Discriminating Anaphoric and Non-Anaphoric Definite Nouns:

A Unified Memory-Based Model

by

Aryn A. Pyke

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Institute of Cognitive Science

Carleton University

Ottawa, Ontario

August 13, 2007

© Aryn A. Pyke, 2007

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-33508-6

Our file Notre référence

ISBN: 978-0-494-33508-6

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

When readers encounter a noun, they rapidly associate it with a mental referent. Definite noun phrases ("the dog") often refer to referents that were previously mentioned in the text (anaphoric use), however such phrases are also frequently used to introduce a new referent into the text (non-anaphoric use; e.g., Poesio & Vieira, 1998). A model of preliminary (first-pass) referent assignment must explain: i) how readers assign referents to non-anaphoric and anaphoric definite nouns, and ii) how readers implicitly determine the status of each definite noun (anaphoric or non-anaphoric).

The present research addresses this Anaphoric Classification Problem, and assesses the potential for anaphoric nouns to be misinterpreted as new referents on encounter (misclassify-as-new error). Two accounts of referent assignment are considered: special-purpose versus memory-based. Under the special-purpose view, readers process definite nouns with an anaphoric bias, and strategically attempt to assign a discourse referent before resorting to new-referent assignment. Under this view, misclassify-as-new errors may be rare, though preliminary assignment may be prolonged for difficult anaphors (e.g., metaphoric anaphors) and non-anaphors (which violate the anaphoric bias). The present research, however, supports an unbiased view of preliminary assignment, in which general-purpose memory mechanisms bring a referent to mind (Gerrig & O'Brien, 2005).

A unified model, MEMBRAL (Memory-Based Referent Retrieval), was developed and computationally implemented to address the classification issue. All referents in memory, including specific referent tokens and generic referent types,

compete for retrieval. The referent with the highest activation (discourse referent vs. referent type) determines whether or not the noun is interpreted as an anaphor. MEMBRAL clarifies the influence of anaphor word choice (including metaphoric anaphors), and predicts that word choice will primarily affect the accuracy rather than duration of preliminary referent assignment.

Four experiments were conducted to investigate the accuracy and duration of preliminary referent assignment (2 eye-tracked reading studies, 2 cloze studies). Results were highly congruent with those in the MEMBRAL simulation. Thus, this research supports the view that preliminary referent assignment to nouns is memory-based, and confirms that a memory-based account is sufficient, in practice, to account for a high rate of success at preliminary referent retrieval.

Acknowledgements

I would like to thank my parents, Sandra and Dale Pyke, for your unflagging love, encouragement and support -- it has meant more than you know, and unfortunately more than I can find the words to say. I both wouldn't have and couldn't have done this without you both. I would similarly like to thank my wonderful sister, Kyra Pyke, although we were in separate universities, in separate fields, and in separate cities, it was a source of comfort to feel like we were, in some way, in this together. You are, as you always have been, wise beyond my years.

I would like to express my sincere gratitude to my committee members, Dr. Robert West, Dr. Bruno Emond, Dr. Masako Hirotsu and Dr. Edward O'Brien, for generously providing your wisdom and your time. Your insight and input are much much appreciated. I also want to express my thanks to Dr. Andrew Brook, whose vision and determination brought this Institute of Cognitive Science into being; and to Dr. Deepthi Kamawar, Dr. Robert Stainton and my Masters supervisor, Dr. Frank Kschischang, for their mentorship and encouragement, which have had a lasting impact.

I am also indebted to my fellow PhD candidates at the Carleton Institute of Cognitive Science, who have shared this journey with me and/or helped to blaze the trail before me. Special thanks go to the old guard: the Latin Theatre Research Group, Matthew Rutledge-Taylor, Jacky Tweedie, Terry Stewart, Sam Scott, and Jerzy Jarmaz, for their companionship on some of the long walks through the snow (uphill, both ways...). And an extra special mention goes to Robert Thomson -- thank you for your strength, insight, sensitivity, calmness in crisis, sense of balance, broad shoulders, and

for being even more stubborn than me.

To my dear friends Catherine Chubey, Louise Vaillancourt, Helen Belepolsky and Paul Chamberlain, I also owe many thanks. You helped me feel at home when I was far from home, your doors and ears were always open, and you reminded me, through your example, of the value of perseverance. Best of all, you largely refrained from asking: "Are you done school yet"?

Last, but certainly not least, to my supervisor, Dr. Jo-Anne LeFevre, I cannot be too erudite in my accolades! Seriously, I can never thank you enough for adopting this academic orphan. Your ethics, frankness, wit and wisdom will always remain a source of both professional and personal inspiration. You know I could go on (and on) but in deference to your preference for brevity I will, albeit reluctantly, leave it at that.

Table of Contents

Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	ix
List of Figures	x
List of Appendices	xiii
CHAPTER 1: Introduction.....	1
Noun Anaphors.....	3
Special-Purpose Search Account.....	3
Memory-Based Account.....	4
The Anaphoric Classification Problem	5
Traditional Framework.....	8
Accounts to Address Anaphoric Classification	14
Investigative Approach and Existing Evidence.....	30
Synopsis.....	37
CHAPTER 2: Research Program.....	39
Main Manipulation: Semantic Overlap	39
Cognitive Modelling Methodology	41
Experimental Methodology	45
Hypotheses	52
Overview of Experiments.....	55

Summary.....	61
CHAPTER 3: Experiment 1.....	63
Method.....	65
Results	67
Discussion.....	68
CHAPTER 4: Experiment 2.....	75
Method.....	77
Results	80
Discussion.....	90
CHAPTER 5: Experiment 3.....	100
Method.....	102
Results	103
Discussion.....	109
CHAPTER 6: Experiment 4.....	116
Method.....	119
Results	123
Discussion.....	136
CHAPTER 7: The MEMBRAL Computational Model	144
What process is being modeled?	144
Criteria for a Cognitive Model	147
Implementation Architecture: Python ACT-R	149
Model Operation.....	154

Simulation Data	163
Contrast with Other Models	174
Limitations of MEMBRAL	188
Summary.....	192
CHAPTER 8: Summary and Conclusions	194
Synopsis of Contributions	194
Future Work.....	199
Concluding Remarks	201
REFERENCES.....	203
APPENDIX A: Stimulus Materials.....	214
APPENDIX B: Descriptive Statistics for Critical Nouns	226
APPENDIX C: Reading Rates for Context Sentences in Experiment 4	227

List of Tables

Table 2.1	Overview of experiments indicating the paradigm, independent variable manipulations, and the key dependent measures which provide information on the accuracy and/or duration of preliminary referent assignment	56
Table 6.1:	(a) Time Course ANOVA Results with Trial (1-39) as covariate to control for within-block 'practice' effects (Exp. 4)	142
	(b) Time Course ANOVA Results with Trial (1-78) as covariate to control for cross-block 'practice' effects (Exp. 4)	143
Table B1:	Mean Statistics for Critical Nouns	226

List of Figures

Figure 1.1: Traditional Account of Noun Referent Assignment.....	9
Figure 1.2: Extending the special-purpose account of anaphor resolution to address the classification problem.....	18
Figure 1.3: A fundamentally flawed conception of how the memory based account could be extended to non-anaphors.....	21
Figure 1.4: Proposed Extension of the Memory-based Account to Address the Anaphoric Classification Issue	22
Figure 2.1: Input-Output diagram for the MEMBRAL Model: Given a context and a critical noun, MEMBRAL will output (assign) a referent.....	45
Figure 4.1: Sentence Response Times (ms) for sentences containing literal category anaphors versus metaphoric anaphors in the present research (Pyke) and prior research (Gibbs, 1990; Lemaire & Bianco, 2003; Noveck et al., 2001; Onishi & Murphy, 1993; Ortony et al., 1978).....	83
Figure 4.2: Time Course Data (ms) on first pass fixations (lower part of bars) and net fixations (total bar heights) for the different word positions in the critical sentence (Exp. 2)	87
Figure 4.3: Synopsis of key time course measures (ms) (Exp. 2).....	88

Figure 4.4: Distribution of where the extra fixation time is spent in the metaphoric versus literal sentences (Exp. 2).....	89
Figure 5.1: Percent of anaphoric completions made for anaphorically- intended nouns (Exp. 3).....	105
Figure 5.2: Difficulty ratings for the anaphoric trials (Exp. 3).....	107
Figure 5.3: Difficulty ratings (1=very easy; 10 =very difficult) by interpretation type (Exp. 3).....	108
Figure 6.1: Percentage of comprehension questions answered correctly (Exp. 4).....	124
Figure 6.2: (a) Time Course (ms) for Critical Sentences in Experiment 4, Block 1 (unmarked).....	132
(b) Time Course (ms) for Critical Sentences in Experiment 4, Block 2 (marked)	133
Figure 7.1: Mean activation levels for the intended referent (R) and the generic competitor (G), before (lower part of bars) and after activation spread from the anaphor (full bars) in the simulation.....	166
Figure 7.2: Comparison of MEMBRAL's preliminary retrieval inaccuracy (misclassify-as-new errors) with human performance data.....	167

Figure 7.3: Simulated and Actual Preliminary Processing Times for Critical	
Nouns.....	171

List of Appendices

APPENDIX A: Stimulus Materials.....	214
APPENDIX B: Descriptive Statistics for Critical Nouns.....	226
APPENDIX C: Reading Rates for Context Sentences in Experiment 4.....	227

CHAPTER 1: Introduction

The capacity for linguistic communication is a fundamental human characteristic, and the processes underlying this ability are a central focus of cognitive research. In particular, how accurate and immediate are our comprehension processes? Some historical models have suggested that certain key aspects of comprehension are deferred until sentence end (e.g., reference assignment and non-literal interpretation; Grice, 1975; Searle, 1979). However, recent evidence suggests that readers and listeners interpret sentences incrementally, as the words become serially available (e.g., Just & Carpenter, 1980; Sedivy, Tanenhaus, Chambers & Carlson, 1999). In particular, human interpreters make a rapid preliminary association of a noun phrase (e.g., "the doctor") with a referent almost immediately after they encounter the noun (Just & Carpenter, 1980; see also Cooper, 1974; Dell, McKoon, & Ratcliff, 1983; Sanford & Garrod, 1989). The term 'referent' is used here in the cognitive sense (as in Gundel, Hedberg, & Zacharski, 2001) to mean a *mental representation* of the person or object in question. Without presently making any claims about the nature of (or the existence of) internal structure, I note that referents are conceptual rather than lexical representations.

At the time a noun is encountered, the processing system does not have access to information in the sentence that occurs after that noun. Thus, a reader's preliminary referent assignment may be influenced only by the preceding context and the noun itself. The objective of the present research is to empirically investigate and computationally model this process of preliminary referent assignment for definite noun phrases (Pyke, West & LeFevre, 2007a, 2007b). In so doing, I identify and investigate

an underlying challenge for this process that has received little attention in the literature (herein: the **Anaphoric Classification Problem**), and, in particular, I explore the procedural basis for a particular type of preliminary error: a **misclassify-as-new error** - where a noun that was intended to refer to a person or object already present in the discourse is instead interpreted as introducing a new referent.

Such comprehension processes are typically rapid and automatic, and tend to occur without conscious effort or deliberation (Ferreira, Baily & Ferraro, 2002; see also Sperber & Wilson, 1995). Certainly, conscious deliberation may sometimes play a role in comprehension, but such pragmatic processing is subsequent to (and operates on) the output of automatic preliminary processing. Preliminary referent assignments may therefore be subject to subsequent adjustment, as the rest of the sentence becomes available, and more global, pragmatic and/or deliberative processes come into play. However, the present focus is on the nature and accuracy of the rapid, automatic, non-conscious preliminary referent assignment process itself, which is invoked immediately on encountering a noun (e.g., Just & Carpenter, 1987; Kintsch, 1998).

In the remainder of this chapter, I will further motivate and characterize the problem of preliminary referent assignment using some example noun phrases in context, and I will describe two alternative accounts of how such phrases are initially processed: the **special-purpose search account**; and the **memory-based account**. Then I will outline my general approach for investigating the accuracy (and the time course) of preliminary assignment, and for arbitrating between the alternative models. The finer details of the research program will be deferred until Chapter 2.

Noun Anaphors

I am primarily, though not exclusively, interested in preliminary referent assignment for nouns used as anaphors. In general, an anaphor is any label (word or expression) that refers back to a person or object that has already been previously mentioned within the current conversation or text. For example, the noun “surgeon” in (2a) is used as an *anaphor* that is intended to co-refer with the *antecedent* “surgeon” in (1).

(1) The patient discovered that his careless surgeon had made an error.

(2a) The surgeon faced a malpractice suit.

(2b) The doctor faced a malpractice suit.

The alternative anaphor “doctor” in (2b) illustrates that the anaphor expression (“doctor”) need not, in general, match the antecedent expression (“careless surgeon”). Nonetheless, readers are often able to re-activate the intended referent almost immediately after encountering an anaphoric noun (Dell, McKoon, & Ratcliff, 1983; Sanford & Garrod, 1989), that is, after the first-pass processing of the noun. What is the nature of the cognitive process that underlies this ability? Two alternative views from the literature will be outlined below: the special-purpose search account, and the memory-based account.

Special-Purpose Search Account

Special-purpose search models suggest that when readers encounter a definite noun phrase, they undertake a strategic search for an antecedent through a mental

representation of the discourse (e.g., Clark & Sengul, 1979; Kintsch, 1998; Kintsch & Vandijk, 1978; O'Brien, Plewes, & Albrecht, 1990). In such special-purpose search models, the discourse might be mentally represented as a proposition network, and the reader might systematically troll backwards through it to find the antecedent that co-refers with the current anaphor. Such special-purpose search accounts might differ in detail, for example, in terms of the hypothesized structure of a discourse representation, and in the order in which the discourse referents are considered as candidates (Cristea & Dima, 2001), however, they share the basic premise that the noun prompts an intra-discourse search process. After all, how else could one account for the fact that when reading a noun like "doctor" in (2a), readers come across the right referent (i.e., the particular surgeon mentioned earlier in the discourse)? As outlined below, the memory-based view of text processing (reviewed by Gerrig & O'Brien, 2005) provides an alternative answer to this question.

Memory-Based Account

Under a memory-based view of text processing, the successful retrieval of the intended referent need not require (nor constitute evidence of) a special-purpose intra-discourse search, because passive general-purpose memory processes often can automatically bring the intended referent to mind. In particular, under the resonance model (e.g., Gernsbacher, 1989; Myers & O'Brien, 1998; O'Brien, Raney, Albrecht & Rayner, 1997), the current information in working memory (i.e., the anaphoric noun) serves as a cue that automatically boosts the activation of other entities throughout long-

term memory -- including, ideally, the intended referent -- in accord with their semantic overlap with the cue. Thus, when the reader encounters the anaphor "doctor" (2b), the surgeon referent may automatically be re-activated via resonance (or spreading activation¹) in virtue of its semantic association with the anaphor concept. Thus, to account for readers' frequent success at referent reactivation during preliminary processing there may be "no need to invoke any process other than general purpose memory processing" (Gerrig & O'Brien, 2005, p. 230).

The Anaphoric Classification Problem

Both the special purpose and memory-based models are plausible accounts of readers' ability to successfully retrieve the intended referent for anaphoric nouns, like "doctor" in (2b). However, both accounts are, as thus far described, functionally incomplete as general accounts of noun referent assignment. In particular, they fail to explicitly address: i) how readers process non-anaphoric, herein *introductory* nouns -- that is, those that introduce a new referent into the discourse, like "patient" and "surgeon" in (1); and more crucially ii) how readers implicitly or explicitly determine whether a particular noun is anaphoric or introductory when they first encounter it; and iii) how accurate such determinations are.

Anaphoric nouns are not a natural kind. Any noun in the lexicon can serve in an anaphoric or introductory capacity. So, how do readers determine whether or not a

¹ For the present purposes there is no need to draw a distinction between the mechanisms of spreading activation and resonance.

given occurrence of a noun is anaphoric? The definite article “the” is often assumed to indicate that the intended referent is already familiar to the reader (Clark & Sengul, 1979; Garnham, 1989; Garrod & Sanford, 1977; Just & Carpenter, 1987; cf. Gundel, Hedberg, & Zacharski, 2001). Additionally, Clark and Haviland (1977) suggest that an assumption of anaphoricity is especially warranted when a definite noun phrase occurs near the start of a sentence, as the grammatical subject. Such cues might, in theory, establish the anaphoric status of the noun prior to its encounter.

In practice, however, grammatical subjects need not always be anaphoric (Francis, Gregory & Michaelis, 1999; Prince, 1992), and corpus analyses have indicated that definite noun phrases are equally likely to introduce new referents into the discourse as to refer to referents that have already mentioned (e.g., Fraurud, 1990; Gundel, Hedberg, & Zacharski, 2001; Poesio & Vieira, 1998). As a case in point, “patient” in (1) is introductory, although it is the grammatical subject of the sentence and is preceded by a definite article. Consequently, cues like the definite article and the sentence-initial position may not be strong or immediate indicators about whether the noun is anaphoric. In principle, then, readers are not privy to the anaphoric status of a noun *a priori*. Thus, when a reader encounters a definite noun phrase, the explicit or implicit determination of the noun's anaphoric status is part and parcel of the preliminary referent retrieval process. And sometimes the reader may initially misclassify an anaphoric noun as non-anaphoric (or vice versa).

In view of this classification issue, a complete cognitive account of the preliminary referent retrieval process for nouns must meet two criteria. First, it must account for

how the reader explicitly or implicitly makes an anaphoric/non-anaphoric classification of a noun. Second, in so doing it should serve to predict, for particular texts and anaphors, when the reader will be apt to (a) misclassify an anaphoric noun as a new referent (herein a **misclassify-as-new error**); and (b) misclassify a non-anaphoric noun as an anaphor (herein a **misclassify-as-anaphor error**).

A central objective of the present research is to highlight and address this classification issue, and to explore the potential for misclassify-as-new errors. The anaphoric classification issue and the potential for (preliminary) classification error have largely been overlooked in psychological accounts of noun processing. To clarify why this is the case, and to better situate the present research with respect to the previous trends in the literature, I provide a brief background of the traditional theoretical framework that has influenced research in this area. Then I discuss how each of the previously mentioned accounts (special-purpose search account vs. memory-based account) has been, or can be, extended to address this anaphoric classification issue for definite nouns. Finally, I outline the present research program to investigate the prevalence and procedural basis for misclassify-as-new errors, and to arbitrate between the two alternative accounts.

Traditional Framework

Much research on noun anaphors was preceded and influenced by research on pronoun anaphors (e.g., Caramazza, Grober & Garvey, 1977; Clark & Sengal, 1979; Ehrlich & Rayner, 1983). However, in contrast to nouns, pronouns are almost inherently anaphoric (Kintsch, 1998), so there was no precedent to consider the classification issue in the pronoun literature. This anaphoric status classification remained sidelined when the research framework applied to pronouns was transferred more or less verbatim to anaphorically-intended nouns: "How do readers resolve the pronoun/noun anaphor?" Referring to anaphorically-intended nouns as *noun anaphors* implicitly makes the anaphoric status of such nouns seem like a given. Furthermore, in the literature, the reader's task upon encountering an anaphorically-intended noun is often dubbed *anaphor resolution*. This terminology, adopted from pronoun research, under-represents the full scope of the reader's (and modeler's) task on encountering a noun, and has presumably diverted attention away from the classification issue.

To the extent that researchers explicitly acknowledged that nouns (vs. pronouns) frequently introduce new referents into the discourse, the classification problem was traditionally believed to be trivial. In particular, the pre-nominal article was assumed to reliably indicate the anaphoric status of the noun. That is, nouns preceded by an indefinite article ("a") were non-anaphoric, whereas nouns preceded by the definite article ("the") were anaphoric. Evidence that nouns preceded by "the" are frequently non-anaphoric is relatively recent (e.g., Poesio & Veria, 1998, Gundel et al., 2001).

Figure 1.1: Traditional Account of Noun Referent Assignment

The traditional framework can be summarized by the process diagrams in Figure 1.1, which reflect the following three assumptions:

Traditional Assumption 1: The pre-nominal article (e.g., "a" vs. "the") indicates the anaphoric status of the noun, and this information then dictates which special-purpose referent retrieval process to apply to the noun.

Traditional Assumption 2: For definite noun phrases (commencing with "the"), referent retrieval is then achieved via a special-purpose 'anaphor resolution process' that involves a search among discourse referents (e.g., Clark & Sengul, 1979, Kintsch & van Dijk, 1978, O'Brien, Plewes & Albrecht, 1990).

Traditional Assumption 3: It is only a matter of time before the right referent should be found within the discourse by this special-purpose 'anaphor resolution process', precluding any need for subsequent re-analysis (Sanford & Garrod, 1989).

This final assumption bears some elaboration in its defense. First, it was consistent with existing evidence, such as that from Dell, McKoon and Ratcliff (1983), who found

that immediately after their noun anaphors were read, the intended antecedent became reactivated in memory (i.e., was facilitated when used as a recognition probe). Further, in view of Assumption 1, readers would never make **misclassify-as-new** errors, and **misclassify-as-anaphor** errors. In principle, there is an additional possible type of error, namely, assigning the *wrong referent* from within the discourse (here, a **discourse-distractor error**). However, the possibility that the wrong discourse referent might be selected was marginalized by a contrast drawn between noun anaphors and pronouns. Pronouns carry little semantic information (i.e., number and gender), and are consequently apt to match more than one candidate referent in the context, whereas nouns held promise as semantically rich and therefore unambiguous anaphor devices. Lastly, Assumption 3 has been implicitly reinforced by a bias inherent in the terminology "anaphor *resolution*" (vs. "referent assignment") which has been used to describe the process/task, and which carries the presupposition that a reader will inevitably assign the intended discourse referent.

Although there are currently grounds to question each of the traditional assumptions, their impetus has continued to influence the design and interpretation of empirical research on referent assignment for definite nouns. For example, in contrast to the present research, most prior research has focused on the time course rather than the accuracy of noun anaphor processing (e.g., Almor, 1999, Clark & Sengul, 1979; Garrod & Sanford, 1977; Duffy & Rayner, 1990; van Gompel, Liversedge & Pearson, 2004; cf. Levine, Guzman & Klin, 2000), because accuracy, along with anaphoric status, has generally been considered a given. Thus, most prior research does not

explicitly entertain a delineation between a preliminary (error prone) assignment phase and a possible subsequent re-analysis phase (cf. Garrod et al., 1994), because referent retrieval for noun anaphors was presumably achieved by the single (thus, preliminary) special-purpose 'anaphor resolution process', which would be launched when a definite noun was encountered, and which would continue until it inevitably succeeded.

Certainly, some factors have been known to induce processing delays within anaphoric sentences, including increasing the distance between the anaphor and the antecedent (e.g., Clark & Senegal, 1979; Duffy & Rayner, 1990), and manipulating the degree of the semantic overlap between an anaphor (e.g., "bird"), and its antecedent (e.g., "robin" vs. "penguin"; Duffy & Rayner, 1990; Garnham, 1989; Garrod & Sanford, 1977; McKoon & Ratcliff, 1980; O'Brien et. al, 1997; van Gompel et al., 2005). However, in keeping with Assumptions 2 and 3, such factors were assumed to prolong rather than undermine the precision of the 'anaphor resolution process'. Consequently, any post-noun processing delays were attributed to spillover from the (preliminary) 'anaphor resolution process' (e.g., Duffy & Rayner, 1990). However, in the present research, I emphasize the potential for preliminary interpretation error (in particular, an anaphorically-intended noun may be initially misinterpreted as a new referent), which introduces an alternative interpretation of post-noun processing delays (i.e., they may result from revision processes when subsequent content in the sentence causes the reader to reconsider their initial interpretation of the noun as a new referent).

In summary, as a consequence of the traditional terminology and assumptions, it was not commonly asked *whether* the reader would correctly interpret (resolve) noun

anaphors, but rather *how* readers resolve noun anaphors. Although the existence of a special-purpose process (Assumption 2) has been called into question by the alternative memory-based account, and the availability of anaphoric status information (Assumption 1) has been called into question by corpus evidence, the impetus remained to refrain from asking *whether* the anaphor will be correctly resolved (i.e., recognized as anaphoric), in favour of asking how long this 'anaphor resolution process' will take.

In contrast, the framework in the present research directs attention to the classification issue and the potential for misclassify-as-new errors in readers' preliminary interpretations of anaphorically-intended nouns. In accord with Just and Carpenter's (1987) immediacy hypothesis, I expect a reader to make a relatively rapid, though possibly error prone, preliminary referent assignment to each definite noun. This preliminary referent assignment will be based on the noun and the preceding context. An alternative hypothesis is that the noun and preceding context may not always provide enough information to enable a referent assignment to be made, and in such cases, readers will refrain from making an assignment until additional information is available. According to this strategic deferral argument, readers may generally avoid preliminary assignment errors by avoiding preliminary assignments in such cases. Such strategic deferral may arguably come into play to prevent **discourse-distractor errors** when an anaphorically-intended noun, such as "bike" in (4), brings multiple discourse candidates to mind (e.g., Jill's bike and John's bike).

(3) Jill got a bike for her birthday, and so did John.

(4) The bike...

However, the strategic deferral argument does not generalize well to argue against the possibility of readers making **misclassify-as-new errors**, which are the focus of the present research. In particular, recall that deferral (vs. immediate assignment/classification) was motivated by the claim that the noun and the context may not always provide enough information to enable a referent assignment to be made. However, in support of the immediacy hypothesis, I argue it should actually always be possible for a reader to assign some referent to a noun without access to any additional information in the sentence past that noun. In particular, whether or not the noun brings to mind an existing referent from the discourse, the noun itself should always provide the reader with sufficient information (e.g., a prototype) to at least allow a new referent interpretation. For example, the reader can readily form an interpretation for the phrase in (5) without requiring access to the remainder of that sentence.

(5) The dog....

Because a large proportion of definite nouns are in fact non-anaphoric (e.g., Faurud, 1990; Poesio & Veria, 1998), new-referent assignments will often be appropriate. Consequently, the preliminary referent assignment process must be flexible enough to handle both introductory and anaphoric definite nouns. Thus, when a noun does not readily yield an anaphoric interpretation, there is no reason for readers to eschew a new-referent interpretation, and carry extra cognitive load by leaving the assignment incomplete. Consequently, readers are expected to make a preliminary referent assignment on encountering a noun, and this choice will effectively classify the definite noun as introductory or anaphoric. The accuracy of such preliminary

classifications obviously depends on the inner workings of the preliminary referent process, and the manner in which it attempts to accommodate (and discriminate between) both anaphoric and introductory definite nouns. Below I will discuss how the special-purpose search account and the memory-based accounts of preliminary referent assignment may be elaborated to address this classification issue.

Accounts to Address Anaphoric Classification

Both the special-purpose search account and the memory-based account can be extended to allow for the interpretation of introductory as well as anaphoric definite nouns. There is also another possible account, an introductory-bias account, which will be outlined first.

Introductory-Bias Account

In recognition that some definite descriptions evoke new referents, Garrod, Freudenthal and Boyle (1994) suggested that perhaps all definite noun phrases might automatically be treated as new referents on encounter. Under this view, readers would make fast and accurate interpretations of non-anaphoric nouns (i.e., nouns actually intended to evoke new referents). However, all anaphoric noun phrases would initially be misinterpreted as new referents, regardless of the degree of the noun's semantic overlap with the intended antecedent. Although possible in principle, a model that initially treats all nouns as new referents seems inconsistent with findings that the intended antecedent is sometimes immediately retrieved (e.g., Dell et al., 1983).

Extending the Special-Purpose Search Account to Non-Anaphors

Although the definite article is not a statistically reliable indicator that a noun is anaphoric, one could argue that readers nonetheless make a default assumption of anaphoricity for definite noun phrases. Thus a reader would automatically launch a special-purpose 'anaphor resolution process' to search for an antecedent upon encountering a definite noun phrase (e.g., Kintsch, 1998).

