demonstratives in general, and demonstrative pronouns in particular, are likely to have abstract referents. Similarly, studies of other genres have found demonstrative anaphors used for propositional referents in spoken descriptions (Linde 1979), with segment antecedents in written narrative (Brown 1983), with clausal antecedents in spoken interviews (Passonneau 1989), and with clausal and clause sequence antecedents in various written genres (Webber 1988). The news commentary articles showed a difference in function between proximal demonstratives (*this*, *these*) and distals (*that*, *those*). In general, among

demonstrative anaphors, proximals had topical referents and distals had peripheral (i.e., non-topical) referents (Table 11). Typically, the anaphors with peripheral referents had antecedents in the same segment, while those with topical referents did not. The topicality correlation for proximal demonstratives results in proximal anaphors being farther from their antecedents, a situation which might be regarded as counter-intuitive. The resulting arrangement is the opposite of what Lyons (1977: 669) describes as the pattern in Latin and Turkish, where the proximal signals closeness in the textual environment as well as closeness in the non-textual environment. The proximal-topical correlation suggests that writers of commentaries use proximals to refer to significant, salient events and participants, and they use distals to refer to entities outside the area of primary attention.[19] The experienced reader presumably uses these topicality clues in identifying the referent of the pronoun. At the same time, the appearance of a proximal demonstrative may serve to reinforce for the reader the topicality of the referent; similarly, a distal signals to the reader that the referent is peripheral in the discourse context, and accordingly deserves no special attentional status. These patterns, we suggest, have become part of the tacit knowledge of the genre, shared by writer and reader, and contribute to the reader's comprehension of the text.

The findings here are at odds with Webber's (1988) claim that most proximal and distal forms can be used interchangeably. If the forms are truly interchangeable in Webber's texts, the fact may indicate a genre difference. However, it is important to note that ordinarily the substitution of *this* for *that,* and vice-versa, does not make the individual sentence ungrammatical, or vaguely inappropriate, in isolation. We discovered the correlation with global topic not by examining isolated instances, but by collecting a large sample in discourse context.[20] Webber proposes an account of deictic reference in which the referent is determined, and only afterwards is the proximal/distal distinction relevant, when the hearer uses it to characterize the speaker's 'psychological distance' from the referent as either close or far away. We suggest that the consistency of the proximal/distal correlation with topicality probably reflects a linguistically significant, though unconscious, choice by the writer. Proximals imply topicality, and distals imply peripherality. The reader utilizes this information in identifying the referent, rather than as an afterthought to the anaphora resolution process.

9. Conclusions

The patterns of anaphora described here are consistent with a psychological model in which pronoun reference is limited to items in short-term memory.

Items accessible to anaphora in general have antecedents inside the discourse segment, consistent with a view of the discourse segment as focus of operating memory. Anaphors with antecedents outside the segment are explicit (full noun phrases rich in lexical content) and/or have topical referents which are maintained in memory because of their importance to the reader and writer. In the news commentary genre, among others, global topics are salient and accessible for anaphoric reference.

The fact that a few pioneering studies of constraints on anaphora focused on individual pronouns in specific genres has tended to foster the expectation that all anaphoric phenomena will be found to be subject to unitary constraints in terms of factors such as recency and focus space. By examining a different genre, by broadening the range of anaphor forms studied, and by distinguishing abstract from individual object reference forms, we have found that these constraints are not independent. Recency is a function of discourse referent type, anaphor form and genre. The existence of two main discourse referent types (individual and abstract, with several subtypes) and three main anaphor forms (pronoun, demonstrative and full noun phrase, including gradations along the explicitness continuum), in combination with a number of different genres, leads to a profusion of possible values for a recency constraint. Similarly, anaphor heaviness is dependent upon whether the genre is segmented, whether the genre has topics, and whether the discourse referent is an event. Thus the choice of heaviness of anaphor is a function of genre, segmentation, topicality and discourse referent type. Finally, we found that segmentation is genre-relative, so a segment-based constraint is a function of genre as well. Our findings suggest that a set of independent competing constraints[21] is not a promising basis for a theory of anaphor resolution.

