Generating Anaphoric Expressions: Pronoun or Definite Description?

Kathleen F. McCoy
Dept. of Computer and Information Sciences
University of Delaware
103 Smith Hall
Newark, DE 19716, USA
mccoy@cis.udel.edu

Michael Strube
Institute for Research in Cognitive Science
University of Pennsylvania
3401 Walnut Street, Suite 400A
Philadelphia, PA 19104, USA
strube@linc.cis.upenn.edu

Abstract

In order to produce coherent text, natural language generation systems must have the ability to generate pronouns in the appropriate places. In the past, pronoun usage was primarily investigated with respect to the accessibility of referents. We argue that generating appropriate referring expressions requires looking at factors beyond accessibility. Also important are sentence boundaries, distance from last mention, discourse structure and ambiguity. We present an algorithm for generating appropriate anaphoric expressions which takes the temporal structure of texts and knowledge about ambiguous contexts into account. We back up our hypotheses with some empirical results indicating that our algorithm chooses the right referring expression in 85% of the cases.

1 Introduction

Anaphoric expressions are an important component to generating coherent discourses. While there has been some work on generating appropriate referring expressions, little attention has been given to the problem of when a pronoun should be used to refer to an object. In most instances the assumption has been that a pronoun should be generated when referring to a discourse entity that is highly prominent (accessible). However, a study of naturally occurring texts reveals that factors beyond accessibility must be taken into account in order to explain the patterns of pronoun use found.

Other researchers have indicated that fuller descriptions tend to be found at the beginning of discourse segments (Grosz & Sidner, 1986; Reichman, 1985), even when the object being referred to is extremely prominent in the preceding sentence. There may be two reasons for this: (1) it could be that the item is not accessible since the "focus space" associated with the previous sentence is "popped" at the boundary (and perhaps an older focus space, with respect to which accessibility should be judged, is restored in its place (Passonneau, 1996)), (2) the use of a fuller definite description is actually marking the discourse segment boundary. If (2) is the case, because other methods for marking boundaries are also possible (and the writer need not use multiple markings), this may help explain why the correlation between discourse segment boundaries and fuller referring expressions is not perfect.

Supposing that discourse segments are an important factor in choosing anaphoric expressions, in order to take advantage of them we must have a clear definition of what a discourse segment is. For generation, the discourse segment boundary must be part of the input to a sentence generator (which, we assume, is responsible for making the referring expression choice). To evaluate proposals for generating referring expressions, the discourse segment boundary must also be recognizable.

While discourse segment boundaries may be important, note that fuller noun phrases are sometimes used when there is no discourse segment boundary (under any reasonable definition of boundary). This may occur when the referent is not accessible because the nearest antecedent is too far away or because it is confusable with another referent. We attempt to explain such instances as well.

In order to determine the circumstances under which to choose a pronoun versus a definite description[1], our tack has been to study naturally occurring examples and to try to hypothesize rules that explain the reference forms in those examples. To date we have concentrated our study on New York Times news articles. We hope to generalize some of our findings to other types of text genres as well.

Consider the following passage from the first several lines of one of the stories we analyzed:

> **Example 1:**
> When **Kenneth L. Curtis** was wheeled into court nine years ago, mute, dull-eyed and crippled, it seemed clear to nearly everyone involved that it would be pointless to put **him** on trial for the murder of **his** former girlfriend,

[1] We use the term definite description to mean either a definite noun phrase or a name.

Donna Kalson, and the wounding of her companion.

It had been a year since **Mr. Curtis** had slammed **his** pickup truck into them, breaking their legs. **He** then shot them both and, finally, fired a bullet into **his** own brain. **Mr. Curtis** lingered in a coma for months, then awoke to a world of paralysis, pain and mental confusion from which psychiatric experts said **he** would never emerge.

One expert calculated **his** I.Q. at 62.....

For convenience, we have indicated all references to the main character in bold.

A surprising thing to note about this passage is that not all of the anaphoric references to *Mr. Curtis* are pronouns even though he is arguably the focus of every sentence included. Some previous work on pronoun generation would predict that a pronoun should be used if the same item remains in focus. Thus it appears that something other than a straightforward application of focusing or other pronoun resolution algorithms is necessary.