What would be the end product of this special-purpose process? One possibility is that the process would *necessarily* output some referent from within the discourse, as in Figure 1.2(a). Because such a preliminary assignment system is not capable of assigning a new referent, it would never make any misclassify-as-new errors. However, many definite nouns are not anaphoric, and such a system would result in a misclassify-as-anaphor error for each non-anaphoric definite noun (i.e., it would be initially associated with some referent within the discourse). The following example (6)-(7) illustrates that such a system is implausible. In particular, the system would be forced to associate a discourse referent (either the man or his car) to the noun "lamppost" in (7).

(6) The man walked towards his car.

(7) The lamppost.....

Alternately, rather than absolutely forcing a reader to select a discourse referent, the system could allow a reader to assign a new referent as a last resort. Thus, the 'anaphor resolution process' might operate as in Figure 1.2(b), so that it could have two alternative output states - it could select and output a referent from within the discourse,

or it could register a failure to do so. In this latter case, the default assumption of anaphoricity would then be abandoned and another process would be launched to initialize a new referent into the discourse (e.g., Just & Carpenter, 1987).

Importantly, because such a system allows new referent interpretations, it is susceptible, at least in principle, to making misclassify-as-new errors in preliminary assignment when processing anaphorically-intended nouns. However, because discourse referents are given preferential consideration, such errors may be very rare. At this level of description, it is not yet possible to predict the absolute likelihood that such a system would make a misclassify-as-new error, however, there are some timing predictions that fall directly out of Figure 1.2(b). In particular, under this account, non-anaphoric definite nouns (or anaphoric nouns misinterpreted as non-anaphoric) will presumably require two stages to process: 1) a strategic search for an antecedent that should fail, followed by 2) some process to associate a new referent to the noun. Thus, in preliminary processing, non-anaphors (and anaphors misinterpreted as non-anaphors) will take longer to process than anaphors. Anaphors misinterpreted as non-anaphoric in preliminary processing will also result in subsequent reanalysis delays if this preliminary misinterpretation is later detected and corrected.

There are some possible objections to this special-purpose account of referent assignment. When readers seemed to treat an anaphoric noun as a new referent, there was no accompanying delay in net sentence reading time (Levine, Guzman & Klin, 2000; see also Budiu & Anderson, 2002). These results suggest that readers may not invest time seeking an antecedent in the discourse before resorting to a new referent

interpretation. A second, more theoretical, reservation about this account pertains to cognitive efficiency. Under this account, an antecedent search is launched for every definite noun phrase. However, because a large proportion of definite nouns are non-anaphoric, the time and effort invested in an antecedent search will frequently be in vain. Having outlined the basic operation and implications of the special-purpose approach to the classification problem, I will next describe how the memory-based account may be adapted to apply to non-anaphoric nouns.

Figure 1.2(a): Extending the special-purpose account of anaphor resolution to address the classification problem. In (a) the process necessarily outputs a referent from within the discourse; in (b) the process will first consider and reject all discourse referents as candidates, and then can subsequently launch a process to assign a new referent.

Extending the Memory-Based Account to Non-Anaphors

In this section I will present an elaboration of the memory-based view to address the classification issue. In particular, how might a memory-based model of preliminary referent assignment accommodate the processing of non-anaphoric versus anaphoric noun phrases, and how prone would the system be to confusing the two? To set the stage for this discussion, I will first briefly recap the memory-based account as it applies to *anaphoric* nouns. The anaphorically-intended noun exerts a relatively immediate impact on the relative activations of referents in memory via general-purpose memory mechanisms (i.e., spreading activation/resonance). Such activation spread is promiscuous in that it is *not* dependent on a particular type of relation between representations (e.g., an IS-A relation, as in, a surgeon IS-A doctor). Instead, any semantic similarity or learned association can provide a conduit for the automatic spread of activation in memory². Thus, reading an anaphorically-intended noun like "doctor" in (2c) should cause some activation to spread to the intended referent -- the reader's representation for the careless surgeon mentioned in (1). However, the noun "doctor" may also result in some activation spread to the reader's representation of the patient, but this latter semantic association would not be as strong. The net resultant association strengths among mental concepts can be estimated via Latent Semantic

² Memory-based accounts often focus on the mechanism of resonance (Hintzman, 1986; Ratcliff, 1978), which raises the activation of each mental representation (referent) in virtue of any conceptual overlap with the cue (anaphor). But any and all automatic spreading activation effects (also) fall, in principle, within the auspices of a memory-based account, irrespective of the nature or level of the association pathway (e.g., within and/or between the lexical and conceptual representation levels). As will be discussed later in more detail, in the present research, the net resultant association strengths are estimated via Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998).

Analysis (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998).

A referent's net activation level during first-pass noun processing owes to two components (e.g., O'Brien et al., 1997): i) the activation boost it gets in virtue of the noun term currently being processed; and ii) its context-dependent pre-noun activation level. An example of a context-dependent influence on activation is recency (i.e., the distance in words or sentences between the anaphor and antecedent; e.g., Clark & Senegal, 1979; Duffy & Rayner, 1990). Together, these two activation factors (current noun and preceding context) will determine the most active referent, which will be the one retrieved and the one initially assigned to the noun.

Discussions of the memory-based view in the noun anaphora literature have not been very explicit about how the account would apply to introductory (vs. anaphoric) nouns, and how the anaphoric status of definite nouns is determined. There seems to be a tacit assumption that the memory-based account could address the classification issue in the same manner applied to the special-purpose account. That is, the overall operation of the system would be structurally similar to that depicted in Figure 1.2(b), but 'memory-based referent retrieval' would replace the 'anaphor resolution process', as in Figure 1.3. Thus, a reader would resort to interpreting a noun as a new referent if and when the memory-based referent retrieval process 'fails'. In the following discussion, I argue that this suggestion is fundamentally flawed, and I will offer an alternative elaboration of the memory-based view to address the classification issue (Figure 1.4).

Figure 1.3: A fundamentally flawed conception of how the memory-based account could be extended to address the classification issue.

Figure 1.4: Proposed Extension of the Memory-based Account to Address the Anaphoric Classification Issue.

The system depicted in Figure 1.3 implicitly misrepresents the scope and nature of the memory-based mechanisms, and, in so doing, fails to recognize the fundamental theoretical and procedural differences between the special-purpose account and the memory-based account. The memory processes are general-purpose rather than task-specific. As a consequence of their automatic, general-purpose operation, memory mechanisms might bring a discourse referent to mind in response to a noun stimulus, however, that is not the sole function or nominal goal of these mechanisms. As such, they cannot technically 'fail' at a task they never explicitly undertook, namely: retrieving a referent *from within the discourse*. From the perspective of these general-purpose mechanisms, which exert their impact throughout memory, the current discourse has no special status. In general, a person's memory is populated not only with representations of the particular referents from the current text, like the patient and the surgeon introduced in (1), but also with other particular referents with which the agent is familiar (e.g., their family doctor, their car), and with generic referent prototypes (e.g., doctor prototype, car prototype).

In the ensuing discussion, mental representations of specific particular referents will be prefaced by "R", for example, [R: my doctor], [R: the incompetent surgeon from (1)], whereas generic prototype referents will be prefaced with "G", for example [G:doctor] and [G:surgeon]. In virtue of the automatic operation of memory mechanisms, all these (kinds of referents will receive boosts in activation in response to a noun (e.g., "doctor"), and all kinds of referents are eligible for retrieval. Thus, sometimes the memory mechanisms may not result in the retrieval of a *discourse*

referent. But that does not imply that they fail to retrieve any referent. Concepts and referents outside those in the current discourse may become strongly activated in addition to, or even instead of, the intended referent (O'Brien & Albrecht, 1991).

The fact that general memory mechanisms might activate referents (including generic prototype referents) outside the current discourse may be considered non-ideal behaviour if such memory mechanisms are viewed solely from the perspective of processing *anaphorically*-intended nouns. However, the potential to retrieve of non-discourse referents clarifies that these general-purpose memory mechanisms could be conceived to support referent assignment, more generally – for both anaphoric and non-anaphoric nouns. That is, such mechanisms may bring about the retrieval of any type of information that would facilitate associating a referent to the noun: be that information about an existing referent, or information (e.g., a prototype) for evoking a new referent into the discourse. Thus, the memory-based account holds appeal as a unified process model appropriate for both anaphoric and non-anaphoric noun phrases. This system is depicted in Figure 1.4. After automatic spread of activation from the noun, the referent with the highest activation level in memory will be retrieved, regardless of whether it is a specific referent or a generic referent prototype.

To clarify the operation of this memory-based model, we can consider an example of how a non-anaphoric noun like "newspaper" in (2d) would be processed.

(1) The patient discovered that his careless surgeon had made an error.

(2d) The newspaper...kept the patient's mind off his pain.

An introductory noun like "newspaper" in (2d), would spread relatively little

activation to the referents in the discourse, [R:patient] and [R:surgeon], because the semantic overlap is low, but it will spread substantial activation to the generic prototype referent [G:newspaper]. The noun newspaper may also spread activation to other referents which are semantically related like [G:magazine], but the concept corresponding to the noun will have a stronger association and will receive more activation. Thus, after spread of activation from the noun, the referent [G:newspaper] will have a higher activation than either discourse referent and than the other generic referents, so [G:newspaper] will be retrieved and the noun will be interpreted as a new referent.

This newspaper example clarifies that this unified memory-based system in Figure 1.4 should always be successful in providing some referent for the noun during preliminary assignment. Regardless of the current discourse, a noun will always, at least, result in activation spread to some corresponding referent prototype, and this generic referent will be retrieved unless a specific referent (from the discourse or from another experience) achieves a higher activation.

Under this memory-based account, the only circumstance that might result in an outright failure to retrieve any referential representation during preliminary processing is in response to an unknown noun like "gryflez", because no association pathways are present to support the spread of activation to any of the particular or prototypic referents in memory. Thus, for an unfamiliar noun, no referent might reach the retrieval threshold (see also Budiu & Anderson, 2006). However, under normal circumstances, I suggest that a memory-based system should never fail to retrieve a referent in

preliminary assignment (cf. Figure 1.3).

However, the retrieved referent may not always be the intended referent. In particular the reader may misinterpret an anaphoric noun as a new referent. The implicit preliminary classification of a noun's anaphoric status will be determined by which referent achieves the highest activation. If the highest activation is attained by a referent from within the discourse, the reader will interpret the noun as anaphoric. If the highest activation is attained by a referent prototype, the reader will interpret the noun as introductory. A misclassify-as-new error will occur when, in response to an anaphorically-intended noun, a generic referent prototype achieves a higher activation than the intended discourse referent.

How prone to misclassify-as-new errors would this memory-based system be? The memory mechanisms themselves are unbiased in the sense that all discourse and prototype referents alike will receive activation from the anaphoric noun. However the discourse referents will have an activation head start, in virtue of having recently been created and/or accessed in memory. Consider the operation of the system for anaphoric nouns like "surgeon" in (2a) and "doctor" in (2b), which are repeated below for convenience.

(1) The patient discovered that his careless surgeon had made an error.

(2a) The surgeon....would face a malpractice suit.

(2b) The doctor....would face a malpractice suit.

When the noun "surgeon" in (2a) is read, activation will spread to all semantically associated referents throughout memory, both particular and prototypic. Thus,

"surgeon" will result in modest activation spread to the particular patient in this context [R:patient], and a large activation spread to the intended referent [R:surgeon], as well as a large activation spread to the surgeon referent prototype, [G:surgeon]. The anaphorically-intended noun surgeon will also result in activation spread to other semantically related referent types like [G:dentist] and [G:scalpel] however, among the referent prototypes, the one which receives the strongest spread of activation from the anaphoric noun will be the concept corresponding to the noun [G:surgeon]. Thus, [R:surgeon] and [G:surgeon] are the referent contenders which received the highest activation. However, the particular surgeon referent [R:surgeon] in the current discourse should have an activation head start because it was recently in working memory and was involved in any integrative comprehension process (e.g., Just & Carpenter, 1987) occurring at the end of (1). Thus, the pre-anaphor activation levels of the discourse referents will generally be higher than the pre-anaphor activation levels of the generic prototypes. Thus, in virtue of this context-dependent (pre-anaphor) component of activation, the discourse referent [R:surgeon] should end up at a higher net activation level than the prototype [G:surgeon] when the spread of activation has settled. Thus, misclassification errors should be rare for such antecedent-match anaphors.

In (2b), the alternative anaphoric noun "doctor" would result in some spread of activation to the patient referent [R:patient], a larger spread of activation to the surgeon referent [R:surgeon], and an even larger spread of activation to the doctor referent prototype [G:doctor]. Note, the intended referent [R:surgeon] will not receive as much

spread of activation in response to the anaphor "doctor", as it would have for the anaphor "surgeon", because there is less anaphor-antecedent semantic overlap. To make matters worse, the noun "doctor" will cause slightly more activation to spread to the generic prototype referent [G:doctor] than it will to the intended referent [R:surgeon]. The net activation level of the intended referent [R:surgeon] will be a function of its pre-anaphor activation level and the activation spread it receives in virtue of its semantic overlap with the anaphor. This discourse referent should have an activation head start (i.e., a higher pre-anaphor activation level than the generic competitor [G:doctor]). Thus, whether or not the intended referent [R:surgeon] is retrieved will depend on whether its context-dependent pre-anaphor component of activation (i.e., head start) is sufficient to compensate for the slightly stronger spread of activation from the anaphor "doctor" to [G:doctor]. Thus, to predict misclassify-as-new errors, it is not the absolute activation level attained by the intended referent that is germane, but its activation level relative to other referents throughout memory, especially that of the referent prototype associated with the noun.

In summary, the present extension of the memory-based account provides a unified procedural treatment that is applicable to both anaphoric and non-anaphoric nouns. In contrast to the special-purpose account in Figure 1.2(b), this system does not perform a prolonged discourse search to assign a non-discourse referent. Consequently, non-anaphors (and anaphors misunderstood as non-anaphors) could be processed in a comparable amount of time as anaphors.

Synopsis of the Alternative Accounts

The above discussion has established that both the special-purpose search account and the memory-based account can be elaborated to address the anaphoric classification issue (i.e., how do readers discriminate and handle introductory versus anaphoric nouns during preliminary referent assignment).

How do these special-purpose and memory-based accounts fundamentally differ? In general, a special-purpose process is one that performs a specific/modular function and is invoked in response to a specific type of stimuli (e.g., a linguistic stimulus, in particular, a definite noun phrase). Under the special-purpose search account, readers approach the interpretation of each definite noun with an anaphoric bias. A goal-directed 'anaphor resolution process' is invoked to seek a discourse referent for a definite noun. The process operates over a circumscribed representation in memory: a representation of the current discourse. The special-purpose system may fail to select a referent from within the discourse, and, after this prolonged search, the reader can then resort to assigning a new referent to the noun.

In contrast, the unified memory-based account attributes the ability to understand anaphors to the general-purpose operation of memory and its associative mechanisms (e.g., spreading activation/resonance), which operate automatically in response to a wide range of stimuli and in support of almost every cognitive task. These mechanisms operate throughout memory and are not confined to the content of the current discourse. Thus, they can also serve to retrieve generic representations (e.g., prototypes) for the instantiation of a new referent into the discourse. Consequently, the memory-based

account can afford a unified procedural treatment applicable to both anaphoric and non-anaphoric nouns.

Having outlined these two processing alternatives to address the classification issue, I next outline my general investigative approach to arbitrate between them, and to obtain evidence about the likelihood of misclassify-as-new errors.

Investigative Approach and Existing Evidence

Both the memory-based model and the special-purpose model are, in principle, susceptible to making misclassify-as-new errors when processing anaphorically-intended nouns. However, such misclassify-as-new errors should be relatively rare under the special-purpose account, because referents within the discourse are given preferential consideration as candidates. Furthermore, it is a given that humans have general-purpose memory mechanisms at their disposal, so the existence of a special-purpose process (to search for an antecedent in the discourse) would only be evolutionarily warranted if it provided a substantial increase in accuracy over that which could otherwise be achieved with the general-purpose mechanisms.

Unfortunately, this deduction -- that a special-purpose system should predict fewer misclassify-as-new preliminary errors than a memory-based system -- does not provide a testable empirical prediction in terms of an absolute expected level of accuracy for preliminary referent assignment. However, once evidence about readers' accuracy of preliminary assignment/classification is available, the ability of a computationally implemented memory-based model to achieve comparable performance would provide compelling support for the memory-based account, on the basis of parsimony.

Thus, the general approach taken in the present research is as follows: i) to evaluate existing evidence and obtain new evidence about the accuracy and duration of preliminary referent assignment for definite noun phrases, ii) to develop a computational implementation of my memory-based model, and evaluate whether such a model can account for the human performance patterns. Thus, I provisionally hypothesized that preliminary referent assignment operates in accord with the unified memory-based account.

Two factors that are generally believed to contribute to difficulty in anaphor interpretation are: i) increasing the distance, in words or sentences, between the anaphor and its antecedent (e.g., Clark & Senegal, 1979; Duffy & Rayner, 1990; Levin, Guzmann & Klin, 2000; O'Brien et al., 1997), and ii) reducing the degree of semantic overlap between the anaphor and its antecedent (e.g., Duffy & Rayner, 1990; Garnham, 1989; Garrod & Sanford, 1977; McKoon & Ratcliff, 1980; O'Brien et. al, 1997; van Gompel et al., 2005). Under a memory-based account, such factors should increase the likelihood that readers will make a misclassify-as-new error, because these factors will have the effect of lowering the net activation of the intended referent in memory at the time the anaphor is processed.

In particular, increasing the distance of the antecedent increases the time since the reader last accessed that referent representation in memory, and so its activation level will have decayed accordingly (i.e., if the referent has not been recently mentioned, its activation level will be relatively low by the time the reader reaches the anaphor). In contrast, reducing anaphor-antecedent semantic overlap has the effect of reducing the

amount of (additional) activation that will automatically spread to the intended referent in response to the anaphor itself. Because the factors of distance and semantic overlap (individually or in concert) result in a lower net activation of the intended referent, they reduce the likelihood that the intended referent will be retrieved, and misclassify-as-new errors should sometimes occur.

Levine, Guzman and Klin (2000) found some support for this claim. In particular, readers seemed to treat a noun anaphor (e.g., dessert) as a new referent when the antecedent was distant and had low semantic overlap with the anaphor (e.g., éclair), and the context included a distractor referent. They reported that in such contexts, readers did not associate the anaphor to either the intended referent or to the distractor. These results are important for helping to explicitly dispel the traditional assumption (Assumption 3) that noun anaphors are always correctly resolved. Aside from this study, however, the literature provides little *direct* information about the occurrence of misclassify-as-new errors during preliminary referent assignment.

There is, however, some additional indirect evidence that such manipulations of distance and/or semantic overlap can result in misclassify-as-new errors. In particular, anaphoric sentences often take longer to process when the anaphor has low semantic overlap with its antecedent (e.g., Duffy & Rayner, 1990; Garnham, 1989; Garrod & Sanford, 1977; McKoon & Ratcliff, 1980; O'Brien et. al, 1997; van Gompel et al., 2005). Such delays may arise when readers detect and correct a misclassify-as-new error made during in preliminary assignment (note, such comprehension processes need not be accompanied by conscious awareness). Admittedly, there may be other

interpretations for (and contributions to) this overall processing delay for anaphoric sentences when anaphor-antecedent semantic overlap is low. However, such net sentence delays were notably absent in cases where readers made misclassify-as-new errors from which they never recovered (Levin, Guzmann & Klin, 2000; see also Budiu & Anderson, 2002), which suggests that such delays are contributed by revision processes rather than say, a prolonged preliminary assignment (e.g., special-purpose account). Thus, processing delays may sometimes result from, and provide evidence of, misclassify-as-new errors in preliminary assignment when semantic overlap between the anaphor and the antecedent is low.

In the present research, my objective was to gather more evidence about the potential effect of semantic overlap on the incidence of misclassify-as-new errors in preliminary referent assignment. Semantic overlap itself has been manipulated in a variety of ways in research on noun anaphora. Most frequently, researchers have manipulated the typicality of the antecedent (e.g., Duffy & Rayner, 1990; Garnham, 1989; Garrod & Sanford, 1977; van Gompel et al., 2005). Adjective modifiers have also been used to manipulate the degree of semantic overlap between anaphors and potential antecedents (e.g., Corbett, 1984; O'Brien, Duffy, & Myers, 1986; O'Brien et al., 1997; Wiley, Mason, & Myers, 2001). In the present research, I kept the context and antecedent invariant, (1), and manipulated semantic overlap by varying the choice of anaphoric noun, as exemplified by the three alternative anaphors below (2a-2c).

(1) The patient discovered that his careless surgeon had made an error.

(2a-c) The surgeon/doctor/butcher faced a malpractice suit.

Thus, I employed antecedent-match anaphors ("surgeon"), literal category anaphors ("doctor"); and metaphoric category anaphors ("butcher"). Under a memory-based account, activation spreads according to semantic similarity/overlap, so at the mechanistic level of processing the literal/metaphoric distinction reduces to a manipulation of semantic overlap (Budiou & Anderson; 2004; see also Lemaire & Bianco, 2003). In particular, because the concept BUTCHER shares semantic overlap with SURGEON (e.g., both are professions that involve cutting flesh), activation will spread between them and may enable the reader to retrieve the intended referent for the anaphor for the anaphor butcher in (2c).

Latent Semantic Analysis (LSA) can provide a quantitative estimate of semantic overlap (for reviews see Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998)³, to gauge the amount of activation that will spread to the intended referent in response to each potential anaphor. In the present example, the estimated strength of association for (surgeon-surgeon) is 1, for (surgeon-doctor) is 0.63; and for (surgeon-butcher) is 0.06 (Latent Semantic Analysis at the University of Colorado at Boulder, 2003). In general, metaphoric anaphors should provide lower levels of anaphor-antecedent similarity than those obtained with other manipulations of semantic overlap. Thus, the desire to probe the performance limits of the preliminary referent assignment process motivated the inclusion of metaphoric anaphors in the materials for the present research.

Under a memory-based account, the retrieval of the intended referent is directly

³ In LSA, a large text corpus is analyzed and lexical concepts are represented as vectors in a multidimensional semantic space. Similarity between concepts is computed as the cosine of the angle between their corresponding vectors. Thus, the maximum similarity (i.e., of a concept to itself) is 1.

dependent on its level of activation, and the spread of activation the referent receives from the noun is, in turn, dependent on the degree of semantic overlap between the anaphor and the antecedent. Thus, the low semantic similarity in the metaphoric case means that that anaphor ("butcher") may not result in a sufficient spread of activation to the intended referent, [R:surgeon], to support its retrieval over the generic competitor [G:butcher]. Consequently, readers may frequently make misclassify-as-new errors during preliminary referent assignment for metaphoric anaphors. Alternately, under a special-purpose account, readers have an anaphoric-bias, and give preferential consideration to the discourse candidates so that they need only recognize (rather than recall) the intended referent. As such, under a special-purpose account, the use metaphoric anaphors may not be expected to greatly undermine the accuracy of preliminary referent assignment.

Existing evidence is inconclusive as to whether or not this strong manipulation of semantic overlap would induce frequent misclassify-as-new errors. Researchers have found that readers usually take longer to process sentences with metaphoric anaphors than corresponding sentences with literal anaphors (Gibbs, 1990; Lemaire et al., 2003; Noveck, Bianco, & Castry, 2001; Onishi & Murphy, 1993; cf. Ortony, Schallert, Reynolds, & Antos, 1978). Under a memory-based account, preliminary errors will be frequent and delays may result as readers use content in the remainder of the sentence to later revise such incorrect preliminary assignments. Alternately, under a special-purpose account, readers may invest additional time in preliminary assignment, and largely avoid preliminary assignment errors, and subsequent reanalysis delays. That

said, it is also possible that even a special-purpose process might be susceptible to misclassify-as-new errors for metaphoric anaphors. In such a case, a delay might still be expected during preliminary assignment, because readers should conduct a prolonged (possibly exhaustive) discourse search, which nevertheless fails, prior to resorting to any new referent interpretation. If readers subsequently revise such erroneous assignments, a reanalysis delay may also contribute to the net sentence processing delay. Thus, net delays for sentences with metaphoric anaphors are compatible with many interpretations – and with both processing accounts. Additional evidence is required to clarify what processes (preliminary referent assignment or revision) and what sentence regions (anaphor vs. post-anaphor) contribute to such delays. Consequently, as detailed in Chapter 2, I used an eye-tracked reading paradigm to gather more detailed evidence about the time course of processing within anaphoric sentences. However, even localized delays are difficult to interpret (e.g., the special-purpose account predicts that preliminary referent assignment will be prolonged for metaphoric anaphors whether or not they are correctly resolved), so I also sought to obtain direct, independent evidence about the accuracy of a reader's preliminary interpretation for literal and metaphoric anaphors. In particular I applied a cloze paradigm (i.e., a complete-the-sentence task). A more comprehensive discussion of the data, methodology and predictions is postponed until the next chapter.

In summary, the existing evidence does not allow for a clear-cut arbitration of the special-purpose and memory-based accounts. Both processing accounts are consistent with longer mean reading times for sentences with metaphoric anaphors, however they

make different predictions about the accuracy of the initial referent assignment, and about how the extra time might be distributed. To evaluate such predictions and to interpret processing delays, I require additional evidence about the distribution of such delays within the sentence, and, ideally, additional independent evidence about the accuracy of the preliminary assignments.

Synopsis

In the present research I investigated the nature of preliminary referent assignment for definite noun phrases. In particular, I raised and addressed an issue that has not received much attention in the literature, the anaphoric classification issue: Because nouns frequently introduce new referents into the discourse, the preliminary referent assignment process must somehow accommodate (and discriminate between) introductory as well as anaphoric nouns. Two alternative processing accounts, the special-purpose account and the memory-based account, make different predictions about the potential for readers to misinterpret an anaphorically-intended noun as a new referent (misclassify-as-new error); and consequently suggest different interpretations for the processing delays in anaphoric sentences involving difficult (metaphoric) anaphors.

To explore the performance limits of the preliminary referent assignment process (and to assist in arbitrating between these accounts), in the present research I compare the processing of non-anaphoric nouns with three types of noun anaphors (literal antecedent-match anaphors, literal category anaphors and metaphoric category anaphors) in terms of the duration and accuracy of preliminary referent assignment.

Chapter 2 provides an overview of the research program and methodology, which includes both empirical investigation and computational modelling. In particular, the present research involved four empirical studies: two cloze studies (Chapters 3 and 5) and two eye-tracked reading studies (Chapters 4 and 6). The research culminated in the creation of a computational ACT-R model of the preliminary referent retrieval process (MEMBRAL: MEMory-Based Referent retrievAL; Pyke, West & LeFevre, 2007a, 2007b), which is presented in Chapter 7, together with the simulation results which establish that the operation of the model is consistent with the human performance measures from the empirical research (noun gaze, regressions, and types of interpretation error). Finally, Chapter 8 will provide synopsis of contributions and outlines some directions for future work.

CHAPTER 2: Research Program

In this chapter I provide an overview of my research program to investigate the nature and accuracy of the preliminary referent assignment process, and to evaluate my hypothesis that this process is memory-based (vs. special-purpose). In particular, I will detail and motivate my manipulations, methods, and measures, and I will develop some empirical-level predications/hypotheses for the memory-based and special-purpose accounts. The present research involves computational modelling and two different experimental paradigms (eye-tracked reading and cloze tasks), which will be described and compared with some alternative paradigms that have previously been applied to research on anaphors.

In detailing the methodology, it is important not to lose sight of the larger agenda that motivates such manipulations and measures (i.e., the 'why'). Consequently, to contextualize the present research program, I begin by reviewing the motives behind my main manipulation, semantic overlap.