Appendix A: Global Topic

For the purpose of this study, we have identified global topics for the 22 articles as follows:

1. Bally agrees to buy back Trump's stock. 2. Brazil suspends interest payments on debt to banks. 3. Finance ministers of US and allies meet and issue statement. 4. The government is investigating insider trading but has not charged Kidder Peabody. 5. Levine is sentenced to prison and fined for insider trading. 6. Microsoft announces a new operating system. 7. Braniff CEO McGee announced aircraft acquisition program. 8. Judge decides case of two students. 9. SEC negotiates with Drexel Burnham Lambert to make changes. 10. Korean factory managers don't like American imports. 11. US jets shoot down Libyan

fighter planes. 12. Reagan blames budget deficit on "iron triangle." 13. SBA program's toughened rules upset minority firms. 14. Writers strike TV studios. 15. Thai civil aviation officials received permission from the cabinet to renegotiate Thailand's air-service agreement with the US 16. The head of the EPA threatens to block building of the Two Forks dam project. 17. James Garner settles suit against Universal City Studios. 18. A study of office lighting was commissioned by Honeywell Inc. 19. Federal officials cited a unit of Lockheed Corp. for health and safety violations. 20. Nestle to launch iced-coffee drink. 21. An audit of the Palo Verde Nuclear Generating Station made recommendations. 22. Shoe manufacturers introduce walking pumps.

Appendix B: Discourse Segmentation

Seven different coherence relations were found, reflecting the articles' content and organization. These coherence relations, each with the total number of those segments identified in the corpus, are as follows:

Evaluation 80
Background 34
Elaboration 22
Topic 20
Cause 5
Constituency 3
Contrast 1

Appendix C: Sample Article

(Wall Street Journal, 3/27/89)
JAMES GARNER SETTLES SUIT AGAINST MCA'S UNIVERSAL
LOS ANGELES

On the brink of a trial that had been expected to bring to light long-debated movie studio accounting practices, actor James Garner agreed to settle a long-pending $16.5 million lawsuit against Universal City Studios, a unit of MCA Inc. Terms weren't disclosed.

When Mr. Garner was the star of the hit 1970's NBC series "The Rockford Files", he negotiated successfully for 37.5% of the net profits from network

airings, reruns and foreign sales of the show, owned by Universal. In 1983 he still hadn't received a penny, so he sued, alleging breach of contract and fraud.

The trial, which was scheduled to begin today in state court here, had been seen as an opportunity to focus on "creative accounting" methods in Hollywood. Actors, directors and producers with clout now seek a percentage of gross revenues instead of net profits because of experiences such as Mr. Garner's.

Mr. Garner had alleged that "The Rockford Files" had grossed nearly $125 million since it began its five-year network run in 1974. He finally received an initial check last December, then another earlier this month, totaling $607,000.

In this article, four discourse segments were identified as follows: Topic, paragraph 1; Background, paragraph 2; Evaluation, paragraph 3; Background, paragraph 4.

Notes

* We wish to thank Joyce McDowell and Sandra Thompson for helpful comments on this study.

1. Discourse segmentation criteria and coherence relations were developed by the second author; data collection and analysis were by the first author.
2. An anaphor can be said to "point" to its antecedent; however, anaphoric reference is to meanings and not to the forms that have gone before (Halliday and Hasan 1976, Sidner 1983). Words do not refer back to other words; people use words to refer to concepts. As observed by Sidner, a noun phrase can be said to specify a cognitive element in the hearer's mind, and a subsequent phrase can be said to co-specify that memory element; this distinction is particularly relevant for *one*-anaphora (not treated in this paper).
3. Six articles are from the February 23, 1987 issue, eight are from the January 5, 1989 issue, and eight are from the March 27, 1989 issue. The articles range in length from 200 to 2,150 words.
4. Our focus was on noun phrases for which there were explicit previous mentions of the referent in the text, rather than the set of noun phrases marked with the definite article. We omitted definite noun phrases which did not have explicit previously-mentioned referents: those requiring bridging from previous knowledge, including associated parts, inferred roles, epithets, and set members (Clark 1977), cataphorics, frame-defined objects (DuBois 1980, Givon 1995), exophoric referents, including homophoric uniques and generics (Halliday and Hasan 1976), Hawkins's (1978) immediate situation use and wider situation use, and "mode" reference (Lockman and Klappholz 1980). We did not include *one*-anaphors, zero-anaphors, ellipsis, or anaphoric expressions inside direct quotations (Payne 1992 also treats differently those participant mentions within quotes, since their continuity pertains more to the quoted discourse than to the discourse in which the quote appears). A text can contain several tokens of a proper name (for example, *Reagan*), and separate tokens can refer to separate individuals (Ronald or Maureen, for example), so that matching the linguistic expression with the