A second thing to note about this passage is that the sentences are generally long and complex and often contain several references to the same character. These types of sentences are very different from those that have been considered by any generation system that has rules for generating pronouns. In addition, it is not clear how most focusing or pronoun interpretation algorithms would handle them.

One hypothesis that might be made is that the underlying structure of the text might affect pronoun generation. Care must be taken in choosing such a structure. The chosen structure must (1) explain the patterns of pronouns found in naturally occurring text, and (2) be based on information available to a sentence generation system. Notice that constructs such as paragraph breaks meet neither of these criteria.

In this work we hypothesize that discourse structure (segmentation) is indeed vital in the decision of whether or not to generate a pronoun. However, we argue that a single definition of discourse segment is not sufficient to explain the patterns of pronoun use found, and seek a more general notion. Here, to distinguish our notion from other notions of discourse segmentation found in the literature (e.g., Reichman (1985) or Grosz & Sidner (1986)), we use the term *discourse thread* to capture the structuring notion to which we refer. We propose that a discourse generally contains multiple threads which run through the discourse and can serve to structure the discourse. In general, a single thread is evident at a particular point in the discourse, but this thread may be replaced by another thread and then picked up again at another point in the discourse (cf. Rosé et al. (1995)). The "threading device" used to structure will be different for different kinds of discourses. For instance, in the kinds of discourses studied in Grosz & Sidner (1986) the threading device may be the intentional structures (and each of their discourse segments would constitute a thread of the discourse). In the discourses that we studied (New York Times articles), threads defined in terms of the time referenced in a clause appeared to be quite prominent.

In this paper we present our preliminary work in uncovering factors that affect pronoun generation decisions. Our work so far has led us to hypothesize several factors including:

Sentence Boundaries – pronouns appear to be the preferred referring form for subsequent reference to an item within the same sentence.

Distance from Last Mention – when the last mention of an item is several sentences back in the text, a definite description is preferred.

Discourse Structure in terms of multiple threads – when the previous reference to an item is in the same thread as the current reference, a pronoun is preferred. A definite description is preferred when the threads are different.

Ambiguity – potential ambiguity must be taken into account when choosing an anaphoric expression in that a pronoun should only be generated if it can be resolved correctly.

In the next section we discuss previous research on pronoun generation. This is followed by a discussion of some anaphoric expressions that do not require considering discourse structure to generation. This is followed by an introduction of time as a structuring device which affects pronoun generation. Next we investigate ambiguous anaphoric references. We follow this with an algorithm which decides when to use a pronoun versus a definite description when referring to some discourse entity. After this, we report on empirical results of the application of our algorithm to a corpus of New York Times articles. Finally, we provide some related work, future research, and conclusions.

2 Previous Work on Pronoun Generation

Few researchers have given serious consideration to the problem of pronoun generation. The most common factor considered has been the accessibility of the referent. If the referent is sufficiently prominent in the preceding text, a pronoun is used. Some early generation work (e.g., McDonald (1980), McKeown (1983), McKeown (1985), Appelt (1981)) used a simple rule to implement this idea based on focus (Sidner, 1979) that roughly stated that if the current sentence is about the same thing that the previous sentence was about, use a pronoun to refer to that thing. As was pointed out above, this rule does not provide a very good match with the referring expressions in our corpus.

Dale (1992) also discussed the generation of pronouns in the context of work on generating referring expressions (Appelt, 1985; Reiter, 1990). Dale specified an algorithm that essentially generated the smallest referring expression that distinguished the object in question from all others in the context.[2] He generated a pronoun (or ellipsis) if one were adequate and if the object being referred to was the center of the last utterance (where the notion of center was defined in a domain dependent fashion). As an example of the kinds of texts he generated consider: "Soak the butterbeans. Drain and rinse *them*." Such an account of pronoun generation, based on center constancy, appears to work quite well in the domain Dale considered. However, as pointed out in Example 1, it does not seem to explain the patterns found in the texts we analyzed.

The centering model (Grosz et al., 1995) itself makes predictions about pronoun generation only in a specific instance – that where Rule 1 is applicable. Centering's Rule 1 states that if any element of the previous utterance's forward looking center list is realized in the current utterance as a pronoun, then the backward looking center must be realized as a pronoun as well (Grosz et al., 1995, p.214). Notice that the *Mr. Curtis* at the beginning of the second sentence in Example 1 is an apparent violation of this rule. But, more generally, we must have a theory that is able to handle all cases of pronoun use.