Main Manipulation: Semantic Overlap

As discussed in Chapter 1, the performance of a memory-based system (spread of activation) is directly influenced by the degree of semantic overlap between the anaphor and the antecedent. In contrast, a special-purpose system specifically seeks an antecedent in the discourse, so its performance might be expected to degrade gracefully (or in some cases not at all, Stewart & Heredia, 2001) to manipulations of semantic overlap. Consequently, to arbitrate between these two views, I required an experimental manipulation that would be apt to reveal the performance limits of the

preliminary referent assignment process.

Thus, a strong manipulation of semantic overlap was applied which included anaphors with very low overlap (i.e., metaphoric anaphors). In particular, I used three types of anaphors: antecedent-match anaphors ("surgeon"), literal category anaphors ("doctor"); and metaphoric category anaphors ("butcher"), as exemplified in (1)-(2a-2c).

(1) The patient discovered that his careless surgeon had made an error.

(2a-2c) The surgeon/doctor/butcher faced a malpractice suit.

In general, metaphoric anaphors should be subject to the same noun processing and referent assignment mechanisms as literal anaphors when they are first encountered because the reader has no a priori cue to launch a distinct, hypothetical 'metaphor processing mechanism'. In particular, under the memory-based account, any decrease in preliminary assignment accuracy for the metaphoric case is not the result of metaphoricity per se, but arises because metaphoricity is correlated with low anaphor-antecedent semantic overlap which, in turn, is correlated with the amount of activation that should automatically spread to the intended referent as a result of reading the anaphor. The degree of anaphor-antecedent overlap can similarly influence accuracy for literal anaphors, as supported by the results of Levin, Guzman and Klin (2000).

For example, under the memory-based account, reducing the semantic overlap between the antecedent ("incompetent surgeon") and the anaphorically-intended noun ("surgeon" vs. "doctor" vs. "butcher") will reduce the amount of activation that will automatically spread to the intended referent upon encountering the noun anaphor. Thus, after spread of activation, the net activation of the intended discourse referent, [R:

careless surgeon], may fall below that of a competing generic referent ([G:surgeon] or [G:doctor] or [G:butcher], respectively). For example, if the anaphor "butcher" causes the reader's prototype referent [G: butcher] to reach a higher activation level than that of the intended referent [R: careless surgeon], the prototype will be retrieved and the reader will make a misclassify-as-new error.

Thus, the three anaphor types in the present research can be considered to represent different degrees of semantic overlap on a continuum, and furthermore, they provide the coverage of a wide range of this continuum (i.e., from high overlap antecedent-match anaphors to low overlap metaphoric anaphors). Thus, this manipulation of semantic overlap should provide useful evidence about the performance limits of the preliminary referent assignment process.

Next I will outline the methodologies used to assess whether a memory-based account is compatible with any degradation of accuracy (i.e., in crease in misclassify-as-new errors) arising from this manipulation. In particular, the present research program involves both cognitive modelling and behavioural experimentation (cloze tasks and eye-tracked reading). Below, I motivate and outline these two complimentary methodological aspects of the present research.

Cognitive Modelling Methodology

The ultimate objective of the present research was to provide a description of the *process* of preliminary referent assignment (vs. a description at the level of behavioural patterns). Consequently, the methodology included the development of a working computation model of preliminary referent retrieval, MEMBRAL (Memory-Based

Referent Retrieval; Pyke, West & LeFevre, 2007a, 2007b). In this section, I will motivate my use of cognitive modelling methodology in general, and the use of the ACT-R cognitive architecture in particular (Anderson, Bothell, Byrne, Douglass & Lebiere, 2004; see also Stewart & West, in press).

Most psycholinguistic accounts of noun processing take the form of verbal descriptions rather than computational models. Developing a computational model that automatically performs a function is a valuable methodology for further detailing the account, and for bringing to light overlooked issues and aspects of the task. For example, although the anaphoric classification issue for definite nouns has not received much attention in the noun anaphora literature, this issue has been addressed in many (non-cognitive) Natural Language Processing algorithms (e.g., Bean & Riloff, 1999; Vieira & Poesio, 2000). Such algorithms, however, often involve multiple passes forward and backward through the text, and do not directly speak to the development of a *cognitive* model of referent assignment for a noun *at its time of encounter* (cf. the READER model of Just & Carpenter, 1987).

In general, computer implementations need not impose any psychological constraints on the character of a model, nor do they necessarily simulate the process in real-time units. Consequently, the cognitive plausibility of one-off, task-specific computational models is sometimes open to question, even when their parameters can be mathematically adjusted to provide a fit with a particular measure of human performance. To address such concerns, I tested the performance of my model against several human performance measures (e.g., noun processing duration, preliminary

accuracy, types of interpretation errors), and I implemented my model using the ACT-R cognitive modeling architecture (Anderson et al., 2004). ACT-R is a programming architecture that has in-built operational constraints that are psychologically motivated, particularly with regards to timing and memory, which makes it well suited as a platform for implementing a memory-based model. This ACT-R architecture has a long-standing track record for supporting the creation of models that explain and predict human performance in a vast range of cognitive tasks, including language processing (e.g., Budiu & Anderson, 2004; Emond, 1997). Further discussion of the implementation and simulation details will be postponed until Chapter 7.

MEMBRAL operates according to the unified memory-based account outlined in Chapter 1. In contrast to non-cognitive Natural Language Processing models, the goal of a cognitive model is not to assign the correct referent, but rather to reflect the (presumably error prone) patterns of actual human performance. In particular, the role of this model in the present research is to address the following question: Can human performance levels (whatever they happen to be) be achieved by a unified memory-based account? If so, this evidence will provide strong support for the unified memory-based account over a special-purpose account.

I mentioned above that the operation of the model will be compared with several human performance measures, however, the primary function of the model is to predict which referent will be assigned (i.e., retrieved) for a given noun in a given context, as shown in Figure 2.1. For a given particular context and a given particular noun (which can be either anaphoric or introductory), MEMBRAL will assign/retrieve a referent.

Among humans, there will be individual variation in performance -- that is, whereas participant 1 might assign the correct referent for a given context and anaphorically-intended noun, participant 2 might make a misclassify-a-new error in their preliminary assignment. Such variation may reflect different (pre-experimental) association strengths in the participants' semantic networks, which would, in turn, affect the pattern/strength of activation spread in response to the anaphor. In MEMBRAL, the effects of individual variability are simulated by adding noise (i.e., small random fluctuations in the activation levels of referents in memory). Consequently, the model can be used to predict the incidence of misclassify-as-new errors (vs. correct co-referential assignments) for under the different conditions of semantic overlap, and notably, for the individual items within each condition.

To evaluate the compatibility of this memory based model with respect to human performance, we require correspondingly detailed information about the incidence of misclassify-as-new errors made by human participants. In the next section, I discuss the empirical methodology for obtaining such data.

Figure 2.1: Input-Output diagram for the MEMBRAL Model: Given a context and a critical noun, MEMBRAL will output (assign) a referent.

Experimental Methodology

The computational modelling methodology outlined above provides a way to generate data about the accuracy of preliminary referent assignment *in theory* (i.e., according to my unified memory-based theory). Now we require methods to obtain data on the accuracy of preliminary referent assignment *in practice* (i.e., by human readers). In particular, how can we determine the nature of a reader's preliminary interpretations and assess the incidence of misclassify-as-new errors?

Because a reader's preliminary referent assignment for a noun occurs on-line, (e.g., mid-sentence), and is not directly observable by an experimenter, information about the nature and accuracy of preliminary assignments is somewhat challenging to empirically obtain. Note that the reader's final interpretation at sentence end may reflect the influence of subsequent process and information. In particular, in the course of

processing the remainder of the sentence after the noun anaphor, readers may detect that the content in the remainder of the sentence is incompatible with their preliminary interpretation (Garrod et al., 1994). In general, when readers detect that an earlier error in interpretation may have been made, they often regress their eyes back to the site of the initial misinterpretation to do an overt reanalysis (e.g., Altmann, Garnham & Dennis, 1992; Meseguer, Carreiras & Clifton, 2002; Rayner, 1990). Consequently, I hypothesized that reader's eye-movements may provide useful evidence about the effectiveness of preliminary referent assignment. In particular, the likelihood of regression to a noun anaphor may provide some indication of the likelihood that it was treated as a new referent in preliminary analysis (i.e., the incidence of misclassify-as-new errors).

Thus, in the present research I employed an eye-tracked reading paradigm (Experiments 2 and 4). In eye-tracked reading, the gaze location of participants is monitored while they read each sentence, which provides data on the incidence of regression (here, I am interested in regressions to the anaphorically-intended noun), and the distribution of processing time within the sentence. Eye-tracking data has never previously been reported for the current manipulation of semantic overlap (literal anaphors vs. metaphoric anaphors)⁴.

There were, however, two studies (Budiou & Anderson, 2002; Janus & Bever, 1985) that used self-paced reading with word-by-word presentation to investigate the

⁴ Giora (1997) mentioned, mistakenly, that a study by Janus and Bever (1985) employed eye-tracking, however their study actually used a word-by-word self-paced paradigm.

allocation of processing time within sentences containing literal and metaphoric anaphors. The evidence obtained, however, was inconclusive with regards to whether there was an increased incidence of misclassify-as-new errors for metaphoric anaphors in preliminary processing. In these studies, the anaphoric noun was overwritten upon presentation of the next word and could not be physically re-read, precluding overt reanalysis (i.e., such studies could not provide regression data). Thus, any reanalysis conducted by the readers should have produced a delay when reading the post-anaphor remainder of the sentence (covert reanalysis). However, these researchers found no evidence of increased processing time (more frequent reanalysis) in the remainder region in the metaphoric condition. These data are, however, subject to interpretation. In particular, there is reason to believe that the time course patterns produced in a self-paced word-by-word study may not reflect those in natural reading (Rayner, 1990; see also von Gompel et al, 2005). In contrast, in eye-tracked reading a participant can see a whole sentence in its entirety, and need not press a button to see each subsequent word, so the eye-tracking paradigm may produce more ecologically valid results.

Aside from issues of ecological validity, the data are still inconclusive. Although the data pattern was consistent across these two studies, the two pairs of researcher argued for very different interpretations of the data. In particular, whereas Janus and Bever (1985) suggested that their readers made correct initial assignments, which precluded any need for revision (compatible with a special-purpose account), Budiu and Anderson (2002) suggested that their readers often made incorrect new-referent assignments to the metaphoric anaphors in preliminary analysis but did not revise them

(compatible with a memory-based account).

These differing interpretations raise an important issue about the limitations of remainder-of-sentence processing time as an indicator of the accuracy of initial assignment. The presence of increased remainder-of-sentence delays (as semantic overlap drops) depends on the presumption that readers will detect and correct their preliminary misinterpretations. In other words, the absence of post-anaphor delays does not necessarily imply that the preliminary interpretations were accurate.

Thus, Budiu and Anderson speculated many preliminary misinterpretations went unrectified, which seems especially likely in cases where their sentences had metaphoric anaphors plus metaphoric verbs. For example, in the context in which it was given, (8) was intended to denote a group of women chatting animatedly.

(8) The hens clucked noisily.

Because such sentences, in and of themselves, are amenable to a coherent though contextually incorrect literal interpretation, they would not supply the reader with a strong, local incongruity to cue revision. Consequently, in the current research, the endings of the anaphoric sentences were designed, as much as possible, to be incongruous under a new-referent interpretation, to ensure that an incorrect preliminary assignments would be accompanied by tell-tale evidence in terms of overt and/or covert reanalysis (i.e., percentage of trials with regressions to the anaphor and/or delays in processing the remainder of the sentence).

In summary, because of the above concerns about the self-paced paradigm and the inconclusive data, an eye-tracking paradigm was used in the present research

(Experiments 2 and 4). This time course data collected in the present research should enable a check of the ecological validity of the time course data obtained with the word-by-word self-paced paradigm (Budiou & Anderson, 2002; Janus & Bever, 1985), and can be compared with eye-tracking data reported for semantic overlap manipulations based on antecedent typicality (Duffy & Rayner, 1990; van Gompel et al., 2005).

Although eye-tracking data can be very valuable in evaluating different processing theories, eye-tracking data itself has some notable limitations. As Altmann (1994) states: “Eye movements may well provide a finer window within which to view local processing difficulties, but the view through that window is far from clear.” For example, time course data do not reveal exactly when a reference assignment has been made, nor do they reveal which referent was initially (or ultimately) assigned.

Thus, an additional task was developed to gain insight into the nature of the referent that a reader has assigned to the definite noun phrase prior to reading the rest of the sentence. In some previous research on pronoun anaphors, participants were presented with the text stimulus only up to and including the anaphor (Gernsbacher, 1989). For example:

(9) John lent Bill money because he... [participant probed to pick an antecedent]

Immediately after the anaphor (i.e., the pronoun “he”), participants were required to state which of the available antecedents (e.g., John or Bill) they believed was the intended one. Although this explicit resolution task provides data on which referent the reader is inclined to assign mid-sentence, the task itself is unsuitable for the present research on *noun* anaphors. By requiring the participant to select an antecedent, the

task presupposes and informs the participant that the critical word is an anaphor. Thus, this task would undermine the present objective to determine whether the reader in fact recognizes whether a noun is an anaphor at the time of encounter (i.e., misclassify-as-new error). The explicit resolution task was suitable for research on pronouns because they are almost always anaphoric (Kintsch, 1998), and if the interest is assessing the possibility of assigning the wrong referent from the discourse (i.e., discourse-distractor errors). So, for the present research on noun anaphors, a cloze task was developed instead.

In general, in a cloze task, part of a sentence is missing and participants must write in words to complete the sentence. For the present purposes, the incomplete sentence will be one that begins with the critical noun. For the example context (1), the endings of (2b) and (2c) would be omitted to produce (2b') and (2c') respectively (nb. each participant saw only one version or the other).

(1) The patient discovered that his careless surgeon had made an error.

(2b') The doctor _____.

(2c') The butcher _____.

The sentence endings generated by the participants should reveal whether or not they associated the critical noun to the intended referent (i.e., the surgeon), as in (2b', completion 1) and (2c', completion 1), or treated it as a new referent, as in (2b', completion 2) and (2c', completion 2).

(2b', completion 1) The doctor didn't apologize.

(2b', completion 2) The doctor who referred Lisa to that surgeon felt bad.

(2c', completion 1) The butcher got his medical degree revoked.

(2c', completion 2) The butcher down the street delivered meat to the invalid.

Incidentally, the above examples are based on actual data from the present research. In regular reading, upon encountering a noun, the reader has access to information in the preceding context and the information in the sentence up to and including the critical noun (e.g., "The butcher"). Hence, interpretations obtained in this cloze task should yield insight about the accuracy of the initial interpretations that would be constructed during regular reading.

An alternative method of assessing whether a particular referent was retrieved in preliminary assignment is to present the antecedent term as a recognition or naming probe, after the anaphor (e.g., Dell, McKoon, & Ratcliff, 1983; Gernsbacher, 1989; Levine et al., 2000; Stewart & Heredia, 2002). Cases in which the anaphor successfully prompts the retrieval of the correct referent contribute to a facilitation effect in the response time to the antecedent probe (i.e., relative to a non-antecedent probe, or relative to if the antecedent-probe were presented elsewhere, such as prior to the anaphor).

Although this probe paradigm has played a productive role in anaphora research, it was not ideally suited for the present purposes. In particular, facilitation effects capture aggregate patterns, but do not provide an absolute measure of the incidence of misclassify-as-new errors for each anaphor. Furthermore, facilitation for an antecedent probe such as "SURGEON" could be expected after the anaphorically-intended noun "doctor" whether preliminary referent assignment had activated the particular intended

surgeon referent [R:surgeon] as in (2b', completion 1); or simply the generic doctor referent [G:doctor] to evoke a new discourse entity as in (2b', completion 2). Thus, cloze data may sometimes provide more direct information about which (kind of) referent was assigned (i.e., particular discourse referent [R:surgeon] or a new referent of a semantically similar type [G:doctor]).

In summary, in the present research, I used cloze tasks as well as eye-tracked reading to gather evidence about the accuracy, duration and nature of the preliminary referent assignment to definite nouns. I next outline some theoretical predictions about the expected data patterns (accuracy and time course) for the memory-based versus special-purpose accounts.

Hypotheses

The ultimate test of memory-based account will be in the compatibility of the human data patterns (e.g., percentage regressions to anaphor), with the performance of the model. However, some course-grained comparative predictions can be made between the memory-based and special-purpose accounts, directly in terms of the expected data patterns for the empirical studies themselves. Obviously the performance of (predictions for) a special-purpose system will be dependent on the details of its implementation (which vary in different version of this account; Cristea & Dima, 2001), however, some general behaviours can be inferred.

Predictions about Accuracy of Preliminary Referent Assignment

Although it is not possible to predict absolute accuracy levels, the performance of a memory-based system will necessarily degrade as semantic overlap is reduced.

Performance in a special purpose system might also degrade in response to a manipulation in semantic overlap (i.e., it could increase the likelihood of rejecting the intended candidate during the discourse search). Thus, degraded performance for anaphors with low semantic overlap could be compatible with either a memory-based or special purpose account. However, it is possible that performance for metaphoric anaphors might not degrade in a special-purpose system if the anaphoric bias is strong.

Stewart and Heredia (2002) found support for the possibility of near perfect performance in the metaphoric case (see also Janus & Bever, 1985). In particular, Stewart and Heredia (2002) conducted a probe naming study on *spoken* metaphoric anaphors (whereas the current focus is on anaphors in text). After listening to a verbal context, participants were required to name visually presented probe words at various points while listening to a sentence with a metaphoric anaphor. Two types of probes were used for the metaphoric anaphors (e.g., “chicken”): (a) semantic associates of the metaphoric sense of the noun (e.g., “coward”), and (b) semantic associates of the literal sense of the noun (e.g., “rooster”). Response time facilitation for each probe was measured relative to its own baseline control word that was matched to the probe for length and production difficulty. Stewart and Heredia found maximal facilitation for the metaphoric associate probes (“coward”) immediately at the offset of the anaphoric noun (“chicken”). Their findings suggest that the listeners had correctly interpreted the metaphoric noun before hearing any more of the sentence. Such facilitation data is not entirely conclusive, however, and speech processing may be different than text processing. However, near perfect performance in cloze tasks (and few regressions)

for metaphoric anaphors would tend to favour a special-purpose account.

Predictions about Duration of Preliminary Referent Assignment

Perhaps ironically, eye-tracking evidence was gathered primarily for the purpose of providing information about the *accuracy* of preliminary referent assignment (e.g., % regressions to anaphor), rather than the *duration* of the preliminary assignment process. However, the time course data may also provide information about the latter. In particular, the time spent reading the noun when it is first encountered (i.e., first-pass fixations) may capture the duration of preliminary referent assignment (Just & Carpenter, 1987). What duration patterns might we expect within and across the special-purpose and memory-based accounts?

The special-purpose search account, has an inherent anaphoric bias, which implies that readers will have to think long (first-pass fixation measure, Experiments 2 and 4) and hard (difficulty measure, Experiment 3) before resorting to a new-referent interpretation of a definite noun. In particular, a reader must conduct a prolonged (possibly exhaustive) discourse search, to reject all discourse candidates, before resorting to a new referent interpretation.

In contrast, the memory-based account does not involve an inherent anaphoric bias. Because the spread of activation is rapid, and all referents in memory (discourse referents and generic referents) compete in parallel for retrieval, the duration of preliminary referent assignment may be relatively invariant across definite nouns (literal antecedent-match anaphors, literal category anaphors, metaphoric category anaphors,

non-anaphors). Some referent, though not necessarily the right referent, should rapidly reach retrieval threshold. Consequently, non-anaphoric nouns (and anaphoric nouns misinterpreted as non-anaphoric) may not require significantly more time to process than anaphoric nouns.

In summary, in accord with my unified memory-based account, I expect that reducing the semantic similarity between the anaphor and antecedent should primarily affect the accuracy rather than the duration of preliminary referent retrieval -- that is, the effect will be to increase the incidence of misclassify-as-new errors. This section provides the groundwork for understanding some of my rationale behind my choice of experiments, and my inclusion of non-anaphors in the stimuli for comparison purposes about duration. Below I will provide a brief description for each experiment, and discuss its role for testing the hypotheses outlined above, and for producing data for comparison with the model.

Overview of Experiments

Table 2.1 provides an overview of the four experiments and their key dependent measures, which provide information on the accuracy and/or duration of preliminary referent assignment. After the table I provide a brief description for each experiment, and discuss its role for testing the hypotheses outlined above, and for producing data for comparison with the model.

Table 2.1: Overview of experiments indicating the paradigm, independent variable manipulations, and the key dependent measures which provide information on the accuracy and/or duration of preliminary referent assignment.

Study (Chapter)	Paradigm	Manipulations	Evidence on Accuracy of Preliminary Assignment	Evidence on Duration of Preliminary Assignment
1 (3)	Cloze Task	Noun Type (3): <ul style="list-style-type: none"> • literal anaphor (doctor) • metaphoric anaphor (butcher) • non-anaphor (nurse) 	% Misclassify-as-new errors	
2 (4)	Eye-tracked Reading	Noun Type (3): <ul style="list-style-type: none"> • literal anaphor (doctor) • metaphoric anaphor (butcher) • non-anaphor (nurse) 	% Regressions-to-anaphor Remainder-of-sentence fixations Subsequent noun fixations	First-pass noun fixations
3 (5)	Cloze Task	Noun Type (3): <ul style="list-style-type: none"> • literal anaphor (doctor) • metaphoric anaphor (butcher) • non-anaphor Determiner (2): the, that	% Misclassify-as-new errors	Difficulty ratings
4 (6)	Eye-tracked Reading	Noun Type (4): <ul style="list-style-type: none"> • antecedent-match anaphor (surgeon) • literal anaphor (doctor) • metaphoric anaphor (butcher) • non-anaphor Colour-marking noun's anaphoric status (2): <ul style="list-style-type: none"> • Block 1 (unmarked) • Block 2 (marked) 	% Regressions-to-anaphor Remainder-of-sentence fixations Subsequent noun fixations	First-pass noun fixations (Noun response time)

In Experiment 1, a cloze paradigm is applied to determine the likelihood that a reader will form a new-referent interpretation for each literal category anaphors (e.g., "doctor", with high semantic overlap with antecedent) versus each metaphoric category anaphor (e.g., "butcher", which have low semantic overlap with the antecedent). In contrast, in accord with my unified memory-based hypothesis, I expect that the incidence of misclassify-as-new errors will be higher for the metaphoric anaphors than for the literal ones (aggregate pattern), but importantly these data (i.e., % incidence of misclassify-as-new errors) also provide insight into the likelihood of misinterpretation for each individual item within each of the conditions. However, these data cannot provide any information on the duration/difficulty of preliminary assignment for the different items.

In Experiment 2, an eye-tracked reading paradigm is used to determine the likelihood that readers will regress to the anaphor for literal ("doctor") and metaphoric ("butcher") category anaphors. These data (% regressions to the each anaphor) also provide quantitative estimates of the relative incidence of misclassify-as-new errors for each anaphoric item, and provide the basis for comparing the accuracy of human performance with that of the memory-based computational model, MEMBRAL. Secondly, Experiment 2 also provides evidence about the duration of preliminary referent assignment (e.g., first-pass noun fixations). Evidence of longer preliminary processing times for non-anaphors (and anaphors misinterpreted as new) over anaphors is compatible with a special-purpose account, whereas the memory-based is compatible with the possibility that preliminary time for referent retrieval may be relatively

invariant across the critical nouns.

Experiment 3, like Experiment 1, involved a cloze paradigm, however, in addition to providing completions of the sentence (which reveal the accuracy of their preliminary assignments), participants in Experiment 3 were also asked to rate difficulty of the task on each trial – which should provide information about the difficulty (and indirectly the duration) of preliminary assignment. That is, do readers have to think long and hard before associating a new referent to a definite noun phrase, as would be predicted by the special-purpose account, with its anaphoric bias (thus the relevant comparison here is between ease of processing for anaphoric versus non-anaphoric nouns, though the latter are rarely included in time course studies on anaphora). This experiment also tests the repeatability of the accuracy patterns obtained in Experiment 1. Finally, this experiment also involved a manipulation of the determiner preceding the anaphors ("the" vs. "that"), however I will defer a discussion about this manipulation until the discussion of this experiment (Chapter 5).

Experiment 4, like Experiment 2, involved an eye-tracked reading paradigm, however Experiment 4 included an additional level of semantic overlap (antecedent-match anaphors, in addition to literal and metaphoric category anaphors and non-anaphors). This experiment therefore provides additional information about the relative likelihood of misclassify-as-new errors (% regressions to the anaphor) for all three degrees of semantic overlap (and for the different items within these levels), which will be compared against the model performance. An additional manipulation was also used in this experiment. In particular, in the second block of trials, the anaphoric status of

the critical nouns was explicitly indicated by font colour (e.g., purple =new referent, orange= anaphor). More details about this manipulation will be presented in the chapter detailing this experiment (Chapter 6).

In summary, cloze and eye-tracked reading paradigms are used in the present research to investigate the accuracy (and, secondarily, the duration) of preliminary referent assignment, and this data provides a basis for comparison with the model. Table 2.1 summarized the key characteristics (manipulations, paradigm, measures) of the four experiments, and clarified the role of each experiment in terms of what evidence it provides about the accuracy and/or duration of the preliminary referent assignment process. To avoid redundancy in the descriptions of the individual experiments, an overview of the stimulus materials will be presented below.

Stimulus Materials

For Experiments 1, 2 and 3, thirty-six stimuli stories were constructed. Each story was four sentences in length, and consisted of three context sentences followed by the final critical sentence, which commenced with a definite noun phrase. For each story there were three versions of the final critical sentence that differed only in the nature of the critical noun, which was either intended as a literal category anaphor, a metaphoric category anaphor or an introductory noun. For example:

(S1) Lisa went in for an operation on her knee.

(S2) An incompetent surgeon performed the procedure.

(S3) Afterwards, Lisa's knee was worse than ever.

(S4) The doctor/butcher/nurse didn't care.

Many of the anaphors were inspired by stimuli used in previous research (e.g., Gibbs, 1990; Onishi & Murphy, 1993; Stewart & Heredia, 2002), which should facilitate a comparison with their results. However, prior studies did not always control for the determiner (e.g., “The”) and the sentence-initial position of the anaphor. For the present research, in anaphoric trials, the noun was intended to refer to a target introduced in the first or second context sentence (e.g., the incompetent surgeon, in the context above). Whenever possible, each critical sentence was designed to be incongruous or incomprehensible as a whole if the anaphors were initially misinterpreted as new referents.

The literal/metaphoric anaphor pairs (e.g., “doctor”/”butcher”, “coward”/”chicken”) were matched as closely as possible for length (literal mean: 6.2 characters, metaphoric mean: 6.2 characters) and frequency (Kucera & Francis, 1967). However, they differed in their degree of semantic overlap with the antecedent, as gauged using LSA. Because the description of a particular referent may be distributed throughout the context rather than concentrated in a single antecedent term, an estimate of anaphor-referent semantic overlap can be computed by finding the LSA similarity between the anaphor and the preceding context as a whole (Budiou & Anderson, 2004). However, such a method gives a single measure of the overlap between the anaphor and the context, and does not differentiate among the degrees of overlap between the anaphor and different referents within the context. Thus, to get a measure of the degree of overlap between the anaphor and the intended antecedent, in particular, the LSA similarity was computed between the anaphor and the set of descriptive terms used for the intended referent in

the context (see also Lemaire & Bianco, 2003). The mean anaphor-referent LSA similarity was 0.39 for the literal anaphors, and 0.15 for the metaphoric anaphors (computed using: Latent Semantic Analysis at the University of Colorado at Boulder, 2003).