appropriate referent can be non-trivial; however, the corpus contained no instances of this sort, and proper names were not identified as anaphors.

5. Statistical significance calculations for the findings in this paper used Fisher's Exact Test, an exact version of the chi-square approximation. For Table 6 below, differences were compared pairwise using an exact test for ordered categories, the Cyrus Mehta Algorithm, an extension of Fisher's Exact Test. For statistical measures, we appreciate the advice and assistance of Margaret Frane.

6. Data from a study of oral conversation show a related pattern. Clausal structures have abstract referents, and the presence of clausal structure within a noun phrase antecedent favors the choice of *that* over *it* as anaphor (Schiffman (Passonneau) 1984).

7. Webber 1988 suggests that non-deictic anaphors (pronouns and full noun phrases) may refer to an associated entity (bridging inference), while deictic anaphors (what we call demonstratives here) may point to something already explicitly included in the discourse.

8. For 28 of the individual full noun phrase anaphors, the most recent antecedent was a synonym, name, pronoun, or descriptive term; prior to this antecedent was a copy of the anaphor. We classified these anaphors as copies, and they are included in the total of 171 individual full noun phrase anaphors.

9. Halliday and Hasan 1976:279 recognize a continuum or cline of cohesive elements, from specific to general, as follows: *copy > synonym > superordinate > general word > pronoun.* Cumming and Ono (1996) suggest that a pre-established taxonomic hierarchy (in the human mental lexicon model or the computer software) will not be adequate for identification of all superordinates, because people establish new supercategories "on the spot," in mid-discourse. If so, a truly intelligent natural language understanding system would need to provide for local amplification of a pre-existing concept hierarchy in the course of processing a text.

10. In computing average distance to antecedent, all anaphors with antecedent 5 or more sentences earlier were counted as having a distance of 5, so that the results would not be greatly affected by the dozen or so anaphors at very long distances. For anaphors with multi-sentential antecedents, the sentence which preceded the anaphor was a part of the antecedent, so they were all assigned a distance of 1. Accordingly, the averages for demonstratives and full noun phrases are slightly conservative, but the overall distinctions among anaphor forms remain.

11. One exception was the pronoun *it* in the idiomatic construction *...there is more to it than the learning curve....* In another exception an intervening sentence between the pronoun and the antecedent consisted of quoted material attributed to the individual:

> *"We have started to make some changes, but these are not fundamental changes," he said. "Lockheed has long had a policy of providing a safe and a healthful workplace for all of its employees." More changes may emerge from further analysis of OSHA's complaints, he said.*

In this case the *he* in the third sentence is coreferent with the *he* in the first sentence. For cases like these, a sentence boundary within a direct quotation is probably best treated as primarily a formal feature of the quotation unit, and only secondarily as a feature of the exposition; conventions for written English allow the insertion of the *he said* between the sentences or at the end of the quotation. In fact, it can be argued that the phrase *he said* here functions primarily as a parenthetical for purposes of attribution of source, and only marginally as the main subject and predicate. For passages with quotations, the generalization of at most one sentence boundary between pronoun and antecedent can be maintained if sentence boundaries within quotations are not counted.

12. Hobbs 1976 found all antecedents in the current sentence or in the previous sentence in a corpus of 200 pronouns from an archaeology book and a newsmagazine; a corpus from a novel in colloquial style showed 6% at a distance of two or more sentences, suggesting a genre-based difference. His sample differs from ours in his inclusion of intra-S antecedents. In a "continuity model," the entities mentioned in a discourse are laid down in memory like beads on a string; in a "discontinuity model," the entities mentioned in the current sentence and one sentence back have a privileged place in working memory, and are readily available to be referred to by nouns and pronouns. As our study shows, a discontinuity model is applicable for pronouns but not full noun phrases.