A pronoun interpretation algorithm based on centering which relied on centering transition preferences was developed in Brennan et al. (1987). Using transition preferences in a pronoun generation rule would cover more cases of pronoun use than is covered by Rule 1, but the application of such transition preferences also proved unhelpful in explaining pronoun patterns in our corpus.

Reichman (1985) and Grosz & Sidner (1986) indicate that discourse segmentation has an effect on the linguistic realization of referring expressions. While this is intuitively appealing, it is unclear how to apply this to the generation problem (in part because it is unclear how to define discourse segments to a generation system). Passonneau (1996) argues for the use of the principles of information adequacy and economy. Her algorithm takes discourse segmentation into account through the use of focus spaces which are associated with discourse segments. Thus, Passonneau explains that a fuller description might be used at a boundary because the set of accessible objects changes at discourse segment boundaries (though she combines this consideration with centering theory which may override the decisions due to segment boundaries). While Passonneau's algorithm seems quite appealing, notice that it provides no explanation of how a discourse segment should be defined. The evaluation that she provided used the discourse segments provided by a set of naive subjects who indicated discourse segment boundaries in her texts. Without such boundaries provided, it is impossible to apply her algorithm. In some sense, the work presented here is consistent with Passonneau's theory. What we attempt to add is a genre-dependent definition of discourse segment (thread) which is well-defined and can be derived from input that any sentence generator must have in order to generate a sentence. On the other hand, we differ from Passonneau in that we do not attempt to make use of focus spaces in generation. Rather we argue for evaluating informational adequacy on the basis of confusable objects near the current sentence in a discourse.

In the following section we hypothesize that discourse structure in terms of multiple threads does have an effect on appropriate anaphoric expression choice and that the temporal structure of a discourse is a particular instantiation of these threads (in the stories that we have analyzed). We hypothesize that if there is a difference in time between the last reference to an entity and the current reference, a definite description is used (and when the time of the previous and the current reference is the same, a pronoun is used).

3 Long- and Short-Distance Anaphoric Expressions

In studying naturally occurring texts in order to identify patterns of pronoun usage, we found consistent patterns over both long distances (i.e., where the last reference to the entity being referred to was more than two sentences back in the discourse) and over short distances (i.e., where the last reference to the entity being referred to was in the current sentence).

In long-distance situations we have found that a definite description is almost always used. In short-distance situations a pronoun is almost always used (except in situations where this pronoun is ambiguous). Thus, the sentence seems to be a very important construct to consider in choosing anaphoric expressions.

We now turn to consider the factors that might be affecting the expression choice in situations other than these. We hypothesize that discourse structure must be considered for such cases.

4 Time-Threaded Discourse Structure

In this section, we describe our approach to using discourse structure for choosing the right referring expression. Since we are working with stories from newspapers we were not able to identify the kind of discourse structure as assumed by Grosz & Sidner (1986), whose dialogues are more task-oriented and have clear intentional goals. Instead, the texts have a structure consisting of multiple story lines which we call threads (cf. also Rosé et al. (1995)). A thread describes a particular part of the story. It can be interrupted by other threads and continued later. Thus the threaded structure is more complicated than the hierarchical (tree-like) structure posited in Grosz & Sidner (1986) and Mann & Thompson (1988). Therefore, we do not think that the texts we looked at can be analyzed using a stack (for example).

We needed to find a structuring device that was part of the input to a sentence generation system and that was recognizable on the surface (so that we could evaluate our algorithm on naturally occurring text). After investi-

[2] This algorithm was later revised in Dale & Reiter (1995) to more adequately reflect human-generated referring expressions and to be more computationally tractable.

gating some work on narrative structure (Genette, 1980; Prince, 1982; Vogt, 1990), we determined that changes in the deictic center of the story (Nakhimovsky, 1988; Wiebe, 1994) not only must be part of the input to a sentence generator (e.g., for appropriate tense generation), but were also both well marked in the text and seemed to have an influence on anaphoric expression choice.