In Experiment 3, the determiner at the start of the critical sentence was also manipulated on the anaphoric trials ("The" vs. "That"). And, in Experiment 4, an additional type of (literal) anaphor was included which was simply the same noun used for the antecedent (an antecedent-match anaphor). Thus, there were three possible noun anaphors for the present example: "surgeon" (literal antecedent-match anaphor); "doctor" (literal category anaphor); and "butcher" (metaphoric category anaphor). Experiment 4 included an additional noun type (antecedent-match anaphor), as well as a manipulation of status marking. This 4(noun) by 2(block) design motivated the addition of six new contexts (each with three anaphors) were also added to the stimulus set, for a total of 42, to supplement the number of items per condition. When necessary to maintain a balance a between stories with anaphoric and non-anaphoric critical sentences, additional stories with non-anaphoric final sentences were included among the stimuli in each experiment. A complete list of stimulus materials is provided in Appendix A.

Summary

Because this research was conducted in the context of a cognitive science degree, it seems appropriate to take a moment to summarize why the above approach is appropriately characterized as cognitive science. The issue of anaphor resolution has

its origins in linguistics, however, in that discipline, anaphora amounts to a relation between an anaphor term and an antecedent term, which are linguistic, non-cognitive constructs. In the current cognitive formulation, the reader is making an association between a noun and referent, which is a mental representation with temporal activation fluctuations and properties beyond those captured in the single, textual antecedent term. Further, the object of investigation is the cognitive *process* of preliminary referent retrieval for nouns. As such, a central objective of this research is to produce a detailed description of the process itself, in the form of a working computational model. Human experimentation contributes to the development of a process description by providing evidence about the accuracy and duration of preliminary referent retrieval. Thus, both the issue of investigation and the current approach are inherently interdisciplinary (linguistics, psychology, computational modeling).

The following four chapters will present the results of my behavioural experiments. A discussion of the computational model is postponed until Chapter 7, when these human performance data are available for comparison purposes.

CHAPTER 3: Experiment 1

The time course for reading does not provide direct information about the nature of the referent initially (or ultimately) assigned to the critical nouns. Thus, the objective of this cloze experiment was to obtain independent evidence about the nature of the preliminary referent assignment made by readers upon encountering the critical nouns (literal category anaphors, metaphoric category anaphors, non-anaphors). As discussed in Chapter 2, the critical noun phrases occur at the beginning of the final sentence in each stimulus story, and, in the present experiment, the remainder of the sentence was omitted and the participant was required to create his or her own ending for the sentence. The sentence ending produced by the participant revealed whether the participant associated the critical noun to a discourse referent or interpreted it as evoking a new referent. For anaphorically-intended nouns, a new referent interpretation constitutes a misclassify-as-new error.

Data about the incidence of misclassify-as-new errors will provide valuable insight for the interpretation and prediction of time course patterns. If such misclassify-as-new errors are frequent in the metaphoric case, post-noun reanalysis processes should be the primary contributor of the net processing delays in sentences with metaphoric versus literal anaphors. However, if such preliminary errors are rare, such evidence will suggest that delays in the metaphoric case occur because preliminary referent assignment is prolonged, but effective (as assumed by Janus & Bever, 1985; cf. Budiu & Anderson, 2002).

Although information about the incidence of misclassify-as-new errors is not, by

itself, sufficient to conclusively arbitrate between the memory-based and special-purpose accounts of preliminary referent assignment, highly accurate performance on the metaphoric trials would lend support to the special-purpose search account. A special-purpose account involves an anaphoric bias and gives preferential consideration to referent candidates within the discourse. Thus, identifying the intended referent effectively becomes a matter of recognition rather than recall (the latter is more difficult and characterizes the process in the memory-based account). Consequently, these properties of the special-purpose account could largely moderate the impact (on accuracy though perhaps not duration) of the reduction in anaphor-antecedent semantic overlap for metaphoric versus literal anaphors.

In contrast, the accuracy of a memory-based system to recall the intended referent would be directly affected by reductions in anaphor-antecedent semantic overlap. Degree of overlap influences the amount of activation that will automatically spread to the intended referent upon reading the anaphor. Thus, the low semantic overlap in the metaphoric cases reduces the net activation achieved by the intended referent, and therefore reduces its likelihood of retrieval.

Note that under the memory-based account, any decrease in preliminary assignment accuracy for the metaphoric case is not the result of metaphoricity per se, but arises because metaphoricity is correlated with low anaphor-antecedent semantic overlap which, in turn, is correlated with the amount of activation that will automatically spread to the intended referent as a result of reading the anaphor. The degree of anaphor-antecedent overlap can similarly influence accuracy for literal anaphors, as supported by

the results of Levine et al. (2000). In general, literal and metaphoric anaphors alike may sometimes be initially processed as new referents, however a greater proportion of misclassify-as-new errors are expected in the metaphoric cases because they tend to have lower antecedent-anaphor overlap. Consequently I expected that the incidence of misclassify-as-new errors in this cloze task would be predicted by anaphor-antecedent semantic overlap, as quantified using LSA.

In the present research, I focus on models and explanations that provide a unified treatment applicable to both metaphoric and literal anaphors. This position was influenced by the principle of parsimony, and by the view that common processing mechanisms can support the processing of literal and metaphoric content (e.g., Frisson & Pickering, 2001; Giora, 1997; 2002, Glucksberg, 2003, Kintsch 1998; 2000; McElree & Nordlie, 1999). However, there are models particular to metaphor processing that can also provide accuracy predictions for comparison purposes. For example, the classical view of metaphor comprehension also suggests that, regardless of prior context, readers are always initially committed to a very narrow ('literal') lexical interpretation, which, in the metaphoric trials, would preclude making any connection to the intended referent. As such, the classical view of metaphor comprehension would predict that all metaphoric anaphors should be treated as new referents during preliminary analysis.

Method

Participants. The participants were eighteen Carleton University students (12 female and 6 male). All participants were native speakers of English. Participants

received either course credit or \$10 for their participation.

Materials. This experiment employed the materials described in Chapter 2 (and listed in Appendix A). For this cloze task, the endings of the fourth sentence (i.e., the words after the critical noun) were not presented to the participants. For example:

Lisa went in for an operation on her knee.

An incompetent surgeon performed the procedure.

Afterwards, Lisa's knee was worse than ever.

The butcher _____.

Three sets of materials were constructed such that each set included 12 stories with the metaphoric version of their final sentences (e.g., "butcher"), 12 different stories with the literal version of their respective final sentences (e.g., "doctor"), and 12 different stories with the non-anaphoric version of the critical sentence (e.g., "nurse"). The versions of the stories were counterbalanced across the sets. Each set also included an additional 12 texts (including three practice texts) with non-anaphoric nouns in the critical position, to provide an equal number of anaphoric and non-anaphoric final sentences. Thus, each set contained 48 texts in total. A random order was generated for the stories, and this same random order was used in each set.

Procedure. Each participant was randomly assigned to one of the three sets of materials, and received the corresponding booklet containing 48 incomplete texts. Participants were instructed to write a coherent ending for the last sentence in each story. No explicit time limit was given, and the experimental session typically took less than 50 minutes.

Results

For all inferential statistics in these experiments, separate analyses were done with participants and items as random variables, and these are reported as F_1 and F_2 , respectively. Effects were significant at $p < .05$, unless otherwise specified.

To assess whether a critical noun had been interpreted as an anaphor or a new referent, two independent coders categorized the participants' use of the noun. The following examples illustrate the coding scheme:

Context: Lisa went in for an operation on her knee.
An incompetent surgeon performed the procedure.
Afterwards, Lisa's knee was worse than ever.

Completions:

The doctor <u>offered to re-do the operation.</u>	[anaphor]
The doctor <u>who referred Lisa to that surgeon felt bad.</u>	[new referent]
The butcher <u>would get his medical degree revoked.</u>	[anaphor]
The butcher <u>down the street brought meat for the invalid.</u>	[new referent]
The nurse <u>gave Lisa some painkillers.</u>	[new-referent]

There was strong inter-coder agreement (above 97%), and the classifications of all questionable cases were resolved in discussion to the satisfaction of both coders. Note, each individual participant saw only one version of each context.

When the critical noun was intended as introductory (e.g., nurse), participants always correctly formed a new referent interpretation (i.e., there were no misclassify-as-

anaphor errors). However, when the critical noun was intended as anaphoric, participants treated it as anaphoric on only 73% of the trials. In particular, misclassify-as-new errors were made in 5% of the literal trials and 48% of the metaphoric trials, $F_1(1,17) = 87.94$, $MSE = 1.93$, $F_2(1,33) = 65.57$, $MSE = 3.50$ (due to a reproduction error in the booklets of stimuli, two of the items were not presented in their metaphoric version so the item analysis was conducted over 34 items). There were only two metaphoric anaphors that none of the participants recognized as anaphoric. That is, for all but two of the thirty-four metaphoric items, one or more participant(s) succeeded in making the intended anaphoric association after having read the noun.

LSA Correlation. Across the full set of literal and metaphoric anaphors, the expected positive correlation was found between referent assignment accuracy and LSA estimates of anaphor-referent semantic overlap, $r(70) = 0.46$, $p < .05$ (1-tailed). However, a partial correlation (controlling for noun type: metaphor category anaphor vs. literal category anaphor) indicated that LSA estimates could not further capture fine-grained performance variation within each category, $r(62) = -.02$, $p = .427$.

Discussion

This cloze task provides estimates of the likelihood that an anaphor will be initially misclassified as a new referent. Without the benefit of access to the original sentence endings, each participant made misclassify-as-new errors for about half of the metaphoric anaphors, and 5% of the literal anaphors. Because even literal anaphors were sometimes treated as new referents, these findings are compatible with those of Levine et al. (2000), and confirm that the initial misclassification of anaphoric nouns as

new referents is not idiosyncratic to metaphoric anaphors. Instead, correct assignment seems to depend on underlying factors like semantic overlap. In particular, the readers' success at recognizing the anaphoric status of the noun was significantly correlated with LSA estimates of anaphor-referent conceptual overlap. As such, the results are generally compatible with a memory-based account, in which the retrieval of the intended referent rather than a generic referent is contingent on the amount of activation the intended referent receives in virtue of anaphor-antecedent semantic overlap.

That said, the accuracy data are also arguably compatible with the special-purpose search account. In particular, the intended referent may be rejected during the discourse search in virtue of low anaphor-antecedent semantic overlap. However, the relatively high frequency of error for the metaphoric trials (about 50%) seems high for a system with an anaphoric bias that only requires the recognition (vs. recall) of the intended referent within the discourse. Onishi and Murphy (1993) established that metaphoric associations are readily recognized when the topic (surgeon) and the vehicle (butcher) are simultaneously considered, as they are when the metaphor is presented in predicative form (e.g., The surgeon was a butcher). In the case of anaphoric metaphors, a discourse search that checks the anaphor (butcher) against possible discourse referents, including the intended one (the incompetent surgeon) should also provide an opportunity for the simultaneous consideration of the two metaphor components, which should have enabled readers to recognize almost all of the intended associations.

An additional reservation about the special-purpose account concerns its ability to explain the misclassify-as-new errors that occurred on literal trials. For example, in a

story about a young male pupil, the anaphor "the boy" was sometime treated as a new referent. Under the special-purpose search account, resorting to a new referent interpretation implies that the system effectively rejected the intended discourse referent, [R: young male pupil], as a compatible candidate for the anaphor ("the boy") during the search. Such behaviour seems implausible, given that the primary advantage and justification for a special-purpose system would be its ability to avoid missing such obvious connections, especially within short four-sentence texts. In contrast, such errors are less surprising under a memory-based account because discourse referents are not given preferential consideration, instead spread of activation leads can lead to the recall of a the generic referent, in this case [G:boy], which competes for retrieval with the intended referent from the discourse [R1: the young male pupil].

In all, the present data suggest that the ability to retrieve the intended referent during *preliminary* referent assignment can be explained under a memory-based account. Thus, the memory-based account, which is steadily acquiring an explanatory track record in studies on (literal) noun anaphora (e.g., Gernsbacher, 1989; Myers et al., 1998; O'Brien, Albrecht, Hakala, & Rizzella, 1995; O'Brien et al., 1997), is also suited to account for the results of the current study containing metaphoric anaphors.

Furthermore, the memory-based framework seems better able account for the data than several models particular to metaphor processing. In particular, the variability in performance (misclassify-as-new errors) across metaphoric items is difficult to explain under the classical view of metaphor comprehension. Because the participants were free to create their own ending for each sentence, there was no direct pressure for them

to abandon a strictly ‘literal’ (and consequently new-referent) interpretation, had that always been their strong immediate inclination. In contrast to the memory-based account, the classical view of metaphor comprehension sheds little light on why metaphoric anaphors were sometimes correctly comprehended and sometimes not in this task.

An alternative view of metaphor processing allows that some common metaphoric terms may have their own distinct entries in the reader’s mental lexicon (e.g., Giora, 1997; 2002; Glucksberg, 2003), so that, for example, “chicken” could be ‘literally’ interpreted as either a bird or a coward. If so, an alternative possibility is that conventionality or familiarity of the metaphoric usage (rather than the underlying semantic overlap) determines cloze performance. To investigate this possibility, I conducted a post hoc norming study with 18 new participants (mean age: 20.2 years), which included a by-item poll as to whether the relevant metaphoric sense (e.g., coward) was judged to be one of the established meanings of the noun (e.g., chicken). These conventionality judgments from the norming group were not, however, predictive of the incidence of misclassify-as-new error for the metaphoric anaphors in this cloze study, $r(34) = .075$, $p = .337$. In particular, the norming group judged, on average, that only 16% of the metaphoric nouns had the appropriate metaphoric sense pre-established in their lexicon. Thus, although conventionality may sometimes play a role in metaphor comprehension, such low numbers suggest that the 52% success rate for the comprehension of the metaphoric anaphors in the present study did not seem to stem from, or hinge on, the intended metaphoric sense being a distinct, pre-established sense

in the reader's lexicon.

In a memory-based model, the activation of a referent is related to its semantic overlap with the anaphor cue, so the novelty of a particular metaphoric usage for a particular participant does not preclude the possibility that the referent may receive sufficient activation for retrieval during preliminary assignment. That said, repeated exposure to a metaphoric usage may have the effect of strengthening the conceptual associations in long-term memory, and this would, in turn, improve performance (Budiu & Anderson, 2006). Thus, in general, an influence of conventionality/familiarity on preliminary referent retrieval is entirely compatible with the memory-based account. However, the conventionality ratings of the norm group were not significant predictors of the results in this study.

Interestingly, a failure to associate the intended referent to a metaphoric anaphor did not always imply a failure to make a metaphoric interpretation of the noun. There were several instances in which participants produced metaphoric, but nonetheless non-anaphoric interpretations for a noun. For example, in a story about people walking along a set of train tracks, the noun "ribbons" was intended as a metaphoric anaphor to denote the tracks (adapted from Gerrig & Healy, 1983). Some participants who failed to make the intended anaphoric connection (i.e., to the tracks), nonetheless produced metaphoric interpretations of the noun (e.g., "ribbons of clouds", "ribbons of sunlight") that were non-anaphoric. Such examples cannot be explained in a framework that necessarily attributes preliminary resolution failure for the metaphoric anaphors to a strictly literal interpretation of the noun.

However, such examples can be explained under a memory-based model. The LSA overlap estimates between ribbons and train tracks and ribbons and sunlight are comparable (0.20 vs. 0.19 respectively), so some individuals may presumably have a stronger conceptual association between ribbons and sunlight than between ribbons and train tracks. Memory-based activation spread involves no delineation between literal versus metaphoric associations between concepts, nor between associations to concepts inside versus outside the discourse. Thus, memory mechanisms could enable the anaphor "ribbons" to bring to mind associations, including 'metaphoric' ones, with concepts that are external to the discourse (e.g., sunlight). The resulting activation may sometimes be stronger than the activation produced by conceptual overlap with the intended referent [R: train tracks].

Overall, the present results are consistent with the suggestion that initial success at assigning the intended referent, rather than a new referent, are grounded in basic memory-based factors like the degree of overlap semantic overlap between the concept associated with the anaphoric noun and the antecedent. In particular, metaphoric anaphors, which have a low antecedent-anaphor conceptual overlap, are more likely than literal anaphors to be initially assigned a new referent. Such preliminary misclassifications imply that in a regular reading task (i.e., reading the original, complete versions of the anaphoric sentences), the reader would either misunderstand the sentence, or would later reanalyze their referent assignment based on subsequent information in the sentence and more global pragmatic considerations.

Consequently, in Experiment 2 (eye-tracked reading), I sought converging time

course evidence that there is a higher incidence of revision when participants read sentences with metaphoric anaphors (i.e., regressions to re-read the anaphor and post-anaphor processing delays). Additionally, although the present results suggest that the *accuracy* of preliminary referent assignment is impacted by the literal/metaphoric manipulation of semantic overlap, they do not provide evidence about the *time course* of the referent assignment process in regular reading. Preliminary reference assignment may also take longer in the metaphoric cases and thus anaphor difficulty might influence *both* accuracy and duration of reference assignment. Eye-tracking data in Experiment 2 were used to test this hypothesis.

CHAPTER 4: Experiment 2

In this experiment, I investigated the time course and accuracy of comprehension for complete versions of the anaphoric sentences using an eye-tracked reading paradigm. The stories were presented one sentence at a time on a computer monitor, and participants signaled their comprehension by pressing the space bar. In keeping with previous data about metaphoric anaphors in particular (e.g., Gibbs, 1990; Noveck et al., 2001; Onishi & Murphy, 1993), and about anaphors with low semantic overlap with the antecedent in general (e.g., Garnham, 1984; 1989; Garrod & Sanford, 1977), a longer sentence response time was expected for the sentences with metaphoric anaphors than for the sentences with literal anaphors. Thus, one goal of this experiment was to determine if I could replicate the effect of longer net reading times for the metaphoric versions of the anaphoric sentences with these stimuli.

The primary goal for this experiment, however, was to explore the distribution of the net processing time *within* the anaphoric sentences, to determine where the extra processing time was allocated in sentences with metaphoric anaphors. Although Frisson and Pickering (1999) collected eye-tracking data for non-anaphoric metaphors and metonyms, to my knowledge, no eye-tracking study has been previously reported that compares metaphoric and literal anaphora. In particular, I wanted to determine whether the fixation patterns would be either: a) compatible with the response time patterns from the word-by-word presentation studies (Budiou & Anderson, 2002; Janus & Bever, 1985) which predict that first-pass noun fixations will be the primary locus of delay; or b) compatible with eye-tracking data from other studies involving antecedent

typicality manipulations of semantic overlap (e.g., Duffy et al., 1990; van Gompel et al., 2005), in which post-noun fixations (and regressions) were the primary locus of the delay.

In light of the findings in Experiment 1, it seems unlikely that first-pass reading time on the noun would be the locus of delay in the metaphoric case, unless readers rarely bother to rectify erroneous initial referent assignments in light of subsequent information in the sentence. To encourage readers to correct any erroneous preliminary assignments, most of the anaphoric sentences were designed to be incongruous under a new-referent interpretation. A comprehension question was presented after each critical sentence to assess the accuracy of a participant's final referent assignment for each critical noun (i.e., after the opportunity for revision). Thus, I expected that readers would usually arrive at an accurate assignment after processing the whole sentence, and that most of the comprehension questions would be answered correctly. Hence, there should be evidence of reanalysis in the time course data.

In particular, two observable time course effects will provide evidence of reanalysis: processing delays in the remainder of the sentence that follows the critical noun (i.e., 'covert reanalysis'), and a greater incidence of regressions to the anaphoric noun (i.e., 'overt reanalysis'). Eye regressions to the site in the sentence at which the interpretation took a 'wrong turn' are common during the comprehension of sentences with temporary syntactic ambiguities (e.g., Meseguer, Carreiras, & Clifton, 2002; Altmann, Garnham & Dennis, 1992), and such behaviour generalizes to referential reanalysis. In particular, Garrod, Freudenthal and Boyle (1994) manipulated the verb

following a noun anaphor (e.g., “The stewardess poured/ordered a drink”), so that in the latter, incongruent case, the verb could induce the reader to consider revising their assignment to a different referent in the discourse (i.e., the passenger). In the incongruent verb case (i.e., when subsequent information in the sentence called the referent assignment into question), they found a significantly higher incidence of regression to the anaphoric noun. Thus, I expected a higher incidence of regressions to the anaphor on the metaphoric trials in the present experiment.

Besides time course effects due to reanalysis, there may also be differences in the time course owing to the preliminary referent assignment process. That is, decreasing semantic overlap might influence both the accuracy and duration of initial referent retrieval. Under the eye-mind assumption (Just & Carpenter, 1980), a difference in the duration of initial referent assignment should be evidenced by a difference in the first-pass fixations on the noun.

Method

Participants. Twenty-four Carleton University students (10 male and 14 female) participated in this study. Participants were given \$10 or course credit for their participation. All participants were native speakers of English, and had normal or corrected to normal vision. Sentence response time data were collected for all participants. Eye-tracking data were collected for a subset of 18 participants using a camera mounted under the monitor. The eye-tracking equipment could not determine gaze location for the other six participants due to astigmatism of the eye or glare from eyeglasses.

Materials and Design. This experiment used the same materials as Experiment 1, with the adjustment that the final sentences of the texts were presented in their entirety (Appendix A). For each story, a TRUE/FALSE comprehension question was also included at the end of the story. These questions were intended to ensure the readers' involvement in comprehending the stories, and to assess the accuracy of the reader's final (vs. preliminary) interpretations of the intended anaphors, after an opportunity for reanalysis. As distinct from the designs of Gibbs (1990) and Onishi and Murphy (1993), in this study the questions for the critical anaphoric trials were designed to assess whether or not the correct anaphoric link had been made. For the example story in Experiment 1, the comprehension question was: "The surgeon who performed the operation felt regret." (answer = FALSE). To avoid biasing the readers to concentrate on content only in the last sentence of the stories, the questions for the non-anaphoric trials related to information found within the first three sentences of the stories. Counterbalancing was done to ensure that an equal number of true and false questions were present in each set of materials and within each condition.

Procedure. The presentation of materials was controlled by a Pentium III desktop computer. The participants were seated approximately 50 cm from the computer monitor. Text was displayed on the 17-inch computer monitor in white, Courier, 30-pixel font against a black background. Participants arranged the keyboard and mouse to enable them to comfortably rest their left hands on the space bar and their right hands on the mouse. Each story was displayed one sentence at a time (left justified and halfway down the screen). All sentences were fewer than 55 characters in order to fit

within a single line on the screen. Prior to the presentation of each sentence, a red fixation cross was displayed briefly (200 ms) where the first character of the sentence would appear, and participants were instructed to look at the cross to prepare for the sentence. For the eye-tracked participants, the fixation cross was displayed until the participant fixated on it.

Participants were instructed to press the space key, with their left hand, as soon as they had understood each sentence. The sentence was then replaced by the fixation cross in preparation for displaying the next sentence. After the fourth and final sentence in each story, a message appeared on the screen notifying the participant that the (T/F) comprehension question would appear next, and reminding them to use the mouse to respond. To further distinguish the questions from the sentences, the questions were presented in orange text. Participants were instructed to press the right mouse button to respond TRUE and the left mouse button to respond FALSE. As a mnemonic device, the letters 'T' and 'F' were affixed, respectively, to the left and right upper corners of the monitor. After the response to the question, the question was replaced by a message instructing the participant to press the space bar for the next story.

For the eye-tracked participants, the monitor-mounted camera for the EyeGaze™ eye-tracking system was trained on the participant's right eye. A chin rest was used to minimize head movement, and at the outset of the experiment, a one-minute calibration procedure was conducted during which the participant was asked to fixate, in turn, on ten yellow dots that appeared sequentially in various locations on the screen. The EyeGaze™ system was interfaced with the Pentium III computer that ran the

experiment. The system calculated and recorded the location of the participant's gaze (screen pixel coordinates) throughout the experiment. The system had a 60 Hz sampling rate, and used the pupil-centre/corneal-reflection method to deduce the position of the eye.

In a few cases, participants had to re-do the calibration procedure partway through the experiment due to tracking failure (e.g., due to significant head movement). Prior to the presentation of each sentence, participants were expected to fixate on a red cross, and the software performed real-time analysis of the incoming eye data to verify this cross fixation. This served as an on-going check throughout the experiment that calibration hadn't been compromised. If the system could not detect the fixation on the cross, an error message was shown instead of the sentence. The experimenter could press a key to re-display the cross. If the error persisted, the one-minute calibration procedure was conducted again, and the experiment resumed with the next trial. The typical duration of the experimental session was 45 minutes.

Results

Reading time on the final target sentence was measured from the initial presentation of the sentence until the participant pressed the space bar. Two trials in which a participant mistakenly pressed the mouse (instead of the spacebar) were excluded from the analyses. As in Experiment 1, separate analyses were done with participants and items as random variables (F_1 and F_2 , respectively). Reported results were significant at $p < .05$ unless otherwise indicated.

Comprehension accuracy. For anaphoric trials, the correct response to each

question required the participant to have correctly associated the critical noun anaphor with the intended referent in the context. Overall, participants responded correctly to 89% of these comprehension questions. This finding suggests that in the majority of trials, the readers had successfully associated the correct referent to the noun anaphors by stories' end (which may sometimes have required a reanalysis of their preliminary referent assignment). Performance on comprehension questions was slightly lower for trials with metaphoric anaphors (86%) than for trials with literal category anaphors (93%), though this difference did not reach significance in the analysis by items, $F_1(1,23) = 4.75$, $MSE = 466.9$, $F_2(1,35) = 1.53$, $MSE = 516.6$, $p = 0.224$. These findings suggest that metaphoric anaphors have a greater likelihood of being misclassified as introductory during preliminary analysis, and that such misunderstanding can sometimes persist in spite of cues in the remainder of the sentence (see also Budiu & Anderson, 2002). Nonetheless, question performance indicates that the majority of metaphoric anaphors were correctly understood by the time the readers had formed their final interpretation (which may have involved reanalysis). Such sentence-end performance rules out the possibility that the metaphoric anaphors in the present materials were inherently incomprehensible to this participant population.

In contrast to the anaphoric trials, in the non-anaphor trials the critical noun introduced a new referent, so the comprehension question could not probe whether participants had made a correct co-referential interpretation. Instead, questions on non-anaphoric trials were about content in the context. Participants answered 86% of these questions correctly, which was not pairwise different from their performance in the two

anaphoric conditions in 3-level (noun type: literal, metaphoric, non-anaphor) ANOVAs ($ps > 0.05$). Note, this measure is not indicative of their comprehension of the critical noun itself. Recall that in Experiment 1, all of these non-anaphoric nouns were appropriately assigned new-referent interpretations (i.e., no misclassify-as-anaphor errors), it was only on the anaphoric trials that participants made errors in preliminary assignment.

Sentence response time. Trials in which the participant responded incorrectly to the comprehension question were excluded from the analysis of sentence response times. A 3-level (noun type: literal anaphor, metaphoric anaphor, non-anaphor) repeated measures ANOVA was conducted by both items and participants. Response times differed depending on the critical noun, $F_1(2,46) = 41.56$, $MSE = 2,130,780$, $F_2(2,70) = 46.39$, $MSE = 4,558,676$. Pairwise Bonferroni comparisons revealed that response times were comparable for the non-anaphoric and literal anaphoric versions of the critical sentences (1940 ms vs. 1832 ms; $ps > .05$), which is consistent with data about non-anaphoric sentences reported by Gibbs (1990). However, among anaphoric sentences, mean response times were significantly longer on the metaphoric trials (2157 ms) than on the literal trials (1723 ms; $p_1 = .027$, $p_2 = .016$). This response time difference was also significant within the subset of eye-tracked participants (2120 ms vs. 1722 ms, $ps > .05$).

As illustrated in Figure 4.1, this pattern of response time differences is congruent with that in several prior studies. These prior studies, however, did not involve eye-tracking, and consequently did not provide data on the time course of processing within

the sentences. Sentence response times are of limited utility in investigating the duration of preliminary referent assignment itself, as distinct from subsequent reanalysis processes. Furthermore, with the exception of Gibbs (1990), these prior studies omitted the contrast with non-anaphoric definite nouns, which might assist in arbitrating between the special-purpose and memory-based accounts. In particular, under a special-purpose account, a reader must first reject the referent candidates within the discourse before resorting to a new referent interpretation for a definite noun. Thus, preliminary referent assignment would be expected to take longer for non-anaphors and metaphoric anaphors interpreted as new-referents than for anaphors. The time course data from the current study are presented in the next section.