13. The additional corpus contained 6000 lines of *Wall Street Journal* text from April 20, 1989.

14. The demonstrative pronoun typically occurred with a verb that was evaluative, as in

> *That means.... That indicates.... That isn't necessarily alarming.... That could cause.... That seems easy enough....*

or explanatory, as in

> *That's because.... That depends on.... That occurs when....*

When the antecedent was not clausal, selectional restrictions on the verb helped to identify a likely non-clausal referent for the pronoun. Reichman-Adar 1984 found parallel data in spoken technical exchanges, where *that* often occurs in "why" or "how" sentences. (She also found *that* marking the end of a discourse section, which we did not; the difference could be genre-related, or could reflect differences in both genre and the discourse structure model used, since her context spaces tend to be smaller chunks than our discourse segments.)

15. The use of full noun phrases instead of pronouns at the beginning of a segment has been found in a range of genres (Hinds 1978, Linde 1979, van Dijk and Kintsch 1983, Reichman 1985, Grosz and Sidner 1986, Guindon *et al.* 1986, Chafe 1987, Fox 1987a, Fox 1987b, Tomlin 1987, Hofmann 1989).

16. Data from stories in Yagua, a language of Peru, suggest that topical participants are more accessible for long-distance anaphora (Payne 1992). When the antecedent was at a distance of more than 8 clauses, the anaphor was typically a noun phrase; however, in some cases the anaphor was a weak form (e.g., a prefix), but for these the referent was a central or major character. Discourse structure was a contributing factor: in one story, the distance between a weak anaphor and its antecedent was 65 clauses, but the intervening clauses constituted an embedded episode.

    Experimental studies (Morrow 1985) show that, in determining pronoun reference in narratives, subjects tended to choose protagonists and were most confident about their decisions when the protagonist was a participant in a foregrounded, rather than backgrounded, event (Hopper and Thompson 1980).

17. On a hierarchy of givenness proposed by Gundel *et al.* 1989, data based on a corpus including a variety of genres showed distal demonstratives ranking slightly lower than proximals in the mid-ranges of a scale with the values: *in focus > activated > familiar > uniquely identifiable > type identifiable.* Referents of proximal demonstratives were more likely to have "activated" status.

18. Although the functions of demonstratives in languages other than English may differ, it is interesting to note that Kirsner and van Heuven 1988 found similar patterns in comparing

proximal and distal demonstrative adjectives in a large written corpus of Modern Dutch. Proximals were used for major referents, relatively important entities retained as topics; distals were used for minor referents, and had shorter distances to their antecedents. Also, proximal demonstratives are associated with topics in English; new noun phrases marked with *this* are more likely to become topics than are those marked with *a/an* (Wright and Givon 1987, Gernsbacher and Shroyer 1989). In contrast, in an experimental study using English mini-narratives with *it* and *that* (Shuster 1988), subjects tended to identify the referent of *that* as the most recent non-topical event.

19. According to Sidner 1983, *this* indicates the main concern and *that* indicates secondary concern; the options allow the speaker to point at the relevant material with the least confusion. Sidner considers demonstrative anaphora in terms of their function in moving the focus of attention from one entity to another. The use of *this* moves the focus to its referent. The use of *that* allows the speaker to re-mention discourse elements without them becoming the focus of the speaker's (and therefore the hearer's) attention.

20. As noted by Reichman-Adar 1984:368, surface linguistic forms are bound to subconscious discourse processing mechanisms, and utterances in violation of discourse reference rules are not visibly ungrammatical. "For example, in many of the cases where a pronominal form is prohibited by discourse rules, a speaker's use of a pronominal in any case, probably would not cause listeners undue difficulty in correctly retrieving the referent intended. On the other hand, their implicit understanding of the discourse structure, which they need in order to adequately model the discourse flow, would be confused."