A shift in the deictic center can be signaled by a shift of topic, a shift of time scale, a shift in spatial scale, or a shift in perspective (Nakhimovsky, 1988; Wiebe, 1994). Since a shift in time scale is often indicated by linguistic means and the time being referred to must be part of the input to a sentence generator, we concentrated on this point. We also acknowledge that the changes in time seem quite important in the news stories that we analyzed. Other genres of text might depend on other kinds of structuring devices.

Changes in time scale or time, as we redefined the category, may require world knowledge reasoning to recognize but are often indicated by either cue words and phrases (e.g., *"nine years ago", "a year", "for months", "several months ago"*), a change in grammatical time of the verb (e.g., past tense versus present tense), or changes in aspect (e.g., atomic versus extended events versus states as defined by Moens & Steedman (1988)).

In considering how time change might affect anaphoric expression choice, we consider the choice for the first mention of a discourse entity in a sentence where that entity has recently been referred to in the discourse. Our hypothesis is that: Changes in time reliably signal changes of the thread in newspaper articles; definite descriptions should appear when the current reference to a discourse entity is in a different thread from the last reference to that entity and pronouns should occur when the previous mention is in the same thread[3].

In order to evaluate this hypothesis, we mapped out the time being referenced in our texts on a clause-by-clause basis. For each clause in the texts we indicated the time which was referred to. We distinguished between events that occurred at a single instance in time (atomic events) and events or states that occurred over a span of time (repeated atomic events, extended events, and states). For atomic events we allowed for both a specific time at which it occurred and for a non-specific time that indicated the range of uncertainty. We allowed time spans to have both specific end points and unspecified end points as well.

An example from our corpus with its associated temporal structure may illustrate these labels, the complexity of the texts under consideration, and how we propose pronoun generation is affected.

[3]Note that the discourse thread may change between the two references in question. This would be signaled by a change in time in clauses between the two references in question. We are interested in whether the two references are in the same (use a pronoun) or different (use a definite description) thread.

Example 2:

(47a) Questioned about the criminal activities of the football club,

(47b) **Mrs. Mandela** maintained

(47c) that **she** had never had any control over them.

(48a) This despite testimony from a half dozen former members

(48b) that they even had to get permission to go in and out of **her** yard.

(49a) **Mrs. Mandela** also said

(49b) **she** had disbanded the club

(49c) after **her** husband asked **her** to, despite evidence to the contrary.

(50) **Mrs. Mandela** faced questions from more than 10 lawyers representing various victims and the panel of commissioners and their investigators.

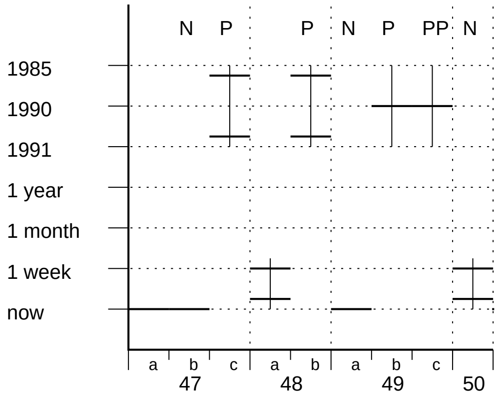

Figure 1: Temporal Structure for Example 2

Figure 1 contains the temporal structure for (each clause) of sentences 47-50 of one of our texts. We observe two threads in this discourse fragment, one dealing with events at the "now"-time of the story, the other telling past events (1985-1991). The corresponding sentences (also broken into clauses[4]) are contained in Example 2.

Notice that sentence 47 consists of three clauses. The first two (47a and 47b) describe atomic events that are taking place at the "now" time of the story (during the proceedings against Mrs. Mandela). The third clause (47c) refers to an indefinite span of time in the past (during which Mrs. Mandela's football club existed). Note, the use of the past perfect in (47c) indicating the change in time and setting.

As Figure 1 illustrates, there is a name (N) reference to Mrs. Mandela in (47b), and a pronominal (P) reference to her in (47c). This pronoun is used even though there is a change in time between (47b) and (47c) and is explained because this is a short-distance reference since it

[4]We take a clause to contain a finite verb.

is the second reference to Mrs. Mandela within the same sentence (a condition that overrules the time change hypothesis).