Figure 4.1: Sentence Response Times (ms) for sentences containing literal category anaphors versus metaphoric anaphors in the present research (Pyke) and prior research (Gibbs, 1990; Lemaire & Bianco, 2003; Noveck et al., 2001; Onishi & Murphy, 1993; Ortony et al., 1978).

Note. The literal vs. metaphoric difference in response times was significant in each study except Ortony et al., 1978.

Fixation durations. Net fixation durations were computed for the sentence as a whole and for different regions within the sentence (article, noun anaphor, remainder of sentence) to determine the degree to which reading, or re-reading, a particular region (e.g., the noun) contributed to the extra time spent on metaphoric sentences. An eye location sample was assigned to an ongoing fixation if the sample's coordinates were within a half of a character width (9 pixels) of the current center of the fixation. Otherwise the sample was assigned to a saccade. Only fixations of at least 65 ms in length were considered valid. Net fixations for a region were computed as the sum of the fixations occurring in that region. Net fixation duration for each word was further partitioned into two components: (a) First-Pass Fixations (sometimes referred to as "gaze duration"), which is the sum of the fixations starting from when the reader first begins reading that region and ending when the reader's eye first saccades out of that region (frequently to the right to read the subsequent word in the sentence). Finally, subsequent fixations in a region are computed by subtracting first pass fixations from net fixations. Of particular interest are first-pass and subsequent fixations on the critical noun itself. Subsequent fixations result if and when a reader returned to re-read the critical noun while processing the remainder of the sentence. I also computed the proportion of trials in which such regressions to the critical noun occurred, in each condition.

Figure 4.2 provides detailed breakdown of the first-pass and net fixations for each the word position in the critical sentence. However, Figure 4.3 summarizes the key time course measures that are particularly relevant for providing evidence of: i) preliminary

error (i.e., evidence of reanalysis: subsequent noun fixations, net remainder fixations), and ii) duration of preliminary assignment (i.e., first-pass noun fixations). As evident from Figure 4.3, the pattern of longer *response times* for sentences with metaphoric versus literal category anaphors was replicated in the pattern of *net fixations* (1675 vs. 1353 ms), $F_1(2,34) = 25.5$, $MSE = 836,637$, $F_2(2,70) = 27.8$, $MSE = 1,860,809$. All critical sentences commenced with the definite article “The”, and fixation times on this article were not significantly different across conditions, $F_1(2,34) = 2.50$, $MSE = 3263$, $p = .136$, $F_2(2,70) = 3.70$, $MSE = 8802$, $p = .063$. This finding is unsurprising because the different versions of the critical sentences were identical at this point. There was a marginally significant effect of noun type on the first-pass fixations on the noun, $F_1(2,34) = 4.43$, $MSE = 11922$, $p = .051$, $F_2(2,70) = 5.14$, $MSE = 20818$, which occurred because readers spent slightly longer (34 ms) on metaphoric versus literal category anaphors ($p_1 = .054$, $p_2 = .034$). However, the first-pass time on the noun did not differ significantly between the non-anaphors and the metaphoric anaphors ($p_1 = .071$, $p_2 = .066$) or the literal anaphors ($p_1 = .083$, $p_2 = .057$). In fact, as indicated by the error bars in Figures 4.3 and 4.2; the time course pattern for processing the non-anaphoric versions of the sentences did not differ significantly from the time course patterns for processing the anaphoric sentences in any pairwise comparisons ($F_1, F_2 < 1$; $ps > .05$), with the exception that readers spent longer on the final word in the sentence when the critical noun was non-anaphoric ($p_1 = .041$, $p_2 = .039$).

In terms of the time course pattern for anaphoric sentences, the majority of the extra time spent in processing sentences with metaphoric versus literal anaphors

resulted from the increase in subsequent noun fixations; that is, the time spent re-reading the anaphor (284 vs. 124 ms), $F_1(2,34) = 22.9$, $MSE=233387$, $F_2(2,70) = 20.6$, $MSE=463556$. This mean difference of 160 ms on subsequent noun fixations was almost five times greater than the 34 ms difference for first-pass noun fixations. Finally, although fixations on post-anaphor words were not significantly longer following metaphoric anaphors versus literal anaphors (or non-anaphors) at any particular word position (e.g., N + 1, N+ 2, etc. in Figure 4.2), the difference in total time spent on the remainder of the sentence region for sentences with metaphoric versus literal anaphors (978 vs. 864 ms, see Figure 4.3) was significant by items and marginally significant by participants, $F_1(2,34)=3.9$, $MSE=115514$, $p=.066$, $F_2(1,70) = 10.1$, $MSE= 361746$. Figure 4.4 provides a breakdown of how the various regions of the sentence contribute to this net fixation difference between the literal and metaphoric sentences.

Figure 4.2: Time Course Data (ms) on first pass fixations (lower part of bars) and net fixations (total bar heights) for the different word positions in the critical sentence [Error bars are 95% CIs].

Figure 4.3: Synopsis of key time course measures (ms). [Error bars are 95% CIs].

Figure 4.4: Distribution of where the extra fixation time is spent in the metaphoric versus literal sentences.

Regressions to Critical Noun. Trials for which the comprehension question was answered incorrectly were excluded from this analysis because I wanted to determine how often regression (overt reanalysis) was required in cases where the reader had ultimately reached the correct referent assignment. Readers backtracked to re-read the anaphoric noun in 64% of the metaphoric trials and 36% of the literal trials, and this difference was significant both by items and participants; $F_1(2,34) = 25.3$, $MSE = 6992$, $F_2(2,70) = 15.88$, $MSE = 9767$. Regressions to the critical noun also occurred on some non-anaphoric trials (29%, which was not significantly less on the anaphoric literal trials, but was less than on the anaphoric metaphoric trials, $ps < .05$). However, the

evidence from Experiment 1 indicates that readers are unlikely to make preliminary classification errors (i.e., misclassify-as-anaphor errors) when processing these non-anaphoric nouns. Thus, the role for regression may be somewhat different in the processing of anaphoric and non-anaphoric sentences. In particular, such regressions may be indicative of inferential (vs. revision) processes to situate the new referent within the story schema and establish its relation with other referents. Sanford and Garrod (1994) suggest that such processes occur subsequent to preliminary of referent assignment.

LSA and Cloze Correlations. Across the 72 anaphoric noun items (36 literal, 36 metaphoric), the likelihood of regression to the noun was inversely correlated with cloze performance from Experiment 1 (misclassify-as-new errors), $r(70) = -.29, p = .01$, and with LSA estimates of anaphor-referent similarity, $r(70) = -.38, p = .001$. Thus, consistent with the memory-based account, degree of semantic overlap predicts the likelihood that reanalysis will be required.

Discussion

The present study replicates and extends the finding that readers often spend longer on sentences with metaphoric anaphors than on sentences with literal anaphors (e.g., Gibbs, 1990; Onishi & Murphy, 1993). Contrary to the response time patterns found with the word-by-word self-paced paradigm (Budiou & Anderson, 2002; Janus & Bever, 1985), these eye-tracking data indicate that the first-pass reading of the noun was not a substantial source of the difference in reading time between the metaphoric and literal sentences. Instead, the reading time difference owed to longer reading times on the

remainder of the sentence, and most substantially, to re-reading the anaphoric noun. These eye-tracking results are important in their own right, as they fill an evidentiary gap in the literature, however, in the context of the present research, I am particularly interested in what information these data can provide about the accuracy, duration and nature of the preliminary referent assignment process.

Accuracy of Preliminary Referent Assignment

The time course data provide information about the accuracy of preliminary referent assignment in the form of reanalysis delays and regressions. After the noun, the remainder of the sentence is identical for the literal and metaphoric versions. In spite of this, a longer net fixation time was spent on this region when the noun was metaphoric (marginal by participants). Thus, even when regressions do not occur, readers may nonetheless have initially misclassified the noun as a new referent and may incur a delay in revising this interpretation without physically looking back to the noun (covert re-analysis). Furthermore, the likelihood of regression corroborates the cloze evidence (Experiment 1) that misclassify-as-new errors occur during preliminary referent assignment, for both metaphoric and literal category anaphors. Such regressions occurred on almost two-thirds of the metaphoric trials, and on about one-third of the literal trials. Although regressions to the critical noun may not always indicate that the reader is revising their preliminary referent assignment (e.g., as in the case of regressions on the non-anaphoric trials), and a reader may sometimes revise their preliminary referent assignment without making a regression to the anaphor (i.e., contributing to the fixation delays I found in the remainder region of the metaphoric

sentences), the likelihood of regression in the present study was predicted by both by the cloze accuracies in Experiment 1, and by LSA estimates of anaphor-referent semantic overlap. Consequently, the data confirm that regressions to the anaphor can provide useful information about of the relative incidence of misclassify-as-new errors during preliminary referent assignment.

Reanalysis was apparently effective at enabling readers to subsequently rectify their initial misinterpretations. Based on results from Experiment 1, the accuracy rate of the initial referent assignment process for these metaphoric anaphors was approximately 52% (i.e., 48% misclassify-as-new errors), whereas the comprehension accuracy assessed at sentence end in the present experiment was 86%.

Thus, the data in Experiments 1 and 2 indicate that that reducing anaphor-antecedent semantic overlap increases the incidence of misclassify-as-new errors in preliminary assignment. This finding is compatible with a unified memory-based account, because reducing anaphor-antecedent semantic overlap reduces the activation that will spread to the intended discourse referent, thereby reducing its chances of retrieval. Although the accuracy patterns are also potentially compatible with the special-purpose search account, data about the duration of the preliminary noun processing (first-pass fixations) lend further support to the memory-based account over the special-purpose account.

Duration of Preliminary Referent Assignment

There was a small increase (35 ms) in the first-pass reading time on the metaphoric versus literal anaphors (marginal by participants). Under a special-purpose account,

this delay could be contributed by those metaphoric trials for which the antecedent was not found during preliminary referent assignment. Such a delay could arise because, despite failing, the search process took a long time (as suggested by Budiu & Anderson, 2002), and/or because the reader was required to *accommodate* the paradox of being unable to locate the antecedent despite the definite article (e.g., Heim, 1982). However, the same effect should have applied to the non-anaphors, but readers did not spend significantly longer on non-anaphors than on literal anaphors.

Therefore, I suspected instead that this extra first-pass time on the metaphoric nouns might be due to a confound of word frequency with noun type (metaphoric or literal) in the stimuli. It was not always possible to perfectly match the literal/metaphoric noun pairs for frequency as well as length and intended referent. As a result, some of the literal nouns were more frequent words than the metaphoric nouns (Kucera & Francis, 1967); and it has been found that less frequent words are typically subject to longer fixations (e.g., Rayner, Ashby, Pollatsek, & Reichle, 2004). After partialling out word frequency, no significant correlation remained between first-pass gaze duration on the noun (literal versus metaphoric), $r(70) = 0.068$, $p = .574$, nor between noun gaze and anaphor-referent LSA similarity, $r(70) = -.159$, $p = .197$. This absence of an effect of semantic overlap on the first-pass time on the noun is consistent with eye-tracking results in which conceptual overlap was manipulated via antecedent typicality (e.g., Duffy & Rayner, 1990; van Gompel et al., 2005).

The unified memory-based account is compatible with a minimal impact of a semantic overlap manipulation on the duration of preliminary referent assignment. In

particular, reducing anaphor-antecedent semantic overlap (i.e., using “butcher” vs. “doctor” to refer to Lisa's incompetent surgeon) reduces the amount of activation that will spread to the intended discourse referent [R: Lisa's incompetent surgeon]. However, this manipulation will not affect the activation levels of the generic competitors (i.e., the anaphorically-intended noun “doctor” results in substantial activation spread to its generic prototype [G:doctor]; and the alternative anaphor “butcher” would result in a comparable amount of activation spread to its generic prototype [G:butcher]). So no matter how far the activation of the intended referent falls below its competitor, the competitor will be readily retrieved and thereby puts a limit on the duration of preliminary assignment.

Reconciling the Present Results with Prior Results

The time course findings from the present eye-tracking study clearly do not coincide with the results obtained in the metaphoric anaphor studies which used the self-paced word-by-word reading paradigm (Budiou & Anderson, 2002; Janus & Bever, 1985). Those studies found that the only extra time spent on the metaphoric sentences occurred during the (first-pass) read of the anaphor, not in the remainder region. Recall that in those studies, the anaphoric noun was overwritten upon presentation of the next word and could not be physically re-read, precluding overt reanalysis. Thus, any reanalysis conducted by their readers should have produced a delay when reading the remainder of the sentence. However, those studies found no evidence of increased processing time (more frequent reanalysis) in the remainder region in the metaphoric condition.

There are two possible interpretations for the absence of a reanalysis delay: i) the participants retrieved the correct referent during preliminary assignment, which precluded any need for reanalysis (Janus & Bever, 1985); or ii) the participants retrieved the wrong referent during preliminary assignment (misclassify-as-new error), but did not subsequently reanalyze and recover from this erroneous interpretation when the full sentence became available (Budiu & Anderson, 2002). I consider these possibilities in turn.

The abnormally slow reading pace associated with the word-by-word paradigm (Rayner, 1998), and the requirement that readers certify their comprehension of the noun (by a button press to proceed to the next word) may have encouraged more attentive initial processing of the nouns than would have been the case in regular, passive reading. This paradigm may therefore have lead to a high incidence of early resolution, and precluded any need for later reanalysis. If so, an artifact of the word-by-word presentation paradigm might be to preemptively shift some burden of anaphor processing from the remainder of the sentence to the first-pass reading of the noun (see also von Gompel, 2005).

However, the difference in reading time for the literal versus metaphoric nouns in the word-by-word studies (about 40 ms) was about ten times smaller than the typical mean difference between sentences with metaphoric versus literal anaphors. And in terms of total sentence reading time (summed across the words or phrases), these word-by-word studies failed to replicate the well-established result that sentences with metaphoric anaphors tend to take longer, overall, to process than those with literal

anaphors when such sentences are presented in their entirety (Gibbs, 1990; Lemaire & Bianco, 2003; Noveck et al., 2001; Onishi & Murphy, 1993).

It seems unlikely that a segmented presentation of the sentence could equalize the processing difficulty of sentences with literal and metaphoric anaphors, so I join Budiu and Anderson (2002) in suspecting that their readers often made inaccurate preliminary assignments that were never corrected, which explains the absence of revision delays while processing the remainder of the sentence. Although Budiu and Anderson did pose questions to evaluate story comprehension, they later determined that a norming group was able to correctly answer many of the questions without exposure to the critical anaphoric sentences. Thus, the absence of increased reanalysis for the metaphoric cases in the word-by-word studies may indicate that many preliminary misinterpretations were never corrected. Further, because prior parts of the sentence disappear from view, a segmented presentation imposes an increased memory load that might deter integration and revision. Thus, in the word-by-word paradigm, readers might be more likely to misinterpret metaphoric anaphors, in particular, because they would normally have relied on returning to re-read such anaphors more frequently.

The regression evidence (present experiment) and cloze evidence (Experiment 1) that metaphoric anaphors are frequently treated as new referents in preliminary analysis would also not have been predicted by the probe study findings of Stewart and Heredia (2002). In very similar contexts, they found that immediately after their metaphoric noun anaphors, the metaphoric sense of the noun was primed (whereas the contextually inappropriate literal sense was not). This finding suggests that their participants were

able to correctly interpret the metaphoric nouns on-line, prior to hearing the next word. To explain this apparent inconsistency with the current results, I note that Stewart and Heredia's study involved *spoken* anaphors, which may be subject to slightly different processing than anaphors in *text*. For example, the duration of a spoken noun tends to be relatively long (e.g., 500 – 600 ms) in comparison to a reader's typical gaze duration on a noun (e.g., 200 ms for the metaphoric nouns in the present study). Although probe words were presented immediately after the spoken anaphor, their *listeners* had had more processing time than would have been invested by a *reader* at the same point in the sentence, and so it may be more likely for listeners to have correctly resolved the anaphor by this point.

Were it just a matter of having a few hundred milliseconds more time however, the cloze task participants in Experiment 1 should have performed as well as listeners, because the cloze task participants were able to take as long as they liked to process each anaphor. Alternately, prosodic cues may account for the more accurate and immediate anaphor resolution in speech versus text. In particular, there may be prosodic cues that can help a listener to differentiate anaphoric from non-anaphoric noun phrases (e.g., Brown, 1983). A more rigorous comparison of noun anaphor comprehension in text and speech remains an interesting avenue of inquiry, though one beyond the scope of the current research.

Summary

The time course data in this experiment suggest that readers spend extra time on the metaphoric sentences because reanalysis is more common for, but not limited to,

metaphoric noun anaphors. This pattern is compatible with the memory-based account and is consistent across the two studies.

In Experiment 1, initial misinterpretation rates are estimated from the percentage of non-anaphoric completions, whereas in Experiment 2, estimates of the initial misinterpretation rates are based on the percentage of trials in which readers return to reanalyze the anaphor. However, these estimates of the initial misinterpretation rates in Experiment 2 were higher than the estimates from Experiment 1. The studies had different groups of participants, which may account for some of the difference. Another possible explanation of this difference is that, in Experiment 1, participants were under no time constraints and the context sentences remained continuously available for re-inspection. In Experiment 2, previous sentences were overwritten. Further, the nature of the production task in Experiment 1 may have resulted in strategic and pragmatic processes that would not normally have been in play on encountering the noun during regular reading. For these reasons, the regressions may be better estimates of the actual rates of anaphoric misclassification, whereas the performance in Experiment 1 may reflect a ceiling effect, especially for the literal trials.

In summary, the data from Experiments 1 and 2 are consistent with a memory-based account, and suggest that preliminary referent retrieval was sometimes inaccurate, and consequently, that subsequent re-analysis was frequently required especially, but not exclusively, for metaphoric anaphors. Contrary to the special-purpose account, participants did not display evidence of an anaphoric bias in interpreting the definite nouns (i.e., there were no delays in first-pass processing for non-anaphoric definite

nouns). In all, the data do not provide support for the claim that there is a special-purpose (anaphorically-biased) referent assignment process triggered by the presence of the definite article "the". Instead, the results are consistent with a memory-based account in which the noun influences the classification of the referent by influencing whether a discourse referent or a generic referent will be most active. However, Gundel and her colleagues (Gundel, 2003; Gundel, Hedberg & Zacharski, 1993, 2001) suggest that the determiner "that" may promote an anaphoric bias in referent assignment, rather than the definite article "the". To test this hypothesis, and to gather additional evidence about the nature and difficulty of preliminary assignments, an additional cloze study was conducted, which is presented in the next chapter.

CHAPTER 5: Experiment 3

This experiment, like Experiment 1, involved a cloze (complete-the-sentence) task. Additionally, participants were asked to rate the difficulty of performing this task for each trial, both anaphoric and non-anaphoric, on a scale of 1 (very easy) to 10 (very difficult). Such difficulty ratings should provide additional information about whether there is any inherent difficulty in forming a new-referent (vs. anaphoric) interpretation for a definite noun phrase because of an anaphoric bias in the preliminary referent assignment process (as assumed in the special-purpose account). Further, to test whether the determiner “that” (vs. “the”) might introduce/increase an anaphoric bias as predicted by Gundel et al., (1993, 2001), I also manipulated the determiner that preceded the anaphoric nouns. Thus, the objectives of this experiment were (i) to provide a confirmation of the performance patterns in Experiment 1, and (ii) to gather additional evidence, in the form of difficulty ratings, about whether there is an anaphoric bias in the preliminary referent assignment process for noun phrases commencing “the” and/or noun phrases commencing with “that”.

The results from Experiment 1 indicated that when readers reach an anaphoric noun in a text, they sometimes treat it as a new referent rather than associating it with the intended discourse referent. Such results are consistent with a unified memory-based model, which suggests that readers do not operate under an anaphoric bias in preliminary referent assignment. Rather, retrieval of a discourse referent (versus a new referent schema) is contingent upon its relative activation level in memory. However, it is also possible that readers were operating under a default anaphoric bias, which was

nonetheless overridden to arrive at any new-referent interpretation. That is, readers may have engaged in a mental search, which involved the consideration and rejection of all discourse candidates, prior to resorting to any new-referent assignment. If so, readers should experience more difficulty when they encounter definite noun phrases that are non-anaphoric (or interpreted as such), than when they encounter definite noun phrases that are anaphoric. In contrast to special-purpose search models, the unified memory-based model does not predict a systematic increase in the difficulty of preliminary referent retrieval for non-anaphoric definite noun phrases (or those misinterpreted as such) relative to anaphoric ones, because in either case, automatic spread of activation will readily bring a referent to mind.

In terms of the determiner manipulation, the alleged anaphoric bias introduced by “that” should have two potential effects relative to “the”, and thus could provide a useful baseline. In particular, “that” could reduce the proportion of incorrect new-referent interpretations on anaphoric trials, and/or it could decrease the ease (if not the frequency) with which such new-referent interpretations are made. Thus, participants might report greater difficulty if they resort to a new-referent interpretation for an anaphor preceded by “That”, than when they form a new-referent interpretation for an anaphor (or non-anaphor) preceded by the “The”. Because the <“The”, literal anaphor> condition may be subject to a ceiling effect (95% correct in Experiment 1), any improvement in accuracy on “That” trials should be especially pronounced in the metaphoric condition. The information about anaphoric status allegedly carried by “That” may simulate the availability of prosodic information about anaphoric status in

speech (e.g., Brown, 1983). On this view, this determiner manipulation may help to reconcile the modest preliminary retrieval performance in Experiments 1 and 2, with the high accuracy of preliminary retrieval reported by Stewart and Heredia (2002) for spoken metaphoric anaphors.

Ideally, then, the determiner manipulation may provide a means to determine the extent to which preliminary interpretation errors can be avoided (and/or interpretation effort reduced) when the anaphoric status of the noun is known a priori. However, parsimony seems to argue against referent retrieval procedures that are different in kind for definite and demonstrative noun phrases. Indeed a memory-based model could apply across all cases. To the author's knowledge, Gundel et al.'s distinction between "the" versus "that" was based on a theoretical analysis that had yet to be empirically tested at the time of this writing. Thus, the present study also serves as such an empirical test. If the predicted effects of determiner on accuracy and difficulty ratings are not found, the data will suggest that readers do not feel constrained by either determiner to associate the ensuing noun with a familiar referent from the discourse, rather than with new referent (during preliminary referent assignment). This outcome would be compatible with a memory-based model that applies across all cases, because in a memory based account, preliminary referent assignment is driven by the noun (and its semantic associations) rather than by the determiner.

Method

Participants. The participants were 40 Carleton University students (17 male and 23 female) from Carleton University. All participants were native speakers of English.

Participants were given \$10 or course credit for their participation.

Materials and Design. This experiment uses the same 36 contexts used for the cloze task in Experiment 1. However, the determiner ("the" vs. "that") was also manipulated for the anaphoric critical nouns. Thus, there were four possible versions of the anaphoric critical noun phrases for versions of the final sentence for each of the 36 contexts (e.g., "The doctor...", "That doctor...", "The butcher...", "That butcher...").

Four sets of materials were constructed such that each set contained 12 texts in each of the four conditions. Each set of materials contained only one version of a given text, and the versions were counterbalanced across the sets. Each set also included 14 texts with non-anaphoric nouns in the critical position (preceded by the article "The"), which resulted in a total of 54 texts per set. A random order was generated for the stories, and this random order was used in each set.

Procedure. Each participant was randomly assigned to one of the four sets of materials, and received the corresponding booklet containing 54 incomplete texts. Participants were instructed to write a coherent ending for the incomplete final sentence in each story. Participants were also asked to rate the difficulty of completing each sentence on a scale of 1 (very easy) to 10 (very difficult). No explicit time limit was given, and the experimental session typically took less than 1 hour.

Results

As in Experiment 1, performance accuracy on trials with intended anaphors was operationalized as the percentage of trials on which participants' completions 'correctly' treat the critical noun as an anaphor. Two coders separately assessed

sentence completions to establish consensus as to which were anaphoric. Inter-coder agreement was very high (98%), and the few outstanding cases were resolved in discussion to the satisfaction of both coders. The participants' difficulty ratings for completing each sentence provide an additional dependent variable, here. For each dependent variable, ANOVAs were conducted by participants and by items to test for any main effects of, and any interaction between, the two independent variables: noun type (literal, metaphoric) and determiner type (the, that).

Performance Accuracy: Percentage of Anaphoric Completions

The performance accuracy percentages under the four anaphoric conditions are shown in Figure 5.1. As in Experiment 1, participants made more anaphoric completions when the (pre-experimental) semantic association between the potential anaphor and the intended antecedent was strong (potential literal anaphor), than when it was weak (potential metaphoric anaphor), $F_1(1,39) = 77.77$, $MSE = 29641$, $p = .001$, and by items, $F_2(1,35) = 3.46$, $MSE = 266.78$, $p = .000$. Participants were only slightly (3%) more apt to make an anaphoric connection when a potential anaphor was preceded by "That" versus "The". However, this small effect of determiner was only significant by participants, $F_1(1,39) = 5.03$, $MSE = 444.44$, $p = .031$, and not by items, $F_2(1,35) = 3.46$, $MSE = 4.00$, $p = .071$. Furthermore, there was no significant interaction between determiner and noun type ($F_1, F_2 < 1$).

The above accuracy analysis applied to the trials in which the critical noun was intended as an anaphor. For non-anaphoric critical nouns, participants were always accurate in treating the noun as a new-referent as opposed to associating a prior

discourse referent (i.e., a misclassify-as-old error), with the exception of a single anomalous item that four participants treated as anaphoric (0.6% of non-anaphoric trials): “The squirrel”, in a picnic context was occasionally interpreted to refer to a man who was a very hearty eater in the story, presumably because he was packing or ‘squirreling’ away the food. In all, participants were far more apt to treat an anaphoric noun as a new referent (misclassify-as-new error) than to treat a non-anaphoric noun as an anaphor (misclassify-as-old error).

Figure 5.1: Percent of anaphoric completions made for anaphorically-intended nouns [Error bars are 95% CI's for the effect of noun type].

Difficulty ratings

Mean difficulty ratings for anaphoric trials were analyzed in a 2(determiner: the, that) x 2(anaphor type: literal, metaphoric) repeated measures ANOVA. As illustrated in Figure 5.2 (a), participants found sentences with potential metaphoric

anaphors more difficult to complete than sentences with potential literal anaphors, $F_1(1,39) = 18.45$, $MSE = 11.14$, $p = .000$, $F_2(1,36) = 7.75$, $MSE = 9.56$, $p = .009$. However, varying the determiner ("the" vs. "that") had no effect on the difficulty ratings, $F_1(1,39) = 0.10$, $MSE = 0.026$, $p = .753$, $F_2(1,36) = .04$, $MSE = .02$, $p = .845$, and there was no interaction of determiner type with anaphor type, $F_1(1,39) = 0.11$, $MSE = 0.11$, $p = .746$, $F_2(1,36) = .21$, $MSE = .07$, $p = .647$.

Additional analyses were done to determine if participants found the non-anaphoric definite noun phrases difficult to process. According to the special-purpose account, readers must conduct and conduct a vain antecedent search overcome an anaphoric bias to process any non-anaphoric definite noun. Thus it is important to determine the extent to which the extra difficulty for metaphoric anaphors (vs. literal anaphors) was due to trials in which the participant made the anaphoric connection versus or due to trials in which they did not make the intended connection. Panel (b) in Figure 5.2 provides a break down of difficulty ratings according to the type of determiner and the type of interpretation that participants made about the noun phrase (i.e., anaphoric or new referent).