21. The adequacy of a given computational approach is an empirical matter (see Walker 1989). A centering approach for pronouns, which computes local focus, has achieved some success (Brennan, Friedman and Pollard 1987, Kameyama 1986). Another algorithm for pronouns has also achieved impressive results using recency, surface order and depth of embedding (Hobbs 1976). How these approaches might incorporate the full range of anaphor forms and genres, in light of the complexities described here, remains to be seen. A strategy that identifies possible referents and assigns them relative probability weights based on a variety of relevant factors (*e.g.*, Asher and Wada 1988, Rich and LuperFoy 1988) could utilize different sets of weights with respect to anaphor form, referent type, and topicality--taking genre into account, as human text comprehenders do. The findings of this study have been used in building a text understanding system that matches a broad range of anaphors to referents (Wada 1994).

References

Ariel, Mira. 1990. *Accessing Noun-phrase Antecedents.* London: Routledge.

Asher, Nicholas. 1987. "A typology for attitude verbs and their anaphoric properties." *Linguistics and Philosophy* 10: 125–197.

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse.* Dordrecht: Kluwer.

Asher, Nicholas, and Hajemi Wada.1988. "A computational account of syntactic, semantic and discourse principles for anaphor resolution." *Journal of Semantics* 6.

Bentivoglio, P. 19983. "Topic continuity and discontinuity in discourse: A study of Spoken Latin-American Spanish." In *Topic Continuity in Discourse,* T. Gívon (ed.), 255–312. [Typological Studies in Language, 3]. Amsterdam: John Benjamins.

Blundell, William E. 1986. *The Art and Craft of Feature Writing, Based on the Wall Street Journal Guide.* New York: New American Library.

Bolinger, Dwight. 1979. "Pronouns in discourse." In *Syntax and Semantics*, vol. 12, *Discourse and Syntax*, T. Givón (ed.), 289–309. New York: Academic Press.

Bosch, Peter. 1983. *Agreement and Anaphora: A study of the role of pronouns in syntax and discourse.* New York: Academic Press.

Brennan, Susan E., Marilyn Walker Friedman, and Carl J. Pollard. 1987. "A centering approach to pronouns." *Proceedings of the 25th Annual Meeting of the ACL, Stanford University*, 155–162.

Brown, Cheryl. 1983. "Topic continuity in written English narrative." In *Topic continuity in discourse*, T. Givón (ed.), 313–341. Amsterdam: John Benjamins.

Chafe, Wallace L. 1974. "Language and consciousness." *Language* 50: 111–133.

Chafe, Wallace L. 1976. "Givenness, contrastiveness, definiteness, subjects, topics, and point of view." In *Subject and Topic*, Charles Li (ed.). New York: Academic Press.

Chafe, Wallace L. 1987. "Cognitive constraints on information flow." In *Coherence and grounding in discourse*, Russell Tomlin (ed.). Amsterdam: John Benjamins.

Clancy, Patricia M. 1980. "Referential choice in English and Japanese narrative discourse." In *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*, Wallace L. Chafe (ed.). 127-202. Norwood, New Jersey: Ablex.

Clark, H. 1977. "Bridging." In *Thinking: Readings in cognitive science*, P. Johnson-Laird and P. Wason (eds). Cambridge: Cambridge University Press.

Cumming, Susanna, and Tsuyoshi Ono. 1996. "*Ad hoc* hierarchy: Lexical structures for reference in *Consumer Reports* articles." *Studies in anaphora*, Barbara Fox (ed.), 69–94. [Typological Studies in Languages, 33]. Amsterdam: John Benjamins.

Dahlgren, Kathleen. 1988. *Naive semantics for natural language understanding.* Dordrecht: Kluwer.

Dahlgren, Kathleen. 1996. "Discourse coherence and segmentation." In *Burning issues in coherence*, E. Hovy and D. R. Scott (eds). Berlin: Springer.

DuBois, John W. 1980. "Beyond definiteness: The trace of identity in discourse." *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*, Wallace L. Chafe (ed.). Norwood, New Jersey: Ablex.

Fox, Barbara A. 1987a. *Discourse Structure and Anaphora: Written and conversational English.* Cambridge: Cambridge University Press.

Fox, Barbara A. 1987b. "Anaphora in popular written English narratives." In *Coherence and Grounding in Discourse*, R.S. Tomlin (ed.), 157–174. Amsterdam: John Benjamins.