(48a) represents a change back to the time of the proceedings (note the discourse-deictic reference (Webber, 1991) "This"). (48b) again points to the time in the past, though this is not explicitly marked linguistically as it was in (47c). Here world knowledge must be used to understand the time referenced in this clause. Note, however, that the time would have to be part of the input to a generator, and thus our rules are completely well-defined from the generation perspective.

The use of a pronoun to refer to Mrs. Mandela in (48b) is warranted by our hypothesis, because the previous mention of Mrs. Mandela in (47c) references the same time as is referenced in (48b).

Because there is a time change between (48b) and (49a), our hypothesis explains the appearance of the proper name in (49a) even though it occurs just after a pronoun (in (48b)) co-specifying the same character.[5] The pronouns in the remainder of (49) are explained because they are subsequent references within the same sentence (despite the fact that they refer to an unspecified time in the past which is different from the time referenced in (49a)). Finally, the use of a name in (50) is again indicated by the change in time between (49c) and (50).

5 Ambiguities

Of course, the choice of referring expression is not only guided by discourse structure, there is also an influence due to ambiguities. Dale (1992) generated referring expressions so that their referents could be distinguished from the other discourse entities mentioned in the context. This strategy can be interpreted as: Generate a pronoun whenever it is not ambiguous. However, how one should define context is not quite clear. For this definition we choose a span of text considered important in our previous work on anaphora resolution (Strube, 1998), and define a referring expression as ambiguous if there is a competing antecedent (i.e. another discourse entity matching in number and gender) mentioned in the previous sentence or to the left of the referring expression in the current sentence. Of the 437 referring expressions in the texts we analyzed, 104 were considered ambiguous by this definition. Of these only 51 were realized as a definite description. Thus a rule which specifies use of a definite description if a pronoun would be ambiguous according to this definition appears to be too strict. Therefore we need to consider ambiguous cases in more detail.

[5]Note that the use of this definite description cannot be explained by a topic shift since there is no topic shift in between the previous text and (49). At least two discourse entities (*"Mrs. Mandela"* and the *"football club"*) are constant, only Mrs. Mandela's husband does not occur in the immediately preceding sentences.

Consider the following excerpt taken from later in the same story that contains Example 1. For convenience, references to the main character appear in **bold**, while those to a competing antecedent appear in *italics*.

> **Example 3:**
> **Mr. Curtis** might have lived out **his** live in obscurity if it were not for *a New Haven television reporter, Jim Hoffer, of station WTNH, who* got an anonymous tip last summer that **Mr. Curtis** had been attending college. Using a hidden camera, a crew taped **Mr. Curtis** going to **his** classes at Southern Connecticut State University, and *a producer* recorded a conversation with **him** in a cafeteria. In a clear, steady voice, **Mr. Curtis** can be heard to say that **he**...

Concentrating on the references to Mr. Curtis, we question when a pronoun can be used, and when the name should be used. To handle cases where competing antecedents occurred, we turned to pronoun resolution algorithms. Our intuition was that a pronoun could be generated to refer to a particular discourse entity if a pronoun resolution algorithm would choose that entity as a referent for the pronoun. To our knowledge, there are only two focus-based pronoun resolution algorithms that are specified in enough detail to work on unrestricted naturally occurring text: Brennan et al. (1987) using the definition of utterance according to Kameyama (1998), and Strube (1998). Strube evaluated the effectiveness of these two algorithms on the task of pronoun resolution in some naturally occurring texts. Because Strube's algorithm showed significantly better results, we have turned to it for guidance in pronoun generation.

The idea is that if we want to refer to a discourse entity, `E`, but there is a competing antecedent, `C`, we look to Strube's algorithm in the following way. If Strube's algorithm would resolve a pronoun to be `E`, we use a pronoun. If, instead, Strube's algorithm would prefer `C` as the referent of the pronoun, we would use a definite description to refer to `E`.

We evaluated this idea along with several other alternatives (e.g., using discourse structure alone, using long- and short-distance rules ignoring the ambiguity, distinguishing between first and subsequent reference within a sentence, using a definite description whenever there was ambiguity) on the ambiguous examples in our corpus. Our analysis showed that the use of Strube's algorithm showed improvement, but it seemed to be too liberal with suggesting pronouns when the competing antecedent was in the previous sentence. Assuming Strube's algorithm reflects human processing of referring expressions (something **NOT** claimed), what this means is that in our texts the writer chose to generate a definite description even though a pronoun would have been informationally adequate to resolve the referent correctly. This occurred most frequently over sentence boundaries. When a competing antecedent is within a sentence, however, Strube's algorithm appears to be quite effective.