The condition of literal anaphors treated as new referents contains too few data points (<5%) to be included in a statistical analysis. However, an analysis restricted to trials in which anaphoric completions were made revealed that the determiner had no effect on the perceived difficulty of making the associations. Thus, the determiner "that" did not increase the ease of making the association relative to "the". However, even when participants made the intended anaphoric association, they gave higher

difficulty ratings when the semantic association between the head noun and the intended antecedent was weak (metaphoric anaphor), than when the association was strong (literal anaphor), $F_1(1,39) = 14.21$, $MSE = 9.62$, $p = .001$, $F_2(1,35) = 6.02$, $MSE = 7.93$, $p = .019$.

Figure 5.2: Difficulty ratings for the anaphoric trials: (a) includes all potential anaphors regardless of the nature of the participant's interpretation; (b) illustrates the breakdown of difficulty ratings according to whether the noun was interpreted as a new referent (hollow markers) or interpreted as an anaphor (solid markers).

Figure 5.3: Difficulty ratings (1=very easy; 10 =very difficult) by interpretation type. [Error Bars are 95% CIs].

Figure 5.3. Provides a summary of difficulty ratings by interpretation type. Because the previous analyses established no effect of determiner, the data have been collapsed across determiner type. Non-anaphors were given the lowest difficulty rating, indicating that participants apparently found nothing inherently daunting about forming new-referent interpretations for definite noun phrases. There was a significant effect in difficulty overall, $F(3,70.4) = 8.22$, $MSE = 13.34$, $p = .001$, which resulted because metaphoric anaphors were rated more difficult to deal with than literal anaphors regardless of whether the participant correctly interpreted the metaphoric noun as anaphoric ($p = .004$) or treated it as a new referent ($p = .004$). Among the metaphoric anaphors, difficulty did not differ with interpretation type (misclassify-as-new vs. interpret-as-anaphoric, $p = .249$). Nor did difficulty differ by interpretation type

between literal nouns interpreted as new referents, and literal nouns interpreted as anaphoric ($p = .153$).

Discussion

This experiment extends and replicates the results of Experiment 1. The data confirm that potential metaphoric anaphors (which have low anaphor-antecedent semantic overlap) are often misinterpreted as new referents on encounter, and that literal anaphors are also sometimes misinterpreted as new referents. However, in this experiment, a determiner manipulation was introduced and a difficulty rating measure was also obtained, which provided additional evidence against an anaphoric bias during preliminary referent retrieval. In the next section, I discuss what the difficulty rating data suggest about the nature of the referent assignment process. The determiner manipulation (“The” vs. “That”) produced very little effect and will be discussed in the subsequent section.

Difficulty Ratings and the Nature of the Referent Assignment Process

Participants rated potential metaphoric anaphors as more difficult to deal with than potential literal anaphors -- both when the intended connection was made for the metaphoric anaphor and when the noun was treated as a new referent. A possible explanation for the increased difficulty in latter, misclassify-as-new case is that readers had to overcome an anaphoric bias (i.e., special-purpose search model) to form a new-referent interpretation of the potentially metaphoric noun. However, the evidence suggests that forming a new-referent interpretation of a definite noun phrase does not pose any particular difficulty, per se. In particular, participants rated non-anaphoric

definite noun phrases as slightly *less* difficult to deal with than anaphoric definite noun phrases, which suggests that participants did not need to overcome any anaphoric bias in forming new referent interpretations. This evidence casts further doubt on special-purpose search accounts, which presuppose an immediate anaphoric bias in the interpretation of definite noun phrases.

Why then, in absence of an anaphoric bias in preliminary referent assignment, would potentially metaphoric anaphors be given relatively high difficulty ratings even when the participants interpreted them as new referents? The cloze task effectively involved two components: assigning a referent to the definite noun phrase, and coming up with a sentence ending to give this referent some role to play in the story. Although assigning a new referent to a definite noun phrase may not pose any particular difficulty, coming up with a role for the new item to play may have been more difficult in the case of metaphoric nouns (interpreted literally as non-anaphors) than for actual non-anaphoric nouns. For example, participants may have had to invest time and effort to come up with a role for a new item like a (literal) butcher to play in a story about a girl's injured knee and her operation. In all, the evidence suggests that the subjective difficulty for metaphoric anaphors interpreted as new referents owes to this latter demand of the cloze task, which occurred subsequent to referent assignment.

The focus of the present research program, however, is on the nature, difficulty and accuracy of the preliminary referent assignment process itself. There is no evidence to suggest that it is systematically more difficult (the present experiment) or more time-consuming (Experiment 2) to form a new referent interpretation of a definite noun

phrase than an anaphoric interpretation. Such findings are inconsistent with special-purpose search accounts, which have an anaphoric bias however the findings are consistent with a unified memory-based model.

When participants were successful at making the intended anaphoric associations, they nonetheless rated the metaphoric trials more difficult than the literal trials. Once the right referent is assigned, coming up with a completion for a given story should, in theory, be equally difficult across the two anaphoric versions, so it would seem that this increased difficulty owes to referent assignment process itself. Under a unified memory-based model, successful assignment in preliminary processing requires that the right referent receive sufficient activation to exceed its competitors, within and outside the discourse. For a given story context, semantic similarity dictates that a literal anaphor will resonate more strongly with the intended referent than a metaphoric anaphor, so that even when the latter produces sufficient activation for retrieval, the absolute level of activation achieved by the referent in response to a metaphoric anaphor will be lower than that produced by a literal anaphor. This absolute activation advantage for literal anaphors might translate into more rapid retrieval and/or more certainty and speed in any potential verification stage. Hence, the memory-based model is compatible with the possibility that even when a metaphoric anaphor is correctly interpreted in preliminary assignment, it may be slightly more difficult to process than a corresponding literal anaphor. That said, there was no significant difference in first-pass time between the two types of anaphors (Experiment 2), though it is possible that part of the preliminary process may spillover onto the subsequent word (e.g., the

verification process; Duffy et al., 1990).

Correctly interpreted metaphoric anaphors may also have been rated more difficult than correctly interpreted literal anaphors in the present cloze study because participants employed more deliberative and pragmatic strategies when interpreting these potential metaphoric anaphors than they would have during preliminary referent assignment in regular reading. For example, a participant may have been initially inclined towards a new-referent literal interpretation, but the difficulty encountered in creating a coherent role for the new referent to play may have induced the participant to re-read the story and/or revise the assignment to arrive at an anaphoric interpretation. Thus, the difficulty ratings for correctly interpreted metaphoric anaphors may (also) reflect the time and effort employed in such pragmatic task-triggered re-analysis processes. Such processes may have resulted in slightly inflated accuracy estimates (Figure 5.1) relative to those that would have occurred in regular reading during preliminary referent assignment.

The Determiner: “That” versus “The”

The determiner manipulation had only a small effect on performance accuracy (3%) that was non-significant by items. The most straightforward interpretation of these results is that, contra Gundel et al. (1993, 2001), the demonstrative “that” does not introduce (or increase) anaphoric bias in the preliminary referent assignment process, in comparison to “the”. The very slight performance improvement in the “that” case could be attributed to other factors. For example, many of critical noun phrases are epithets, and “that” may commonly precede, and aid in the recognition of, an epithet

(e.g., “That weasel”). An advantage for the “That” construction in identifying epithets is independent of issues of anaphoricity, because in general, an epithet can also be used to introduce a new entity into the discourse.

“That” clearly did not induce an anaphoric bias of the strength to strictly compel the participant to pick a referent from within the discourse (e.g., the discourse candidate with the strongest semantic association to the anaphor). However, it is possible that “that” might nonetheless have induced an anaphoric bias that involved an increased, though not complete, reluctance to resort to a new-referent assignment. The difficulty ratings, however, do not support this suggestion. Participants did not report any greater difficulty when they formed a new referent interpretation for a noun preceded by “That” than for a noun preceded by “The”.⁵

Thus, contrary to one hypothesis of Gundel et al. (1993, 2001), the present study provides no evidence to suggest that the demonstrative determiner “That”, introduces an

⁵ Admittedly, time course evidence from an eye-tracking study would provide useful additional evidence about whether the preliminary referent assignment process is different in kind for nouns preceded by “The” versus “That”. The present study indicated that, regardless of the determiner, readers often initially formed new-referent interpretations for metaphoric anaphors, so readers of complete versions of these anaphoric sentences would subsequently have to perform a reanalysis in a comparable number of “the” and “that” trials. However, for “that” trials, readers may also have invested extra time and effort, albeit largely without benefit, during preliminary referent assignment, to overcome the alleged anaphoric bias. Thus, overall sentence processing time should be longer for sentences with metaphoric anaphors preceded by “That” versus “The”, if the former induces an immediate anaphoric bias (e.g., a special-purpose antecedent search). Under Just and Carpenter’s (1980) eye-mind assumption, a prolonged referent assignment process in the “that” case should be evidenced by, and localized to, longer first-pass times on the noun. However, even if referent assignment spills over onto the subsequent word, an anaphoric bias in the “that” case could still be evidenced by longer net processing times on the sentences. If neither “the” nor “that” introduces an immediate anaphoric bias (or they introduce a comparable level of bias), then the duration of preliminary reference assignment, the incidence of reanalysis, and consequently the time course patterns should be comparable for the two determiners. Given that “that” had no effect on difficulty ratings and only a marginal effect on accuracy in the present study, the most straightforward conclusion is that it does not introduce an anaphoric bias, and consequently, the dissertation committee ruled that a further, time-course study was not well motivated.

anaphoric bias into preliminary referent assignment. That said, the current account is in accord with Gundel et al. in terms of the claim that the definite determiner, “The”, (also) does not induce an anaphoric bias in preliminary referent retrieval. However, there are some subtle, yet important, distinctions to be made between their account and the current one regarding the definite article. In both treatments the hypothesis that the definite article produces no anaphoric bias was motivated and supported by corpus evidence that definite noun phrases are frequently non-anaphoric. In the present research, the absence of an anaphoric bias in preliminary processing was further motivated by and bound with my computational model about the nature of the referent retrieval process – namely, a unified memory-based model, in which new referent schemas compete for retrieval in parallel with discourse referents. In contrast, Gundel et al.’s treatment does not provide any mechanistic proposal about the referent retrieval process. They are proposing a linguistic theory about the meaning of the word “the”, rather than a procedural model of referent assignment. In particular, they suggest that the definite article “the” (in contrast to “that”) does *not* convey the information that the referent for the noun phrase is already familiar to the reader. Their claim is somewhat stronger than the present claim in this respect. In particular, while the present evidence suggests that “The” does not induce an immediate and observable anaphoric bias during the *preliminary referent assignment process*, the present treatment leaves open the possibility that the definite article may nonetheless convey a latent anaphoric bias that may impact subsequent pragmatic and integrative processes (which are outside the present scope).

Conclusions

In all, the data thus far provide no evidence of an anaphoric bias in preliminary referent assignment for nouns preceded by the definite (or the demonstrative) determiner. Participants experienced no particular difficulty in associating new referents to definite noun phrases (present experiment), nor did they take longer to process non-anaphoric definite noun phrases than anaphoric definite noun phrases (Experiment 2). The relative success at handling literal and metaphoric anaphors during preliminary processing can be readily explained by an unbiased, unified memory-based model of referent retrieval, which is driven by the noun rather than the determiner.

Unfortunately, the determiner manipulation cannot provide a way to decouple the relative contributions of the two factors which contribute to the challenge of interpreting metaphoric anaphors: low semantic similarity to the antecedent, and absence of a priori knowledge about the anaphoric status of the noun. In Experiment 4, another attempt was made to decouple these contributions within an eye-tracked reading paradigm. In particular, a manipulation was introduced to explicitly mark the anaphoric status of the anaphoric and non-anaphoric nouns (using font colour) in one block of trials.

CHAPTER 6: Experiment 4

This experiment extends the preceding research in two respects: i) an additional anaphor type is included which exactly matches the antecedent term - resulting in 4 types of critical noun (antecedent-match anaphor, category anaphor, metaphoric anaphor, and non-anaphor), and ii) a font colour manipulation is introduced to explicitly convey the anaphoric status (anaphoric vs. non-anaphoric) of the critical noun in the second block of trials.

Under a memory-based account, the amount of activation spread by the anaphor to the intended referent depends on the semantic association between the anaphor and antecedent – thus the activation spread to the intended referent will be strongest when the anaphoric term (concept) is identical to original antecedent term (concept). In cases where the anaphor does not exactly match the antecedent that introduced the referent, the degree of spreading activation received by the referent will depend on the strength of the semantic association between the anaphor and antecedent concepts. Consider (10a-c) which lists three possible anaphors for the antecedent “terrier” in (9).

(9) John took his terrier to the park.

(10a-c) The terrier/dog/monster terrorized the birds.

The LSA estimates of the association strengths for the metaphoric condition (terrier-monster), the category condition (terrier-dog) and the matching condition (terrier-terrier) are 0.14, 0.72 and 1, respectively. Thus, a memory-based model predicts greater referent activation, and thus higher first-pass retrieval likelihood for category anaphors than metaphoric anaphors, and, in turn, higher first-pass retrieval

likelihood for antecedent-match anaphors than for category anaphors, though little time course data is available about this latter contrast. In their Experiment 1, McKoon and Ratcliff (1980, see also Garnham, 1989) found that net reading time was shorter for sentences containing antecedent-match anaphors than for sentences containing category anaphors, however this difference did not reach significance in their second experiment. Further, sentence response times provide no information about the distribution of processing within the sentence (e.g., first-pass noun time). Consequently, the present experiment compares the processing of these three types of anaphors (antecedent-match, category and metaphoric) and non-anaphors using eye-tracked reading. Under a memory-based account, the likelihood of correct referent retrieval is highest for antecedent-match anaphors, so the likelihood of regression (and the processing time on the remainder of the sentence) is expected to be the lowest when the noun is an antecedent-match anaphor, and highest when the noun is a metaphoric anaphor.

In addition to the manipulation of noun type, in this experiment I also manipulated whether the anaphoric status of the noun was marked or unmarked. In the marked condition, the font colour of the critical noun in each final sentence indicated its anaphoric status - that is, the noun appeared in one colour (e.g., red) if it was anaphoric, and another colour (e.g., green) if it was non-anaphoric. Although such marking does not naturally occur in text, the motivation for such marking was similar to that for the determiner manipulation in Experiment 3.

For an anaphoric noun, the information that it is anaphoric could lower the 'threshold' set by the processing system for accepting a discourse candidate as a match

for the anaphor. In special-purpose search models, each discourse referent, including the intended one, will come up for consideration in the course of a search through the discourse representation. In the experimental stimuli, the intended referent is almost always the best candidate within the discourse in terms of LSA compatibility with the anaphor, even when it is metaphoric (see Appendix B). Thus, readers could achieve near perfect first-pass performance, even for metaphoric anaphors, if they used a special-purpose search with a permissive threshold. If so, the discrepancy between regression likelihood (and post-noun processing time) for metaphoric versus both types of literal anaphors (antecedent-match and category) should be reduced in the marked block.

What effect of marking might be expected under a unified memory-based account? The retrieval mechanism is common to anaphoric and non-anaphoric nouns, and is automatic and cognitively impenetrable. The memory-based account, per se, only encompasses the retrieval of a referent, and does not speak directly to the nature of any subsequent verification process. However, crucially, retrieval is a pre-requisite for assignment. Thus, under a memory-based account, marking a noun as anaphoric would provide only a minimal level of improvement in first-pass accuracy for low semantic overlap anaphors (i.e., metaphoric anaphors), because the memory mechanisms will frequently fail to retrieve the right referent for consideration in the first place (because it would be less accessible than a new referent schema).

How might marking affect the processing of non-anaphoric nouns? Special-purpose search models predict that readers make a default assumption of anaphoricity for

definite nouns, so providing information that a noun is non-anaphoric could reduce first-pass processing time by preempting a vain antecedent search. In contrast, for a unified memory-based model, the referent retrieval process is the same regardless of the anaphoric status of the noun – so marking would not be expected to provide much of a processing advantage in the case of non-anaphoric nouns. The automatic memory mechanisms can rapidly result in the retrieval of the appropriate generic referent without recourse to any additional information.

Method

Participants. Twenty-four undergraduate students from Carleton University participated. All participants were native speakers of English, and had normal or corrected-to-normal vision, and normal colour vision. Participants were given \$10 or course credit for their participation.

Materials. The stimuli in this experiment consisted of the 36 anaphoric stories in Experiments 3, plus 6 additional anaphoric stories. In this study, each of the 42 anaphoric stories had three different versions of its final sentence corresponding to three different anaphor types: antecedent-match, category, and metaphoric. The materials also included 30 non-anaphoric stories (for which the critical noun in the final sentence was not an anaphor). In all stories, the fourth sentence was the critical sentence and commenced with a definite noun phrase: “The (noun)”. These stimuli are presented in Appendix A, and descriptive statistics about the lengths (i.e., number of letters), corpus frequencies, and LSA anaphor-referent similarities for the different types of critical nouns are presented in Appendix B. To summarize, the different versions of the

anaphor differed systematically in their LSA similarity with the intended referent, $F(2,70.24) = 36.092$, $MSE = 1.143$, $p = .000^6$. Comparisons using the Bonferroni adjustment confirmed that the LSA means for the three types of anaphors differed significantly pairwise. There were no significant differences in length between the anaphoric and non-anaphoric critical nouns, $F(1,154) = .038$, $MSE = .14$, $p = .846$, or between the three types of anaphors, $F(2,82) = .52$, $MSE = 1.50$, $p = .600$. However, the metaphoric anaphors had lower log frequencies than both types of literal anaphors, $F(2,82) = 4.72$, $MSE = 2.82$, $p = .011$.

Six sets of materials were constructed, each containing 6 practice texts, the 30 non-anaphoric stories and 42 anaphoric stories (i.e., 14 anaphoric texts in each of the 3 anaphor types - antecedent-match, category, metaphoric). Each set was partitioned evenly into 2 blocks of 39 trials. Each set of materials contained only one version of a particular anaphoric text, and the versions were counterbalanced across set and block. The non-anaphoric items were divided into two groups and counterbalanced across block so that each item was seen by half the participants in the marked block and half the participants in the unmarked block. A random order was generated for the stories within each block of each set, and this random order was used for each participant who received that set of materials.

For stimuli in Block 2 (marked anaphoric status), the font colour of each critical noun phrase was set according to whether the noun was anaphoric (e.g., red) or non-

⁶ The assumption of sphericity was not satisfied according to Mauchly's Test ($p = .026$), therefore the appropriate Greenhouse-Geisser adjustments were applied to the tests of significance.

anaphoric (e.g., green). As a control for the effect of colour itself, the critical nouns in the unmarked block (Block 1) were also presented in colour, (e.g., orange or purple) however these colours were assigned randomly so that there was no correlation between the anaphoric status of the noun and its colour. All font colours were counterbalanced across condition and participant.

Procedure. Each participant was randomly assigned to one of the 6 sets of materials. The procedure is similar to that described in Experiment 2 with the exception that for the final, critical sentence of each story, the noun phrase (in coloured font) was displayed first, and participants were instructed to press the space bar as soon as they had comprehended the phrase. Then the remainder of the sentence (in white font) was appended to the noun phrase, so that the full sentence was displayed upon the screen, and participants were required to press the space bar again once they had comprehended the full sentence.

Each participant first received the unmarked stimuli, Block 1. For this unmarked block, participants were told, truthfully, that the colour (e.g., purple or orange) of the head noun in the critical sentence had no meaning/relevance. After completing Block 1, participants took a short break, and then read a description of the colour marking scheme for the marked block, Block 2. In particular, participants were told that nouns in one colour (e.g., green) were introducing a new object or character into the story, whereas nouns in another colour (e.g., red) “referred to an item/character already present in the story” (i.e., an anaphoric noun). Participants were also instructed that an item or character might be referred to with a different label than that used to initially

introduce it in the story (e.g., “Loblaws” might later be referred to as “the store”, and although the labels differ, “store” still refers to an item already present in the story).

Participants then performed a training exercise to memorize the colour marking scheme (e.g., red = anaphoric, green = non-anaphoric), and to confirm that they could visually discriminate the relevant colours. On each trial of the training exercise they were presented with a dummy noun phrase “The XXXXX” located in the same position on the screen that the critical noun phrases appeared in the main experimental task. The dummy noun phrase was presented in one or the other marking colours (e.g., red or green) in a randomized order. On the lower left and right of the screen respectively were the status choices “old item” and “new item” (in white font). The participant was instructed to press the left mouse button if the colour of the dummy phrase indicated the status choice on the left of the screen, or the right mouse button if the colour of the dummy phrase indicated the status choice on the right of the screen.

To prevent the participant from merely making an association of the phrase colour to a particular mouse button (rather than to its old item vs. new item meaning) the order of the status choices at the bottom of the screen were also randomized from training trial to training trial. For each training trial the participant was given feedback about the correctness of their response (e.g., “CORRECT: Red refers to a NEW ITEM”). Participants were initially presented with 20 training trials, and training was complete if no more than one response was incorrect. Otherwise, additional sets of 12 training trials were presented, until no more than one response in the set was incorrect. Then, participants performed the same training task for 12 training stories that had colour-

marked critical nouns (vs. "XXXXX" in the first training exercise). After training was completed, participants completed the second half of the experiment (Block 2 – marked). No explicit time limit was given. The full experimental session typically lasted one hour.

Results

Comprehension Question Accuracy

Figure 6.1 presents a breakdown of question answering scores (% correct) in the various conditions. As in Experiment 2, on anaphoric trials, questions were designed to test whether the anaphoric association had ultimately been made in the final sentence. Overall, participants answered the vast majority of these questions correctly (95.6%). Thus, regardless of any preliminary misinterpretation, participants had made the intended association by the time they had finished processing the critical sentence (or the question). Participants had better question answering performance in the block for which the anaphoric status of the critical noun was marked as opposed to unmarked (97.6 % vs. 94.0 %), $F_1(1,23) = 4.67$, $MSE = 442.33$, $p = .041$, $F_2(1,41) = 6.83$, $MSE = .08$, $p = .012$. To prevent participants from strategically focusing on the content in the final, critical sentence, the questions on the non-anaphoric trials related to content in the first three sentences of the stories. There was no significant effect of block (noun marking) on the question answering performance on non-anaphoric trials, $F_1(1,23) = 1.66$, $MSE = 82.99$, $p = .210$, $F_2(1,28) = 2.02$, $MSE = .009$, $p = .166$.

Thus, the improvement in question answering performance on anaphoric trials in the second (marked) block cannot be attributed to a general practice effect, in which

overall comprehension and question answering performance were improved. Instead, the marking specifically improved the readers' ability to make the anaphoric associations. A 2(noun type: anaphor, non-anaphor) by 2(block: unmarked block, marked block) repeated measures analysis confirmed the interaction of noun type (anaphor vs. non-anaphor) with block, $F_1(1,23) = 5.54$, $MSE = 22.58$, $p = .030$, $F_2(1,306) = 3.67$, $MSE = .044$, $p = .056$. This analysis also indicated, unsurprisingly, that readers tended to perform better on questions which focused on the most recent content in the final sentence of the story (i.e. questions for anaphoric trials) than on questions that focused mainly on earlier content (i.e., questions for non-anaphoric trials), $F_1(1,23) = 10.27$, $MSE = 227.18$, $p = .005$, $F_2(1,306) = 4.28$, $MSE = .051$, $p = .039$.

Figure 6.1: Percentage of Comprehension Questions Answered Correctly. Panel (a) illustrates the improvement in Block 2 on anaphoric trials, and the interaction of block with anaphoricity. Panel (b) provides a breakdown by anaphor type [Error Bars are 95% CI's for Bonferroni pairwise comparisons].

Among the anaphoric trials, there was comparable question answering performance on trials with literal and metaphoric anaphors (96.1% vs. 95.2%), thus, as in Experiment 2, the effect of anaphor type on question performance was not significant, $F_1(2,46) = .47$, $MSE = 24.72$, $p = .630$, $F_2(2,82) = .25$, $MSE = .003$, $p = .783$. An interaction of anaphor type with block was significant in the analysis by participants, $F_1(2,46) = 3.30$, $MSE = 198.57$, $p = .046$, and marginally significant in the analysis by items, $F_2(2,82) = 2.78$, $MSE = .036$, $p = .068$. In particular, as evident by the pairwise Bonferroni comparisons in Figure 6.1(b), the overall improvement in performance on anaphoric trials in Block 2 resulted from an improvement for both types of literal anaphors (antecedent-match anaphors, $p = .028$; category anaphors, $p = .003$). However question answering performance did not differ significantly between blocks for the metaphoric anaphors ($p = .713$).

Why might question answering performance improve in the marked block for trials involving literal anaphors? In the marked block, participants were explicitly told to expect that the critical noun phrase might refer to a referent from the story with a different label than was used earlier (e.g., surgeon vs. doctor). Being explicitly told that the label for a referent would sometimes be altered within the story may have reduced potential confusion in interpreting the questions themselves. In particular, the label for the referent in the question sometimes reverted back to the original label. For example, the critical sentence: "The doctor didn't care." could be followed by the question: "The surgeon who performed the operation expressed regret", which may have raised the reader's suspicion that a deliberate distinction was being drawn between

the doctor and the surgeon, and thus undermined the readers' confidence in their correct co-referential interpretation, and suspect that a distinction was being drawn between the doctor and the surgeon. This concern would be reduced in the marked block.

For the metaphoric trials, questioning answering did not improve in the marked block. The meta-level expectation that a new item would sometimes be introduced "out of the blue" in the final sentence of a story may have reduced the readers' expectations about the coherence and relevance of this final sentence, which might have reduced the readers' tendency to recognize and rectify preliminary misinterpretations in the metaphoric case.

Time Course

Noun type by block analyses were done for several time course measures: i) net fixations for the sentence as a whole; ii) net fixations in the remainder region (cumulated from the word following the noun, Noun+1, until sentence end), iii) the first pass and net fixations for each word position in the critical sentences (Figure 6.2), iv) regressions to reanalyze the noun, and v) the duration of subsequent noun fixations when such regressions occurred.

In contrast to Experiment 2, the present experiment was blocked and the presentation order of the stories varied across participant. Preliminary analyses (see Appendix C) of the reading rates for the critical sentences and the context sentences indicated that trial order, and the total number of characters displayed on the screen significantly influenced reading rate. That is, participants increased their overall reading pace on long sentences (and noun phrases) relative to short ones, and on later

trials relative to early trials. To prevent the variance from these factors from obscuring the effects of the experimental manipulations (in view of the conservative cell size of $n=7$ items of each anaphor type per participant per block), the 4(noun type: antecedent-match anaphor, category anaphor, metaphoric anaphor, non-anaphor) $\times$ 2(block: unmarked block, marked block) ANOVAs were conducted between-groups rather than as repeated measures in order that trial number could be included as a covariate (along with sentence and word length). Participant was also included as a random variable to control for individual differences in reading speed. For each fixation measure, two such ANOVAs were done: F_{in} to control for trial order (1-39) within each block, and F_{bet} , to control for absolute trial order across blocks (1-78). Table 6.2 at the end of the chapter summarizes all the results.

Net Sentence Fixations and Net Remainder Fixations

Absolute remainder fixations and sentence fixations were lower in Block 2 (Table 6.2a, F_{in}), however, after controlling for cross-block practice effects (Table 6.2b, F_{bet}) these measures were actually relatively higher in Block 2, due to the contribution of fixations on the noun and the last word in the sentence.

The type of noun also affected the overall reading time for the sentence (F_{in} , F_{bet}). In particular, pairwise comparisons indicated that sentences with metaphoric anaphors took significantly longer to read (3105 ms) than sentences with the other 3 types of nouns, that is – sentences with literal category anaphors (2809 ms), antecedent-match anaphors (2855 ms) and non-anaphoric nouns (2844 ms). This pattern applied across blocks, as well as within each block. When the analysis was restricted to the remainder

of the sentence after the anaphor, however, noun type did not have a significant effect on net fixations. As will be seen below, the longer net sentence fixations for metaphoric anaphors were due to longer subsequent noun fixations.