Gernsbacher, M. A. and S. Shroyer. 1989. "The cataphoric use of the indefinite *this* in spoken narratives." *Memory and Cognition* 17: 536–540.

Givón, Talmy. (ed.), 1983. *Topic Continuity in Discourse: A quantitative cross-language study.* Amsterdam: John Benjamins.

Givón, Talmy. 1989. *Mind, Code, and Context: Essays in pragmatics.* Oakdale: Lawrence Erlbaum.

Givón, Talmy. 1995. "Coherence in text vs. coherence in mind." *Coherence in Spontaneous Text*, Morton Ann Gernsbacher and Talmy Givón (eds), 59–115. Amsterdam: John Benjamins.

Grimes, J.E. (ed.). 1978. *Papers on Discourse.* Dallas: Summer Institute of Linguistics, Inc.

Grosz, Barbara J. 1977. "The representation and use of focus in a system for understanding dialog." In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence Cambridge, MA*, 67–76. Los Altos, California: William Kaufmann.

Grosz, Barbara J., A. Joshi, and S. Weinstein. 1983. "Providing a unified account of definite noun phrases in discourse." In *Proceedings of the International Joint Conference on Artificial Intelligence, Vancouver, B.C.*

Grosz, Barbara J. and C. Sidner. 1986. "Attention, intentions, and the structure of discourse." *Computational Linguistics* 12: 175–204.

Guindon, Raymonde. 1985. "Anaphora resolution: Short-term memory and focusing." In *Proceedings of 23rd Annual Meeting, Association for Computational Linguistics*, 218–227.

Guindon, Raymonde, Paul Sladky, Hans Brunner, and Joyce Conner. 1986. "The structure of user-adviser dialogues: Is there method in their madness?" *AACL Proceedings*, 224–230.

Gundel, Jeanette K.; Nancy A. Hedberg; and Ron Zacharski. 1989. "Givenness, implicature and demonstrative expressions in English discourse." *Papers from the Parasession on Language in Context at the Twenty-fifth Regional Meeting of the Chicago Linguistic Society*, Randolph Graczyk, Bradley Music, and Caroline Wiltshire (eds). Chicago: Chicago Linguistic Society.

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

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

Hinds, John. 1977. "Paragraph structure and pronominalization." *Papers in Linguistics* 10: 77–99.

Hinds, John. 1978. "Anaphora in Japanese conversation." In *Anaphora in Discourse*, John Hinds (ed.), 136–179. Edmonton, Canada: Linguistic Research.

Hirst, Graeme. 1981a. *Anaphora in Natural Language Understanding: A survey.* Berlin: Springer. [*Lecture Notes in Computer Science* 119].

Hirst, Graeme. 1981b. "Discourse-oriented anaphora resolution: A review." *ACL* 7: 85–98.

Hobbs, Jerry R. 1976. *Pronoun Resolution.* Technical Report 76–1, Department of Computer Science, City College, City University of New York.

Hobbs, Jerry R. 1985. *On the Coherence and Structure of Discourse.* CSLI Report #CSLI-85–37.

Hofmann, Thomas R. 1989. "Paragraphs and anaphora." *Journal of Pragmatics* 13: 239–250.

Hopper, Paul J. and Sandra A. Thompson. 1980. "Transitivity in grammar and discourse." *Language* 56: 251–299.

Kameyama, Megumi. 1986. "A property-sharing constraint in centering." In *Proceedings, 24th Annual Meeting, Association for Computational Linguistics*, 200–206.

Kirsner, Robert S. 1979. "Deixis in discourse: An exploratory quantitative study of the Modern Dutch demonstrative adjectives." *Syntax and Semantics*, vol. 12, *Discourse and Syntax*, T. Givón (ed.), 355–375. New York: Academic Press.

Kirsner, Robert S. and Vincent J. van Heuven. 1988. "The Significance of demonstrative position on Modern Dutch." *Lingua* 76: 209–248.

Levinsohn, Stephen H. 1978. "Participant reference in Inga narrative discourse." *Anaphora in Discourse*, J. Hinds (ed.), 69–135. Edmonton, Canada: Linguistic Research.