Thus the rule we settled on acknowledged the importance of sentence boundaries and is shown in Figure 2.

1. *If* this is the first occurrence of X in the current sentence and
   (a) *if* there is a competing antecedent in the *previous* sentence, use a definite description;
   (b) *if* there is a competing antecedent in the *same* sentence (i.e., to the left) and
      i. *if* Strube's algorithm would resolve a pronoun in this position to be X, use a pronoun;
      ii. *else* use a definite description;
2. *if* this is a subsequent occurrence of X in the current sentence and
   (a) *if* there is an intervening competing antecedent, use a definite description;
   (b) *if* there is no intervening competing antecedent, use a pronoun.

Figure 2: Realization of X when Competing Antecedents Exist

6 Anaphoric Referring Expression Generation Algorithm

In the previous sections we have argued that in some instances of anaphoric expression choice, the threaded discourse structure must be taken into account. For the particular texts that we analyzed, we argue that threads defined in terms of the time referenced in a clause are appropriate for use. Other kinds of discourses will also exhibit a threaded structure, but the threads themselves might be defined by different means. In addition to discourse structure, the potential for ambiguity must be considered. Additionally, cases of long- and short-distance anaphora should be handled independently of the threaded discourse structure considerations.

Based on these findings, we propose the algorithm for realizing anaphoric expressions shown in Figure 3. Note in this algorithm we refer to the notion of a discourse thread which might be defined differently for different kinds of texts.

7 Empirical Data

We applied the algorithm described in the previous section to three texts from the *New York Times*. Articles ranged from a frontpage article to local news. We applied the algorithm to all references to persons in these texts. The algorithm was correct in 370 cases (84.7%), and wrong in 67 cases (15.3%). In Figure 4 we show the distribution over the rules specified in the algorithm.

In order to interpret the results of the algorithm, we must have some comparison. We use a simple scheme which does not consider threaded discourse structure and which handles ambiguous cases very conservatively.

1. *If* this is a **long distance anaphoric reference** (i.e., if the previous reference to X was more than two sentences prior) use a **definite description**;
2. *else*
   if this is an unambiguous reference (i.e., there is no competing antecedent) and this is an **intra-sentential anaphor** (i.e., this is not the first mention of X in the current sentence) use a **pronoun**;
3. *else*
   if this is a **thread change** (i.e., the previous reference to X occurred in a thread different from the one in which the current reference occurs) use a **definite description**;
4. *else*
   if there is a **competing antecedent** (i.e., another object in the previous or current sentence that matches the type and number of X) use the **rule found in Figure 2**;
5. *else*
   for the **remaining cases** (i.e., unambiguous cases when time remains the same) use a **pronoun**.

Figure 3: Algorithm for Generating an Anaphoric Referring Expression for Discourse Entity X

This scheme is shown in Figure 5 for comparison purposes.

The results of applying these rules give 343 correct cases (78.5%) and 94 incorrect ones (21.5%). Hence our algorithm reduces the error rate by 28.9%.

8 Related Research

A significant amount of work in linguistics has investigated the use of different kinds of anaphoric referring expressions in discourse and their relationship to ease of comprehension. See Arnold (1998) for a discussion of several of the factors involved in referring expression choice. In many cases the various factors seem to af-

Rule Name		number	percentage
All Rules			
	correct	370	84.7%
	wrong	67	15.3%
Long Distance (1)			
	correct	46	97.9%
	wrong	1	2.1%
Intra-sentential (2)			
	correct	168	96%
	wrong	7	4%
Time Rules (3 & 5)			
	correct	116	72.5%
	wrong	44	27.5%
Ambiguous (4)			
	correct	40	72.7%
	wrong	15	27.2%

Figure 4: Results of the Algorithm

1. *If* this is a **long distance anaphoric reference** (i.e., if the previous reference to this item is greater than two sentences prior) use a **definite description**;

2. *else*
if there is a **competing antecedent** (i.e., another object in the previous or current sentences that matches the pronoun which would be used to refer to this entity) use a **definite description**;

3. *else*
the anaphoric expression would be **unambiguous** so use a **pronoun**.