Fixations for Each Word Position in Critical Sentence

Figure 6.2 depicts the first-pass time (lower half of bars) and net fixations (total bar height) spent on each word in the critical sentences in: (a) the unmarked block, and (b) the marked block. Fixations on the noun and on the last word in the critical sentence are inflated relative to the other words in part because readers were required to press a response key to signal their comprehension at these junctures. The details of these ANOVAs for the first-pass and net fixations are presented in Table 6.2 at the end of the chapter (e.g., F -statistics and p -values for the main effects and interactions).

The F_{in} ANOVAs controlled for intra-block trial order effects (each block had trials 1-39), so the estimated marginal means for each block were normalized to reflect the midpoint of that block (e.g., trial 19.5). As evident in Figure 6.2, absolute fixation durations were shorter in Block 2 (marked) relative to Block 1 (unmarked), for each word position except the definite article, $F_{in}(1,23.05) = .493$, $MSE = 16532.53$, $p = .490$. To explore the possibility that some of this reduction in reading time on the Block 2 critical sentences was due to practice effects that persisted across block (rather than due to anaphor status marking, per se), additional ANOVAs were done which included absolute trial presentation number as a covariate (Block 1 trials number: 1-39 (block 1); 40-78 (block 2), so marginal means for each block were normalized to a common presentation position - the midpoint of the whole experiment (i.e., trial 39.5). These

ANOVAs, F_{bet} , are presented in Table 6.2(b). In this latter case, any remaining effect of blocking should be due to the marking manipulation rather than practice. In both sets of analyses, participant was included as a random factor to control for the variance associated with individual differences in reading speed.

Effects of Noun Type. The relative time course patterns are similar across block for the different noun types, and there were no significant interactions of noun type with block in any of the analyses. For each word position, planned comparisons (using the Bonferroni adjustment) were made among the four different noun types, and the 95% confidence intervals are shown by the error bars in Figure 6.2. As indicated by the ANOVA results in Table 6.2, there was no main effect of critical noun type on the first-pass reading time for any word in the sentence other than the critical noun itself ($ps > .05$). There was a significant effect of noun type on first-pass time on the noun. In particular, first-pass noun time was shortest on antecedent-match anaphors, and longest on metaphoric anaphors, and this contrast was the only pairwise-significant difference in first-pass noun gaze ($p = .033$). As in Experiment 2, the first-pass processing time for non-anaphoric nouns did not differ from that of anaphoric nouns (all $p > .05$).

Because the critical noun types were not perfectly matched in terms of their frequency of use in English (see Table B1 in Appendix B), an additional analysis was conducted which included the noun's corpus frequency as a covariate. A main effect of noun type on first-pass noun gaze remained after controlling for frequency, $F_{in}(3,74) = 3.28$, $MSE = 141296.2$, $p = .026$, and the pairwise significance pattern was unaltered (i.e., metaphoric anaphors were read more slowly than antecedent-match anaphors, p

=.005, but there were no other significant differences).

Thus, as expected, the antecedent-match anaphors produced the shortest first-pass gaze on the noun itself. However, this first-pass reading time for the antecedent-match nouns was not significantly shorter than the time spent on literal category anaphors (nor on non-anaphors). Further there was no increase in processing speed for any subsequent word position (or for the sentence overall) for sentences with antecedent-match in comparison to the other three noun types, in fact, the absolute values were often slightly higher for antecedent-match anaphors.

Effect of Block. As seen when comparing Figure 6.2(a) with 6.2(b), the absolute fixation durations were significantly shorter in Block 2 (marked) relative to Block 1 (unmarked), for each individual word position except the definite article. In the F_{bet} ANOVAs, however, which control for cross-block practice effects, the effect of block disappeared for the words between the noun and the final word in the sentence ($p > 0.05$, for both first-pass and net fixations). However, an effect of block/markings remained, in the opposite direction, on the first-pass times on the noun, $F_{bet}(1,85.76) = 6.65$, $MSE = 1831813.68$, $p = .012$ and on the last word in the sentence, $F_{bet}(1,179.365) = 13.49$, $MSE = 1533329.13$, $p = .000$. Thus, after controlling for trial order ‘practice’ effects, readers spent longer on the noun (and the final word in the sentence) when information about the anaphoric status of the noun was encoded in its font colour (Block 2) than when the status of the noun was not explicitly encoded in its colour (Block 1). Note that this effect cannot be attributed to the presence of colour per se, because the noun phrases in the unmarked block were also in colour, but the colour in

the unmarked block had no informational content. The font colour of the definite article, “The”, in the Block 2 also indicated anaphoric status of the ensuing noun, however readers did not spend longer processing the definite article in the second block, $F_{bet}(1,175.77) = .072$, $MSE = 10309.80$, $p = .789$, perhaps because the processing of such a short frequent word was highly automatized, and the information about anaphoric status was also available in the font colour of the noun itself.

Figure 6.2(a): Time Course (ms) for Critical Sentences in Block 1 (unmarked).
The lower portion of each bar represents first-pass fixations and the total height of the bar represents net fixations [Error bars are 95% CIs].

Figure 6.2(b): Time Course (ms) for Critical Sentences in Block 2 (marked). The lower portion of each bar represents first-pass fixations and the total height of the bar represents net fixations [Error bars are 95% CIs].

Regressions to Critical Noun

Overall, there was an effect of noun type (antecedent-match, category, metaphoric, non-anaphor) on the likelihood of regression, $F_{in}(3,71.9) = 7.51$, $MSE = 1.60$, $p = .000$. In particular, among the anaphoric nouns, participants were more likely to make regressions to the metaphoric anaphors (68%) than the category anaphors (44%, $p = .000$) or antecedent-match anaphors (46%, $p = .001$). . Unexpectedly, however, there

was no reduction in regressions for antecedent-match anaphors relative to category anaphors ($p = .542$). Overall, the likelihood of regressing to an anaphoric noun was inversely correlated with the LSA anaphor-referent similarity, $r(df=991) = -.065$, $p=.041$. There was no significant interaction of anaphor type with block, $F_{bet}(2,46.6) = 0.18$, $MSE = 0.38$, $p = .839$.

To determine if readers who did not regress had spent more time during first-pass processing, an analysis of first-pass noun times was redone which included whether or not a regression occurred as a factor (for a discussion of regression-contingent analysis see Altmann, Garnham, & Dennis, 1992; Altmann, 1994; Rayner & Sereno, 1994). First-pass time on the nouns was not related to (predictive of) whether the reader would subsequently regress, $F_{bet}(1,33.42) = .026$, $MSE = 3836.46$, $p = .873$. Thus, there was no evidence of a trade-off pattern between first-pass time and likelihood of regression.

Participants returned to re-read the critical noun on a smaller percentage of trials when the anaphoric status of the noun was marked than when it was unmarked, $F_{in}(1,23.1) = 29.53$, $MSE = 10.97$, $p = .000$. This effect was present both when the critical noun was anaphoric (32.9 % vs. 53.9 %, $p = .000$), and when the critical noun introduced a new referent (35.9 % vs. 50.2%, $p = .000$),. However, this main effect of block did not persist in the analysis to extract practice effects across blocks, $F_{bet}(1,174.0) = 0.25$, $MSE = 0.06$, $p = .619$. Thus, the overall reduction in regressions for Block 2 may have been due to an overall tendency for faster reading and fewer regressions as participants progressed through the experiment (even within the unmarked block).

However, there was marginally significant interaction of anaphoric status (non-anaphoric vs. anaphoric) with marking, $F_{bet}(1,23.1) = 3.08$, $MSE = .590$, $p = .093$. In particular, there was a larger drop in regressions in the marked block for anaphoric nouns than for nouns introducing new referents. Thus, marking the status of the noun especially reduced regressions for anaphoric nouns, presumably because marking reduced the likelihood that anaphors would be initially misinterpreted as new referents. In contrast, marking did not provide as substantial a processing advantage for nouns that actually did introduce new referents.

Subsequent Noun Fixations

There was a significant effect of noun type on subsequent noun fixations, $F_{bet}(3,74.3) = 5.43$, $MSE = 65601$, $p = .002$. In particular, the mean subsequent noun fixations were significantly larger for the metaphoric anaphors than for the other three types of nouns (there were no significant differences among antecedent-match anaphors, literal category anaphors and non-anaphors). There was no effect of block and no interaction. Note that in this analysis, the mean times included fixations of 0 ms for trials in which the reader did not regress.

The following analyses of subsequent noun fixations are based upon the subset of trials for which regressions to the critical noun did occur. There was a marginally significant trend for participants to spent less time re-reading the critical noun when its anaphoric status was marked (Block 2), than when it was unmarked (Block 1), $F_{in}(1,42.6) = 3.32$, $MSE = 7447007.57$, $p = .075$. However, this trend disappeared when absolute trial number was included as a covariate to control for practice effects,

$F_{bet}(1,217.3) = 0.57$, $MSE = 190095.14$, $p = .812$. As in Experiment 2, among the anaphoric nouns there was a trend for readers to spend less time re-reading literal (category) anaphors than metaphoric anaphors. However, here the effect of noun type on the time spent re-reading anaphoric nouns was not significant, $F_{in}(3,132.2) = 1.051$, $MSE = 194361.74$, $p = .372$, $F_{bet}(3,128.1) = 0.92$, $MSE = 179529$, $p = .435$. Finally, there was no interaction of noun type with block, $F_{in}(3,85.7) = 1.05$, $MSE = 204580.80$, $p = .428$, $F_{bet}(3,84.6) = 0.98$, $MSE = 190920.64$, $p = .404$.

Discussion

The results in this experiment replicate and extend the time course findings from Experiment 2. Overall, the results are consistent with a unified memory-based account of preliminary referent retrieval. In particular, there was no evidence of an anaphoric bias during preliminary referent assignment. First-pass processing time on the noun was not inflated (e.g., due to a vain discourse search, and an anaphoric bias override) for non-anaphors versus anaphors. Nor did first-pass processing time differ on the word immediately following a non-anaphor versus an anaphor (were one to hypothesize that preliminary referent assignment may spill over).

Anaphor Type

Readers did spend longer (first-pass) on metaphoric anaphors than on antecedent-match anaphors. It is possible that this processing time advantage in the antecedent-match case may be explained by lexical priming - the antecedent-match term was previously mentioned in the context, whereas the metaphoric term gets little priming from the context.

As in Experiment 2, remainder processing time and regression likelihood were higher for critical sentences with metaphoric anaphors than sentences with either type of literal anaphor. This pattern is consistent with the memory-based model, which predicts that the lower the semantic overlap between the intended referent and the anaphor, the more likely that a new referent schema will be retrieved and assigned instead. The contrast between the two types of literal anaphors, however, was not as clear-cut.

Sentences with antecedent-match anaphors took as long to process overall, and involved as many regressions to the anaphor, as sentences with category anaphors. How can this pattern be reconciled with a memory-based account?

It is possible that correct preliminary referent retrieval was somehow as likely in the case of antecedent-match anaphors as for category anaphors. At first blush, this suggestion seems incompatible with the memory-based theory, since antecedent-match anaphors (e.g., terrier-terrier) obviously have a stronger semantic association with the referent and seemingly should increase its activation more than a category anaphor would (e.g., terrier-dog). However, some readers may have a tendency to mentally represent referents at the basic category level (Rosch, 1978) when they first encounter the antecedent. Thus, the antecedent “terrier” might be initially encoded as a dog. Hence, this mental referent could resonate more strongly with the category anaphor “dog”, than the anaphor “terrier”, despite the fact that the latter anaphor matches the original antecedent term. Thus, if readers often represent referents at the basic category level, it is not surprising that category anaphors would be comparably effective at prompting the retrieval of the referent. Antecedent-match anaphors might have had

short first-pass processing times, not because they always prompted more rapid referent retrieval than category anaphors, but because lexical access was facilitated by the presence of the exact same noun earlier in the text.

There is also a possible pragmatic explanation for post-noun delays and regressions in the antecedent-match case: the repeated name penalty (e.g., Gordon, 1993). Roughly, the use of an antecedent-match anaphor may strike the reader as pragmatically inappropriate or excessive (e.g., violating the Gricean maxim of quantity) when a pronoun or less specific noun may have sufficed to identify the referent. For example, studies (e.g., Gordon et al., 2004) have shown an increase in post-anaphor delays and regressions for sentences like (13a) which involve an antecedent-match anaphor (dog-dog), in contrast to sentences like (13b) in which a pronoun anaphor was employed (dog-it).

(12) John took his terrier for a walk.

(13a) The dog barked at the birds.

(13b) It barked at the birds.

These results are typically interpreted according to centering theory (Grosz, Joshi & Weinstein, 1995), which argues for a soft constraint that anaphoric references to an entity in discourse focus should be realized by a pronoun. This suggested constraint is violated by the antecedent-match anaphor in (13a), and thus, under their account, some post-hoc processing cost is incurred. Strictly speaking, however, this traditional characterization of the repeated name penalty, as a contrast-in-kind between nouns and pronouns, does not shed light on why the antecedent-match noun anaphors in the

present study should cause more post-noun delays and regressions than the category noun anaphors -- both are nouns and both would therefore violate the suggested constraint, were the antecedent in discourse focus (which it typically is not in the present study).

The Informational Load framework of Almor (1999; see also Areil, 1990; Sperber & Wilson, 1995) however, is more general, and suggests that there are gradations of specificity among noun anaphors. Under his view, when the referent is highly accessible (in focus), an anaphor with high anaphor-antecedent similarity (i.e., one which contains an excess of 'information') may be more difficult to process than an anaphor with lower anaphor-antecedent similarity. Again, however, the referents in the present study were rarely in discourse focus. It remains possible that pragmatic assessments of the anaphor may have contributed to post-noun processing delays -- however, they fall outside the current focus on the preliminary referent retrieval mechanism, and there is little information and consensus on the underlying procedural details and time course of such effects (though they necessarily occur after a referent has already been retrieved). There is no evidence in the present study that any pragmatic delay-inducing effects impacted first-pass processing of antecedent-match anaphors (see also van Gompel, Liversedge, & Pearson, 2005).

Anaphor Status Marking

Although the marking manipulation was somewhat artificial, and readers reported that they were not always consciously aware of processing the colour of the critical noun in the marked block, marking did seem to exert some effect on processing. In

particular, readers spent longer than expected during first-pass processing of the noun in Block 2, when its colour conveyed its anaphoric status (after controlling for trial order practice effects). Furthermore, comprehension question performance was selectively improved for anaphoric trials in the marked block. However, marking did not seem to selectively benefit preliminary performance on metaphoric trials (i.e., there was no interaction of noun type with marking to suggest that the likelihood or duration of reanalysis was particularly reduced), despite the fact that cloze data in Experiments 1 and 3 suggest that there was the most room for improvement (preliminary misclassifications) on the metaphoric trials. Further, although comprehension question performance improved for the other two types of anaphors, it did not improve for metaphoric anaphors in Block 2.

What, if anything, do such marking effects suggest about the nature of the preliminary referent assignment process? While certainly not conclusive, the evidence is consistent with a memory-based resonance model of preliminary referent retrieval that is cognitively impenetrable. In all, marking did not seem to elevate the level of first-pass comprehension for the metaphoric anaphors to that for *spoken* metaphoric anaphors (Stewart & Heredia, 2001). Recall that metaphoric but not literal associates were facilitated immediately following their metaphoric terms. It is possible that this marking manipulation was too artificial and novel to simulate any possible prosodic cues about anaphoricity that listeners might leverage. However, even if such cues are present in speech, they might not be expected to exert an immediate impact on preliminary referent retrieval if an automatic memory-based mechanism is in play.

Instead, listeners may have benefited not from prosodic cues about anaphoric status, but about prosodic cues that a figurative interpretation is intended. There is evidence that such cues are present in idioms (Baum & Titone, 2005). If such cues generalize to metaphoric anaphors, they could, in theory, influence the lexical access stage, which precedes referent retrieval.

Table 6.2(a): Time Course ANOVA Results with Trial (1-39) as covariate to control for within-block ‘practice’ effects

Region	Fixation	Noun Effect (df1 = 3)				Block Effect (df1=1)				Interaction (df1=3)			
		df2	F	MSE	p	df2	F	MSE	p	df2	F	MSE	p
“The”	1 st pass	71.0	1.57	17696	.204	23.1	0.48	7596	.506	69.5	.593	5439	.622
	Net	69.9	3.59	117468	.018	23.1	8.88	493228	.007	69.6	.385	11765	.764
Noun	1 st pass	70.3	7.16	929456	.000	23.0	16.5	10526551	.000	69.5	1.57	274753	.204
	Net	70.1	11.9	241282	.000	23.0	41.86	841308	.000	69.5	.556	200109	.468
Remainder	Net	71.6	1.88	774349	.141	23.0	53.12	33351189	.000	69.4	.903	322727	.444
Noun+1	1 st pass	72.3	1.07	67747	.365	23.0	10.04	1071326	.004	69.4	.591	62103	.623
	Net	71.7	0.65	88474	.584	23.0	28.4	219098	.000	69.4	.445	48388	.772
Noun+2	1 st pass	73.0	2.06	170370	.113	23.1	1.92	124931	.179	69.5	.902	50484	.445
	Net	73.0	3.34	321194	.024	23.1	18.6	1568507	.000	69.5	1.86	128723	.149
Noun+>2	1 st pass	81.7	1.08	37747	.362	30.0	0.00	43211	.999	51.6	0.31	15645	.820
	Net	82.4	6.63	54774	.000	33.8	10.91	508811	.002	51.6	.124	9562	.945
Last	1 st pass	73.6	1.16	93284	.332	23.1	19.81	355436	.000	69.4	1.34	136180	.268
	Net	73.5	1.11	89865	.350	21.1	19.35	185250	.000	69.4	1.35	133735	.266
Sentence	Net	69.5	7.32	36057010	.000	23.0	77.37	882566810	.000	69.1	.031	151314	.993

NB: bold *p* values are significant (alpha .05); net characters on screen and word/region length were also included as covariates

Table 6.2(b): Time Course ANOVA Results with Trial (1-78) as covariate to control for cross-block ‘practice’ effects

Region	Fixation	Noun Effect (df1 = 3)				Block Effect (df1=1)				Interaction (df1=3)			
		df2	F	MSE	p	df2	F	MSE	p	df2	F	MSE	p
“The”	1 st pass	69.8	1.29	14260	.285	175.8	0.07	738	.789	69.5	0.54	4936	.656
	Net	71.9	3.67	121005	.016	232.0	0.70	28781	.402	69.6	0.38	11687	.706
Noun	1 st pass	70.3	7.23	938982	.000	85.8	6.65	1831813	.012	69.5	1.58	275307	.203
	Net	72.9	11.62	2820499	.000	103.6	2.67	1070496	.106	69.6	0.89	206479	.453
Remainder	Net	70.1	1.78	737663	.160	148.2	10.28	3921220	.000	69.4	0.89	320238	.449
Noun+1	1 st pass	72.3	1.07	67359	.368	184.3	0.26	18141	.608	69.4	0.59	36701	.623
	Net	71.9	0.92	3820654	.213	101.5	2.16	971326	.091	69.6	0.89	206479	.453
Noun+2	1 st pass	73.0	2.06	170508	.112	388.0	0.13	8441	.717	69.5	0.91	50621	.443
	Net	73.0	3.34	320534	.024	337.2	0.17	12916	.682	69.5	1.84	129062	.148
Noun+>2	1 st pass	81.7	1.08	40688	.363	245.3	0.76	30495	.384	51.5	0.31	50968	.819
	Net	32.4	6.60	362541	.000	408.2	3.43	189824	.065	51.6	0.12	9102	.949
Last	1 st pass	73.6	1.16	93507	.331	179.4	13.50	1533329	.000	69.4	1.31	133206	.277
	Net	73.5	1.12	90174	.348	171.2	13.24	1515473	.000	69.4	1.32	130800	.276
Sentence	Net	69.5	7.32	36057010	.000	23.0	77.37	882566810	.000	69.1	0.03	151314	.993

NB: bold *p* values are significant (alpha .05); net characters on screen and word/region length were also included as covariates

CHAPTER 7: The MEMBRAL Computational Model

The theoretical and empirical evidence thus far suggest that preliminary referent assignment to nouns is fast and fallible, and does not involve a strategic search or elimination process to arrive at an introductory (new-referent) interpretation. Such evidence is consistent with a memory-based framework for text processing (Gerrig & O'Brien, 2005). Thus, the memory-based theoretical framework informed the design of the proposed computational model.

In this chapter I will: i) clarify the nature and scope the task being modeled, ii) enumerate the criteria for a cognitively plausible model; iii) outline the functionality of the proposed model; and iv) present simulation results for the performance of the model on the experimental stimuli (Appendix A) and compare these results with human performance data. I will also discuss how the model provides a procedural basis for additional empirical effects in the literature, beyond those observed in the present paradigm (e.g., distance effects). Finally, I will provide an overview of two other recent cognitive models that have addressed the processing of noun anaphors (Budiu & Anderson, 2004; Lemaire & Bianco, 2003), and I will compare them with the proposed model, and show that they do not fulfill several of the desired criteria.

What process is being modeled?

At the outset, it pays to clarify what the proposed model is (and isn't) a model of. Nominally, the process being modeled is the rapid, automatic, and pre-conscious referent retrieval process that is triggered when a reader encounters a noun. In the model, what is retrieved from memory in response to a referring expression is a mental

referent (in particular, the most active referent in memory), which may either be specific (e.g., your own dog "Rover") or generic (e.g., a dog prototype). Thus, the operation of the model will predict which specific or generic referent will be initially retrieved:

- i) for a noun
- ii) at its time of encounter
- iii) by automatic memory-based processes.

That said, the local process at the time the noun is encountered is in some sense the tip of the iceberg. The present model must also simulate the processing of the preceding context (serially, in real-time units) in order to establish the activation levels of the various generic and specific referents in memory just prior to when the critical noun is encountered.

The automatic, unified operation of the model is conducive to processing both literally-intended and metaphorically-intended nouns. Further, as is the case with a human reader, the model tackles the classification problem: that is, it can process nouns intended as either anaphoric or introductory (introducing a new referent into the discourse). Ideally, when 'reading' nouns intended as anaphors, the model will retrieve the specific intended referent from the discourse, and when 'reading' nouns intended as introductory, the model will retrieve an appropriate generic referent prototype. That said, it is important to emphasize that the model is not an artificial intelligence model that aims to achieve optimal performance, but rather it is a cognitive model that aims to reflect the sometimes error-prone performance of actual human readers. As evidenced

by the cloze data in Experiments 1 and 3, and the regressions in Experiments 2 and 4, the preliminary referent retrieval process may sometimes result in the retrieval of the wrong referent. In particular, a noun intended as an anaphor may result in the retrieval of a generic referent rather than the specific intended referent from the discourse. For each stimulus context and critical noun phrase, the proposed model will predict the likelihood that the wrong referent will be retrieved during preliminary processing.

Such erroneous preliminary assignments may, in practice, be subsequently rectified by the reader. The non-local processes by which a reader subsequently detects that an erroneous preliminary assignment was made are, at present, beyond the scope of the model, though it may be expanded to include this capacity. The model does, however, simulate the overt re-analysis process (regression to the critical noun to retrieve another referent), which is triggered at sentence end if the wrong referent was retrieved.

As mentioned in the introduction chapter, the focus of the present model is on preliminary referent retrieval for *nouns*. It is possible, as suggested by Gernsbacher (1989), that the same memory-based process may apply to other types of referring expressions like proper names (e.g., "John") and pronouns (e.g., "he", "it"). However, non-noun referring expressions fall outside the current focus. In contrast to nouns, which can be introductory or anaphoric, pronouns are almost exclusively anaphoric (Kintsch, 1998), so the classification problem is side-stepped, and the problem space is typically constrained to referents already mentioned in the current discourse. Pronouns are also subject to special syntactic constraints (e.g., Principles A and B in Binding Theory: Chomsky, 1981; see also Sturt, 2003), which further constrain their

interpretation, so there is reason to assume that there may be a special-purpose process for associating referents to pronouns (see Emond, 1997 for a pronoun processing model in ACT-R). The present focus however, is on referent retrieval for nouns (which may be literal or metaphoric, and anaphoric or introductory) so the critical referring expressions in my stimuli are nouns⁷. In future, a pronoun processing module, could be merged into the present noun processing model.

Criteria for a Cognitive Model

Reference assignment is addressed by several non-cognitive Natural Language Processing algorithms (e.g., Bean & Riloff, 1999; Vieira & Poesio, 2000), however, they involve multiple passes forward and back through the entire text, and do not directly speak to the development of a *cognitive* model of *on-line* processing. Thus, to set the stage for the proposed model, I outline the general criteria for a psychologically plausible model of referent retrieval for nouns. These criteria can then serve as a framework for articulating the properties of the proposed model, and for comparing and contrasting it with two other models in the literature.

Incremental Processing: The system should be able to make use of information in the discourse as it becomes serially available, roughly word-by-word.

Appropriate Representational Units: In line with criterion 1, the unit of analysis

⁷ However, the contexts do contain some names and pronouns. The model faithfully follows the same processing steps for non-nouns as for nouns (e.g., productions for lexical access, spread of activation, and referent retrieval), however, non-noun referring expressions are tagged with the referent ID (e.g., "John=1 went to the store and he=1 bought a soda. The drink...was cold"), so the appropriate referent is retrieved. As is the case in referent retrieval for nouns, the retrieval latency will be influenced by the activation level of the referent.

for the processing system must be smaller than a complete sentence. The cognitive task being modeled is to associate a definite noun phrase with a mental representation of its referent, so the system must, minimally, have representations (though possibly atomic ones) for individual nouns and individual potential referents, in order to model the task of associating the former to the latter.

Context Sensitivity: The model should allow and account for the influence of: (i) preceding context sentences, and (ii) preceding parts of the current sentence on the processing of the current noun. To account for the processing of a particular anaphor, the model ideally should also, as a pre-requisite, model the processing of the prior context, to get the cognitive system into the appropriate state -- so that relative accessibilities of specific and generic referents reflect the influence of prior context.

Real-Time Simulation: Ideally, the reading process should be simulated in real-time units (e.g., ms vs. 'cycles'). Cognitive effects like priming (spreading activation) are time-sensitive and subject to decay. The system should ideally model such memory effects (fluctuations in activation over time), and should simulate processing of the text at a representative reading rate (e.g., 150 ms/word).

Appropriate Problem Space: Models necessarily abstract away from some level of detail, but care must be taken to ensure that the problem space in the model is not artificially skewed or trivially sparse. The characterization of the task and the representation of the problem space (e.g., range of possible referent choices) should be sufficiently rich to reflect the interpretation challenge facing a real reader, and so permit key types of possible errors. Because readers do not know a priori whether a current

definite noun phrase is anaphoric or introductory, the explicit or implicit determination of this property is part of the referent assignment process, and thus should be part of the model. Ideally, the model should be able to operate on and discriminate between (though not always accurately) both anaphoric and non-anaphoric definite noun phrases. In this vein, the memory system must be populated with not only the correct referent, but also the other particular referents encountered in the discourse (and in, in principle, in prior life), and with generic referent prototypes.

At the time of this writing, I am unaware of a previous model that fulfills all of these criteria. Other models have alleged to 'process' texts with literal and metaphoric anaphors (e.g. Budiu & Anderson, 2004; Lemaire & Bianco, 2003), however they fulfill some of these desired criteria but not others, notably the criteria having referents as representational units, and having an appropriate problem space. In view of these criteria, the proposed model was implemented in ACT-R as will be discussed below. Then the model description and simulation results will be presented. Finally, I will contrast the present model with these previous models in the literature in terms of the above criteria.