Li, Charles N. and Sandra A. Thompson. 1979. "Pronouns and zero-anaphora in Chinese discourse." *Syntax and Semantics*, vol. 12, *Discourse and Syntax*, T. Givón (ed.). New York: Academic Press.

Linde, C. 1979. "Focus of attention and the choice of pronouns in discourse." *Syntax and Semantics*, vol. 12, *Discourse and Syntax*, T. Givón (ed.), 337–354. New York: Academic Press.

Lockman, Abe, and A. David Klappholz. 1980. "Toward a procedural model of contextual reference resolution." *Discourse Processes* 3: 25–71.

Lyons, John. 1977. *Semantics*, vol. 2. Cambridge: Cambridge University Press.

Mann, William C. 1988. "Rhetorical structure theory: Towards a functional theory of text organization." *Text* 8: 243–281.

Mann, William C., and Sandra A. Thompson. 1987. "Rhetorical structure theory: A framework for the analysis of texts." *Papers in Pragmatics* 1.1: 79–105. Also available as Information Sciences Institute Research Report 87–185, 4676 Admiralty Way, Marina del Rey, CA 90292–6695.

Morrow, Daniel G. 1985. "Prominent characters and events organize narrative understanding." *Journal of Memory and Language* 24: 304–319.

Passonneau, Rebecca J. 1989. "Getting at discourse referents." *Proceedings, 27th Annual Meeting, Association for Computational Linguistics*, 51–59.

Payne, Thomas E. 1992. *The Twins Stories: Participant coding in Yagua narrative.* Berkeley: University of California Press.

Pu, Ming-Ming. 1995. "Anaphoric patterning in English and Mandarin narrative production." *Discourse Processes* 19: 279–300.

Reichmann, Rachel. 1978. Conversational coherency. *Cognitive Science* 2: 283–327.

Reichman-Adar, Rachel. 1984. "Technical discourse: The present progressive tense, the deictic 'that,' and pronominalization." *Discourse Processes* 7: 337–369.

Reichmann, Rachel. 1985. *Getting Computers to Talk Like You and Me.* Cambridge, Mass.: MIT Press.

Reinhart, Tanya. 1983. *Anaphora and Semantic Interpretation.* Chicago: The University of Chicago Press.

Rich, Elaine and Susann LuperFoy. 1988. "An architecture for anaphora resolution." In *Proceedings of the Second Conference on Applied Natural Language Processing, Association for Computational Linguistics.*

Schiffman, R.J. (Passonneau). 1984. "The two nominal anaphors *it* and *that*." *CLS* 20: 322–357.

Schuster, Ethel. 1988. *Pronominal Reference to Events and Actions: Evidence from naturally-occurring data.* Philadelphia: Department of Computer and Information Science, School of Engineering and Applied Science, University of Pennsylvania.

Sidner, C. 1983. "Focusing in the Comprehension of Definite Anaphora." *Computational Models of Discourse*, M. Brady and R. Berwick (eds), 267–330. Cambridge, Mass.: MIT Press.

Tomlin, Russell S. 1987. "Linguistic reflections of cognitive events." In *Coherence and Grounding in Discourse*, R.S. Tomlin (ed.), 455–479. Amsterdam: John Benjamins.

Van Dijk, Teun A. 1977. "Sentence topic and discourse topic." *Journal of Slavic Philology* 1: 49–61 (reprinted in T. A. van Dijk, 1981, Studies in the pragmatics of discourse, The Hague: Mouton).

Van Dijk, Teun A. and Walter Kintsch. 1983. *Strategies for discourse comprehension.* New York: Academic Press.

Wada, H. 1994. "A treatment of functional definite descriptions." *Proceedings of the COLING, Tokyo, Japan.*

Walker, Marilyn A. 1989. "Evaluating discourse processing algorithms." *Proceedings of the 17th Annual Meeting, Association for Computational Linguistics*, 251–261.

Webber, Bonnie Lynn. 1988. "Discourse deixis: Reference to discourse segments." *Proceedings of the 26th Annual Meeting, Association for Computational Linguistics*, 113–122.

Wright, S. and T. Givón. 1987. "The pragmatics of indefinite reference: Quantified text-based studies." *Studies in Language* 11: 1–33.