Figure 5: Simple Algorithm

fect the *accessibility* of a referent (where accessibility is intended in a broad sense to cover both "topic accessibility" (Givon, 1983) and accessibility due to factors such as recency of mention). Basically, the more accessible a referent the more underspecified a referring expression should be. Accessibility explains the apparent "name-name penalty" as examined in Gordon & Hendrick (1998), for example.

Our work argues that factors beyond accessibility must be considered in anaphoric expression choice. It is consistent with work such as Vonk et al. (1992) whose experiments indicate that a referring expression "... that is more specific than is necessary for the recovery of the intended referent ... marks the beginning of a new theme concerning the same discourse referent."(Vonk et al., 1992, page 304). They argue that such overspecified expressions are serving a discourse function of indicating boundaries. This work does not define what a discourse segment boundary actually is. On the other hand, using the definition of time change as a boundary condition, our work is consistent with their hypothesis. Interestingly, Vonk et al. (1992) found that in discourses where a theme change was well marked by other means (e.g., by a preposed adverbial phrase or a subordinate clause indicating time or place) that pronouns were much more common even though a new theme was begun. Presumably such phrases mark the theme change well, and thus it is not necessary to also mark the change via an overspecified description.

Approaches which define discourse segments on the basis of reference resolution (Sidner, 1979; Suri & McCoy, 1994; Strube & Hahn, 1997) are not useful for our purposes because they require referring expressions for recognizing segment boundaries. In contrast to these approaches, we define segment boundaries independently from reference resolution so that in this respect our work is in line with Grosz & Sidner's (1986) definitions.

9 Future Directions

In analyzing our data, there are several places for further consideration. One problem is that our rule which indicates a definite description should be used in a time change overgenerates definite descriptions. Following Vonk et al. (1992) we plan to investigate whether definite descriptions might best be viewed as boundary markers and whether other markers of discourse boundaries (e.g., preposed adverbial phrases) are found in places where our algorithm suggests a definite description because of a time change but a pronoun appears in the text.

In addition to evaluating more texts under this scheme, there are several places where we will attempt to tighten our methodology. One of these is in the time analysis. Our current analysis distinguishes between four types of time and is driven by both semantic cues in the text (e.g., adverbial time phrases) and changes in tense. Nakhimovsky (1988) also uses changes in "time scale" as a marker for changes in time. We plan to investigate this to see whether it explains more of the examples. Nakhimovsky also describes several other markers for a setting change, and these will also be investigated to see if they are indicative of definite description use.

Another line of future research involves further investigation of the ambiguous cases. Our current rule was developed by evaluating several different possibilities (e.g., using time change rules, different pronoun resolution algorithms) and selecting a rule that explains most of the cases. Still, the number of ambiguous cases is fairly small and analyzing more texts and concentrating on cases where the current rule makes an incorrect prediction may lead us to a more robust rule.

Finally, further thought must be put into evaluating the algorithm. In particular, our current evaluation methodology presupposes that the human writer has chosen the best anaphoric expression. It may be interesting to see if there are differences between reading time, eye movements, or comprehension of stories with the human-generated expressions and those using our algorithm. Thought must be put into setting up such experiments and into interpreting results. For example, a faster reading time may not indicate a better choice of referring expression (since the writer may have been interested in an effect other than ease of comprehension). On the other hand, such experiments have the potential for producing a more adequate evaluation of the methodology than does mimicking the human-produced text and should be looked into further.

10 Conclusions

Pronouns occur frequently in texts and have been hypothesized to play a significant role in text coherence. Yet, pronoun generation has not been studied in detail. If future natural language generation systems are to produce coherent, natural texts, they must use rules for generating pronouns that produce pronouns in roughly the same places that human-produced texts do. Moreover, the rules must be based on information that would be available to a sentence generator. At the same time, in order to evaluate rules, they must be based on information that can be gleaned from a text.

In this work we have argued that discourse structure in terms of multiple threads provides an explanation for patterns of pronoun use in naturally occurring text. As a particular instantiation of a threaded discourse structure, we looked at changes in setting, as indicated by changes in time. That is, even in places where a pronoun would be unambiguous, a definite description might be used when the time of the sentence is different from the time of the sentence in which the previous mention was made. This hypothesis provides an explanation for many of the uses of definite descriptions found in the studied texts. Other uses of definite descriptions occur because of ambiguities. We have suggested a rule which addresses when such ambiguities should not preclude the generation of a pronoun. Our scheme appears to be a reasonable explanation for the patterns of pronoun use found in our corpus.

Acknowledgments. This work was done while the first author was visiting the Institute for Research in Cognitive Science and while the second author was a postdoctoral fellow there (NSF SBR 8920230). We would like to thank Jennifer Arnold and the centering group at UPenn for helpful discussions. We would also like to thank the anonymous reviewers.

References

Appelt, D. E. (1981). *Planning Natural-Language Utterances to Satisfy Multiple Goals*, (Ph.D. thesis). Stanford University. Also appeared as: SRI International Technical Note 259, March 1982.

Appelt, D. E. (1985). Planning English referring expressions. *Artificial Intelligence*, 26(1):1–33.

Arnold, J. E. (1998). *Reference Form and Discourse Patterns*, (Ph.D. thesis). Stanford University, Department of Linguistics.

Brennan, S. E., M. W. Friedman & C. J. Pollard (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics,* Stanford, Cal., 6–9 July 1987, pp. 155–162.

Dale, R. (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. Cambridge, Mass.: MIT Press.

Dale, R. & E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

Genette, G. (1980). *Narrative Discourse: An Essay in Method*. Ithaca, N.Y.: Cornell University Press.

Givon, T. (1983). Topic continuity in spoken English. In T. Givon (Ed.), *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam, Philadelphia: John Benjamins.

Gordon, P. C. & R. Hendrick (1998). The representation and processing of coreference in discourse. *Cognitive Science*, 22(4):389–424.

Grosz, B. J., A. K. Joshi & S. Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. & C. L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Kameyama, M. (1998). Intrasentential centering: A case study. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in Discourse*, pp. 89–112. Oxford, U.K.: Oxford University Press.

Mann, W. C. & S. A. Thompson (1988). Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

McDonald, D. D. (1980). *Natural Language Production as a Process of Decision Making Under Constraint*, (Ph.D. thesis). MIT.

McKeown, K. R. (1983). Focus constraints on language generation. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence,* Karlsruhe, Germany, August 1983, pp. 582–587.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge, U.K.: Cambridge University Press.

Moens, M. & M. Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Nakhimovsky, A. (1988). Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43.

Passonneau, R. (1996). Using centering to relax Gricean constraints on disocurse anaphoric noun ophrases. *Language and Speech*, 39(2):229–264.

Prince, G. (1982). *Narratology: The Form and Functioning of Narrative*. Berlin: Mouton.

Reichman, R. (1985). *Getting Computers to Talk like You and Me*. Cambridge, Mass.: MIT Press.

Reiter, E. (1990). Generating descriptions that exploit a user's domain knowledge. In R. Dale, C. Mellish & M. Zock (Eds.), *Current Research in Natural Language Generation*. London: Academic Press.

Rosé, C. P., B. DiEugenio, L. S. Levin & C. Van Ess-Dykema (1995). Discourse processing of dialogues with multiple threads. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge, Mass., 26–30 June 1995, pp. 31–38.

Sidner, C. L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English*. Technical Report AI-Memo 537, Cambridge, Mass.: Massachusetts Institute of Technology, AI Lab.

Strube, M. (1998). Never look back: An alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, Vol. 2, pp. 1251–1257.

Strube, M. & U. Hahn (1997). Centered segmentation: Scaling up the centering model to global referential discourse structure. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society,* Palo Alto, Cal., 7–10 August 1997.

Suri, L. Z. & K. F. McCoy (1994). RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics*, 20(2):301–317.

Vogt, J. (1990). *Aspekte erzählender Prosa: Eine Einführung in Erzähltechnik und Romantheorie* (7th ed.). Opladen: Westdeutscher Verlag.

Vonk, W., L. G. Hustinx & W. H. Simons (1992). The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3/4):301–333.

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

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.