Implementation Architecture: Python ACT-R

The criteria outlined previously motivated the choice of ACT-R (Anderson et al., 2004) as a suitable cognitive modeling platform for the present model. In particular, the proposed model was actually implemented using the python extension of ACT-R (Stewart & West, in press), because, as will be discussed below, the memory-based character of my model motivated my colleagues and I to enhance the functionality and

psychological plausibility of the ACT-R architecture itself (Pyke & West, 2007; West, Stewart, Pyke, Emond, 2006). The ACT-R architecture is predicated on a Unified Theory of Cognition (Newell, 1987) - that is, on the belief that our performance on a vast range of tasks can be accounted for parsimoniously by a common cognitive system whose operation relies on general-purpose mechanisms and principles. As such, ACT-R is compatible, in spirit, with the memory-based text-processing framework (Gerrig & O'Brien, 2005), in which referent retrieval for nouns occurs in virtue of general-purpose memory mechanisms, rather than requiring a special-purpose search process.

As mentioned in Chapter 2, the ACT-R architecture has proved conducive to modeling a vast range of cognitive tasks. The data accrued by the ACT-R modelling community has provided theoretically and empirically motivated constraints for the values/ranges for its key adjustable parameters (which reduces concerns about arbitrariness and generalizability of a model's design). In the present model, key parameters are set to the recommended 'universal' defaults (e.g., noise=0.3, decay=0.5, production time=50 ms).

ACT-R supports two concurrent levels of functionality:

- (i) a production system that carries out sequences of situation-driven productions (i.e. if-then rules) that serve as the procedural building blocks (steps) for various tasks.
- (ii) a dynamic memory system which contains the various representational units (called chunks) upon which the productions act. The memory system can simulate the real-time fluctuation of activation of a representational unit. For example, the effect of recency and frequency of use on the activation level of representation R_i is quantified by

A_{Ri} , where t_j is the time since use j , and d is a constant whose default value is .5 (Anderson et al, 2004). Activation level is thereby affected by all prior uses of the referent, including those that occurred before the current discourse.

$$(E1) \quad A_{Ri} = \ln \sum_{j=1}^N t_j^{-d}$$

Thus, ACT-R affords sufficient functionality to fulfill almost all of the desired model criteria: i) incremental processing -- cognitive operations (productions) can be executed at a rate of one per 50ms, so a model can perform several operations (e.g. recognize a word, spread activation, retrieve referent) during the typical reading time for each incoming word (150 ms); ii) representational units – the modeller can specify the types and numbers of representations in memory, for example: noun-chunks, generic referent chunks, and specific referent chunks, iv) problem space – the architecture allows the modeller to define the task (set of productions) and populate the memory (set of chunks) as appropriate, iii) real-time simulation – the rate at which productions are performed is paced to reflect the time for the mind to perform a single simple operation, and the memory system can simulate the activation fluctuations of the chunks (referents) over time according to (E1).

However, the traditional ACT-R architecture has a limitation in its ability to model a desired type of context sensitivity – in particular, it does not lend itself to modeling any cumulative priming effects from prior stimuli (here, words) on the processing of the current stimuli. The content that is currently in the system's focus (e.g., the anaphor noun-chunk) can spread activation to other representations in memory (e.g., referent-

chunks). However, in traditional ACT-R, the boost in activation received by any memory chunk (referent) from a stimulus (word) is instantly removed as soon as the stimulus is no longer in focus. This limitation of the architecture undermines its ability to model cumulative priming effects from prior words on the activation levels of referents in memory.

To clarify this architectural limitation, consider how a lexical priming effect might be modelled using traditional ACT-R. In a priming task, a prime word such as "dog" is presented, followed by a target word such as "cat". The agent must then make a lexical decision about (or name) the target word. At the point "dog" is being processed (is in focus), the ACT-R architecture can be programmed to allow activation to automatically spread from the current word (dog) to an associated word/concept (cat) in memory. However, when the new stimulus (i.e., the word "cat") is presented, any influence (activation boost) afforded by the preceding stimulus immediately ceases, because it is no longer in focus – and thus the preceding word, "dog", could not facilitate the processing of the current input, "cat". However, in traditional ACT-R, there is a work-around: the effects of priming can be modeled by placing the stimulus immediately preceding the current stimulus in a special mental buffer. The architecture then functions as if both words are simultaneously in focus. So, at a given moment, activation will automatically spread from both the contents in the special buffer (say, the preceding stimulus, dog), and the contents in focus (the current stimulus, cat). Hence, the previous stimulus is treated more or less as if it is still simultaneously present, and thus it can facilitate the processing of the current stimulus.

While sufficient to model facilitation in the simple two word (prime-probe) paradigm, this traditional work-around of retaining the immediately preceding stimulus (word) in a special buffer does not scale up nicely to model cumulative effects of priming from several preceding words or sentences. Working memory is a limited resource (Miller, 1956), so only so much information can be plausibly 'buffered' by a reader (and consequently by a cognitively plausible model). Furthermore, empirical evidence suggests that, even within a sentence, readers do not cognitively 'buffer' prior words in this manner, because if they did, they would not so frequently have to physically regress their eyes back to re-read these prior words (e.g., the eye-tracking research in the current studies). Finally, when prior stimuli are 'buffered' in traditional ACT-R, the temporal character of the stimulus stream is lost. It is as if everything in the buffer (say, the preceding M words or sentences) is processed simultaneously, because all elements (words) in the buffer have equal spreading activation influence, even though some elements were input more recently than others (e.g., the word preceding the current word, versus a word several words back).

Thus, the limitations of the traditional ACT-R architecture motivated my colleagues and I to introduce a persisting spreading activation functionality into python ACT-R, without the requirement for indefinite buffering of prior stimuli (Pyke, West & LeFevre, 2007; West, Stewart, Pyke, Emond, 2006; see also Stewart & West, in press). The capability for limited buffering is still available, however, it is not leveraged in the present model. Instead, the spread of activation of an input stimulus decays gradually rather than instantaneously when the stimulus is removed. Thus, local spreading

activation effects can cumulate (e.g., the successive stimuli "towel", "waves", "sand".... may cumulatively prime "beach"), but due to decay, the activation contributions will effectively be weighted according to order and timing of the preceding inputs. The mathematical activation equations are discussed more fully later on. We believe that this modification augments rather than circumvents the psychological plausibility of the architecture, and will significantly enhance its general ability to model tasks involving streams temporal streams of stimuli, including non-linguistic ones.

Model Operation

The following sample story will be used to illustrate the operation of the model. As in Experiment 4, there are three possible types of anaphors (antecedent-match, category, and metaphor) for the target referent (the terrier).

(14) John had a terrier.

(15) It was white and shaggy.

(16) He took it to the park.

(17) The terrier/dog/mop barked at the birds.

The memory system contains two types of referents: i) specific people or objects the agent is acquainted with from earlier in the present discourse (e.g., [R2: white, shaggy, terrier]), and more generally, from any prior life experience (e.g., one's own mother), and ii) generic referent templates (e.g., the generic dog referent, [G:dog]).

In the proposed model, the local referent retrieval process in play while reading the anaphoric noun is basic and strictly memory-based: When 'reading' the noun, activation automatically spreads from the noun to all referents throughout memory (both generic

and specific), and then the most active referent is retrieved. Thus, in effect, all referents in memory (both generic and specific) simultaneously compete for retrieval. The anaphor term determines the relevant (most active) generic referent in play. If the anaphor is "mop", the intended referent [R2: shaggy, white, terrier] competes not just with the other specific referents from the story (and from prior life experience) but also with the generic referent [G1:mop], which will receive a large boost in activation due to spread from the corresponding noun "mop". For the version of the story with the anaphor "dog", the relevant generic competitor is [G:dog], which will automatically receive considerable spreading activation from the noun "dog", even though the noun is intended to refer to the specific referent [R2:white, shaggy, terrier].

Any impact of prior discourse processing will be entirely mediated by its lingering effect on the activation levels of the various referents in long-term memory (e.g., lingering activation spread from prior words). Thus, this memory-based process does not navigate through some complex hypothetical discourse structure representation (though it is possible one exists). Rather the local, preliminary referent retrieval process is simply sensitive to raw referent activation levels, and in principle, a non-discourse (generic or specific) referent is as eligible for retrieval as any specific discourse referent, provided that it becomes the most active referent in memory due to: i) spread from the noun, and ii) any lingering activation effects of prior context processing. Thus, to model the preliminary referent assignment for a particular noun, it is also necessary to simulate the processing of the preceding discourse, but only so far as is necessary to approximate its effect on the immediately pre-noun activation levels of the

various key referents in long term memory (i.e., the specific people and objects in the discourse, and the relevant generic competitor).

Factors Affecting the Activation Levels of Referents

As mentioned above, each referent's accessibility during preliminary noun processing is affected by two components: i) the activation boost/spread from the noun term currently being processed; and ii) its 'context dependent' pre-noun activation level. Below I outline how the discourse context, up until the critical noun, influences the activation levels of all the generic and specific referents in memory.

1. Spread of Activation from Pre-Anaphor Words. Just as the anaphor resonates with, or spreads activation to, referents, I assume that such activation spread generally occurs as each content word in the discourse is encountered. The activation boosts received by referents may persist (as do lexical priming effects, e.g., Collins & Loftus, 1975) even when the reader progresses on to the next word. Although such priming effects decay, the sequence of pre-anaphor words may exert a cumulative effect on each referent's activation. Thus, in (18), the spreading activation from "dove" will increase the activation level of the mental referent [G:pool], even before reaching the noun "pool", which will further increase its activation.

(18) John dove into the...pool.

2. Recency and Frequency of Use of the Referent. These general factors affect any mental representation's accessibility. Evidence indicates that the further back that an antecedent is (in sentences, and consequently in time), the more challenging it is to process the anaphor (e.g., Clark et al., 1979; Duffy & Rayner, 1990; Levine, Guzman,

& Klin, 2000). A referent referred to many times in a text, and/or referred to in the sentence preceding the critical anaphor should be more active, *ceteris paribus*, than a referent mentioned only once several sentences back.

3. Sentence Wrap-Up Effects. Just and Carpenter (1980) suggested that integrative processes occur at sentence end, which is why readers spend tend to spend relatively longer on the final word in each sentence. These wrap-up processes may result in sentence-end activation effects (Balogh, Zurif, Prather, Swinney, & Finkel, 1998). Probe studies suggest that a referent mentioned early in a sentence may also produce facilitation effects at sentence end (e.g., Dell et al., 1983; McKoon & Ratcliff, 1980). In the MEMBRAL model, the processes at sentence end result in an activation boost of the specific referents mentioned in the sentence.

4. Discourse Dependent Associations. In addition to pre-existing associations like those we are modeling with LSA, discourse dependent associations may be formed in memory. Spread of activation through such associations may produce intermittent (yet cumulatively significant) activation contributions to an intended referent during pre-anaphor processing. For example, each sentence (and proposition) in a discourse may contain several referents. In Dell et al. (1983, see also McKoon & Ratcliff, 1980), two referents that have appeared in a common sentence are called *companions*. The comprehension process appears to forge a lasting association between companion referents in memory, possibly during sentence wrap-up, such that when a referent is subsequently encountered, its companion(s) from prior sentences also become re-activated right away, and to a comparable degree (Dell et al., 1983). Thus, in the

proposed model, whenever a referent is mentioned, its companions are also boosted in activation, thereby making them more accessible as referent candidates for up-coming nouns.

Note that for the stimuli in the present research, the three versions of each story are identical up until the anaphor (antecedent-match, category, metaphor). Thus the *pre-anaphor* activation levels of the intended referent and other specific referents in the story are the same across story versions. For example: [R1:John]: 2.4, [R2:terrier]: 1.4, [R3:park]: 2.8. However, the *post-anaphor* activation levels of the referents will vary depending on the anaphor.

Overview of Operation

0. LTM is seeded with generic referents for various discourse concepts, including, importantly the antecedent concept and also (if different) the anaphor concept. For the "mop" version of our example story, the simulated reader's memory will be seeded with [R1:John], [G1:terrier], [G2:park], [G3:birds] and [G1:mop]. The referent [R700:DaVinci] is also seeded in memory as a representative of the legion of referents a real person's memory would already contain prior to reading the current story.

1. Words of the story are then processed serially.

2. Each content word (e.g., noun, adjective, verb) automatically spreads activation to all referents in memory, both specific and generic, according to the LSA similarity between the word and the referent.

3. If the current word is a referring term (noun, name, pronoun), a referent is retrieved from memory. The referent retrieved for each noun will be the most active one

(after the spread of activation from the noun), be it specific or generic. In the latter case, the generic referent is used to create a new specific referent to associate with the noun. The assigned referent is automatically boosted in activation (in virtue of its current use), and activation also spreads from it to its companion referents from previous sentences (see below). A fan effect is applied, so if the current referent had n companions in the previous sentence, the weight factor for activation spread to each companion will be $1/n$.

4. At the end of a sentence during wrap-up, the sentence's referents become reactivated and mutually associated.

Some sources of activation furnish a relative advantage to specific discourse referents, to enable them to win out over a generic referent. Sentence wrap-up effects and discourse-specific associations (companion spreading) augment the activation of specific referents mentioned in the discourse, whereas generic referents receive no activation boost from these effects. The frequency and recency of usage of a specific referent in the discourse will selectively bolster its activation level (relative to its generic counterpart). Thus, the activation level attained by the generic competitor will depend on the automatic spread of activation it receives from the pre-anaphor words and from the anaphor term itself. In contrast, specific referents mentioned in the text will also get activation boosts each time they are retrieved and assigned, and also during integrative sentence wrap-up processes.

Activation Equations

For our example story, besides the original descriptor used to introduce a referent

(e.g., "terrier"), other properties of the referent may also be explicitly mentioned in a text (e.g., white and shaggy). Activation from the word being read spreads automatically to each attribute explicitly mentioned in the text for each referent [R1:terrier,white,shaggy]. Each attribute in turn contributes activation to the referent representation as a whole. The association strengths that quantify the activation spread from the word (concept) being read are estimated by LSA associations. Thus, in the present model the weight of activation spread from the word currently being read (word w) to a referent, R_1 , via its M explicit attributes (a_k) is:

$$(E2) W_{wR1} = \sum_{k=1}^M LSA < w, a_k >$$

In traditional ACT-R the spreading activation boost received by chunks (referents) in memory is a temporary additive boost that is instantly removed when the source of the activation spread (e.g., the current word) is no longer in focus. Thus, in traditional ACT-R the net activation of a referent could be expressed as (E3), which reflects referent usage effects (t_j is the time since the j^{th} mention of the referent in the text) plus spreading activation from only the current word (or buffered words).

$$(E3) A_{Ri} = \ln \left(\sum_{j=1}^N t_j^{-d} \right) + \text{Spreading Activation from Only the Current Input}$$

However, our augmented version of the ACT-R architecture (Pyke, West, & LeFevre, 2007a, 2007b) allows for a more gradual decay that follows the form of the standard base-level usage equation. Thus, the activation of R_i is at a given point in time depends on both usage effects and the cumulative (though decaying) spreading activation contributions from prior words in the context. Thus, the activation of

referent R_i , A_{Ri} , can be expressed in (E4), where t_w is the time since the w^{th} spread (i.e., the w^{th} word) which had spreading weight W_w for the current chunk (referent R_i).

$$(E4) \quad A_{Ri} = \ln \left(\sum_{j=1}^N t_j^{-d} + \sum_{w=1}^M W_w t_w^{-d} \right)$$

The post-spread activation (1 ms after the spread) of a referent is given by equation (E5), where d is the adjustable activation decay parameter in ACT-R. The present model adheres to the recommended default value of $d=0.5$, thus (E5) becomes (E6).

$$(E5) \text{ Post Spread Activation} = \ln \left(e^{\text{preBoostActivation}} + W * (.001)^{-d} \right)$$

$$(E6) \text{ Post Spread Activation} = \ln \left(e^{\text{preBoostActivation}} + W * 31.6227766 \right)$$

When the current word is a noun, the referent with the highest post-spread activation, as determined by (E4), will be retrieved. In contrast to a traditional spreading activation contribution that is additive, (E4) is non-linear. The result is that a given weight of spread will produce a relatively modest increase in net activation for referents that are already relatively high in activation, but a bigger absolute boost in activation for referents that are initially low in activation. Thus, when an object is mentioned out of the blue, like a "catapult", the relatively dormant generic referent for catapult in memory, [G:catapult], gets a big activation boost from the noun "catapult", allowing the generic referent to achieve a higher activation than the referents that may otherwise have been initially more active in virtue of having been mentioned more recently, in the current discourse.

Regression and Memory-based Reanalysis

For the experimenter, regressions to the critical nouns provided a useful indication

of possible error in preliminary referent assignment, however, what function do such regressions serve for the reader? I hypothesized that a regression to the noun might serve to re-engage the memory-based referent retrieval process, at a time when the target referent's accessibility may be increased due to priming from post-anaphor words. For example, consider a hypothetical critical sentence (20) containing the anaphorically-intended noun "mop" which refers to a white shaggy terrier introduced in (19).

(19) John opened the door for his white shaggy terrier.

(20) The mop barked at John.

When the noun "mop" is encountered, the model simulation predicts that the most accessible referent will often be the generic mop referent [G:mop] not the intended referent [R3:terrier,white,shaggy]. However, the remainder of the sentence will contribute considerable activation to the intended referent. For example, "barked" will spread activation to [R3:terrier,white,shaggy]. Furthermore, the mention of John will also boost [R3:terrier,white,shaggy] through companion associations, because [R1:John] and [R3:terrier,white,shaggy] were mutually present in a previous sentences.

Thus, I considered the hypothesis that overt re-analysis might simply consist in re-engaging the automatic memory-based process, and that such a process might be sufficient to automatically retrieve the right referent 'in retrospect'. To explore this hypothesis, an extended version of the simulation was done. In the extended version of the simulation, a coherence check was conducted at the end of the critical sentence. Although the mechanisms and criteria for such a coherence check are themselves

beyond the current scope, it is a given that readers 'somehow' detect a problem, and that detection often results in a regression. In the present simulation, a regression to re-read the noun was automatically (omnisciently) triggered at sentence end if the wrong referent had been retrieved during preliminary processing. The incorrect referent initially retrieved was inhibited, and the critical noun was then 're-read' - engaging the same memory-based process which occurred during first-pass read: activation was spread to (un-inhibited) referents in memory, and the most active referent was retrieved. The efficacy of this simple memory-based re-analysis is discussed below with the simulation results.

Simulation Data

Method

The model was used to predict which referent would be retrieved *during the preliminary processing* of the critical nouns in my experimental stimuli (Appendix A). To obtain percentage accuracy rates for retrieving the correct referent, the model was run 100 times on i) the 30 non-anaphoric stimuli and, ii) on each version (antecedent-match, category, metaphor) of each of 42 anaphoric stimuli stories. The noise parameter in ACT-R was enabled to simulate the variability introduced by different 'participants' (noise = 0.3).

There are several types of data produced by the model. First, the model predicts the percentage of times that the incorrect referent will be retrieved during preliminary processing. These percentages were compared with human data from my eye-tracked reading studies (Experiment 2 and 4), in which regressions to reanalyze an anaphor are

indicative of the incidence of preliminary error. Second, the model predicts the relative times to perform preliminary processing for the four different types of critical nouns (antecedent-match, category, metaphoric, and introductory). These times are compared to the times that human readers spent on the nouns during first-pass processing (Experiments 1 and the unmarked block of Experiment 4). The model also predicts the relative frequency of the types of errors made during preliminary processing for anaphoric nouns (misclassify-as-new errors vs. discourse-distractor errors) and non-anaphoric nouns (misclassify-as-old errors vs. discourse-distractor errors). The types of preliminary errors made by the model will be compared with the types of errors evidenced in the human cloze experiments (Experiment 2 and 3).

Results

Performance on Non-Anaphoric Critical Nouns. The noun "television" in (22) exemplifies a non-anaphoric (introductory) noun after a context sentence (21):

(21) Jo-Anne lost a chess game with her grandfather.

(22) The television distracted her.

The model achieved near-perfect performance (96%) in retrieving the appropriate referent for the critical introductory nouns in the present stimuli. That is, after spread of activation from the noun, the appropriate generic referent, here [G:television], became the most active referent in memory (vs. other generic referents and specific discourse referents). The retrieved generic referent was then used to initialize a new referent to represent the particular television being introduced into the discourse. The mean pre-noun activation levels of the appropriate, generic referents were low (generic referent:

0.86 vs. specific discourse referents: 2.20), because in general the appropriate generic referent had received little activation spread from the prior words in the context, due to relatively low semantic association. However, after the spread of activation from the critical noun, the activations of the appropriate generic referents surpassed those of the discourse referents (generic referent: 3.53 vs. specific discourse referents: 2.75).

Performance for Anaphoric Critical Nouns. Figure 7.1 depicts the mean pre and post anaphor activation levels of the intended discourse referent (R) and the relevant generic competitor (G). As shown in the figure, the intended referents reached their highest absolute activations upon 'reading' antecedent-match anaphors. Note that the mean pre-anaphor activations of the intended referent (R) are the same across the three anaphor types (2.5), because for all three versions of each story, the text is identical up until the critical noun anaphor. However, after the anaphor is encountered, activation levels vary across the three contexts. Ignoring the activation levels of the generic competitor for the time being, the figure shows that the intended referent achieves the highest absolute activation level after an exact match anaphor (e.g., terrier) and the lowest activation level after a metaphoric anaphor. However, it is not absolute activation level but activation level relative to the generic competitor that determines whether or not the intended referent will be retrieved.

Figure 7.1: Mean activation levels for the intended referent (R) and the generic competitor (G), before (lower part of bars) and after activation spread from the anaphor (full bars).

Figure 7.2 indicates how frequently the model retrieved the incorrect referent during preliminary noun processing (misclassify-as-new errors). Among those simulation trials in which an incorrect referent was retrieved, this incorrect referent was almost always the generic referent corresponding to the noun, rather than another discourse referent (the latter occurred in less than 1% of incorrect retrievals). Figure 7.2 also shows the human regression likelihoods, which were averaged across the two eye-tracking studies (Experiments 1 & the unmarked block of Experiment 4). The pattern of simulation results is highly congruent with the regressions made by readers. Across the 126 stimuli (42 contexts * 3 anaphor types), the human likelihood of regression was correlated with the likelihood of correct referent retrieval in the simulation, $r(124) = -.308$, $p = .000$ (2-tailed), and with the post-anaphor activation levels of the intended referent in the simulation, $r(124) = -.333$, $p = .000$ (2-tailed). As

seen in Figure 7.2, the model performance coincides closely with the human regression patterns for the metaphoric and category anaphors. However, for antecedent-match anaphors, humans regressed on more trials (50%) than would be expected in light of the minimal number of preliminary referent retrieval errors predicted by the model (<10%). As touched on in the discussion of Experiment 4, the regressions in the antecedent-match case may result not only from preliminary errors but may have been inflated due to a pragmatic appropriateness check: the anaphor seems over-specified (e.g., Almor, 1999; Kennison & Gordon, 1997). Such pragmatic processing is beyond the scope of the present model. However, these simulation results suggest that only a small proportion of the regressions in the antecedent-match case result from preliminary errors.

Figure 7.2: Comparison of MEMBRAL's Preliminary Retrieval Inaccuracy (misclassify-as-new errors) with Human Performance Data.

Performance on Anaphoric Nouns after Re-analysis. If the wrong referent was initially retrieved, re-analysis (re-reading) of the critical noun was triggered at sentence end. Such re-reading simply re-engaged the memory-based retrieval process that occurred during preliminary processing, however the incorrect referent is inhibited and the intended referent may be more active than it was on initial encounter because it has since been primed by the post-anaphor words. After allowing for such reanalysis, the overall performance accuracy of the model in retrieving the intended referent for the anaphoric stimuli increased from 56% to 80%. Thus, memory-based processes may play a productive role during overt reanalysis as well as in preliminary referent retrieval. However, human readers may also employ other types of processes during reanalysis.

Duration of Preliminary Processing. In the MEMBRAL model, spread of activation plus referent retrieval takes 50 ms (ACT-R's default interval for a processing step) plus a variable retrieval delay, which is dependent on the activation level of the referent that is retrieved. In ACT-R, the retrieval time, T , is characterized by the equation (E7), where A is the activation level of the referent at the time of retrieval, and F is a latency factor that can be set by the modeler, however, MEMBRAL uses a typical default value ($F = 0.35$).

$$(E7) T = Fe^{-A} \quad (\text{Anderson et al., 2004})$$

Figure 7.3 shows the mean duration of preliminary noun processing for the different types of critical nouns in the simulation. Note that these preliminary processing times also include processing steps that occur prior to referent retrieval, such

as lexical access and syntactic parsing. The inner workings of these precursor processes are beyond the scope of the current model, however MEMBRAL includes placeholder productions for these steps, so that they contribute a total of 100ms to preliminary processing time. In total, there is only a 4 ms difference in between the shortest and longest mean processing times, for the antecedent-match anaphors and non-anaphoric nouns respectively.

Figure 7.3 also presents the human data on first-pass noun duration from the two eye-tracking studies (Experiment 1 and the unmarked block of Experiment 4). As shown in the figure, the pattern of results in the simulation is congruent with the pattern of results in the human data. For human readers, there will be some variability in lexical access times for the various nouns as a function of factors like corpus frequency (Rayner, Ashby, Pollatsek, & Reichle, 2004). Because such variability in lexical access (prior to the referent retrieval phase) is not captured in the present model, the human data presented in Figure 7.3 are the estimated marginal means, after controlling for noun length and frequency by including these factors as covariates. As an additional note, the absolute values for the first-pass noun fixations in Experiment 4 are inflated relative to those in Experiment 2 and those in the simulation because readers were required to press a key to display the remainder of the sentence – thus, the durations in Experiment 4 included an extra delay component associated with this mechanical key-press. However, within each data set (Experiment 2, Experiment 4 and the simulation), the relative pattern is consistent: there was no significant difference in preliminary processing time across the different types of critical nouns, including non-anaphoric

ones.

Note that in Experiments 2 and 4, readers' first-pass noun fixations were used to gauge the duration of preliminary referent assignment (i.e., the immediacy hypothesis was applied, Just & Carpenter, 1980). One could argue, however, that preliminary assignment may spillover, and may vary with noun type as predicted by a special-purpose account. If spillover does occur, it would be difficult to differentiate preliminary processes from early reanalysis processes. However, there are additional reasons to support the view that preliminary assignment may not fluctuate much in duration and may be completed during first-pass analysis: i) even if the noun does not reactivate the intended referent, the noun being read would activate its generic referent, so a reader always has enough information to assign a referent upon reading the noun; ii) the MEMBRAL simulation demonstrates that a referent assignment can be made in the typical first-pass noun time; iii) in the present research there were no significant differences in first-pass times on the word following the noun to suggest spillover or support a special-purpose account (e.g., longer times on non-anaphors vs. anaphors); and iv) readers did not report having more difficulty (which may reflect duration) in processing non-anaphoric versus anaphoric definite nouns (Experiment. 3).

Figure 7.3: Simulated and Actual Preliminary Processing Times for Critical Nouns

In all, the simulation results in conjunction with the human data support the claim that manipulating anaphor word choice primarily affects the accuracy rather than the duration of preliminary processing. This pattern is compatible with the unified memory-based account. In particular, for any noun, the presence of its corresponding generic referent prototype places a cap on the time spent in preliminary referent retrieval. The generic referent prototype competes with the intended referent for retrieval, so as soon as a manipulation allows the generic referent to achieve a higher net activation than the intended referent, it is the activation of the generic referent rather than the intended referent that determines retrieval latency.

Discussion

The model simulation provided estimates of the pre and post-critical-noun activations of referents in memory, and likelihood that the correct referent would be retrieved during the preliminary processing of the noun. The simulation results provided a good agreement with the human data patterns from the preceding experiments in terms of both the duration and accuracy of preliminary processing. In particular: i) preliminary errors (misclassify-as-new) made by the model were more frequent for metaphoric than literal anaphors; ii) the likelihood of human regression to each particular noun was correlated with the likelihood of incorrect preliminary retrieval in the simulation; and iii) there were no notable differences in preliminary processing time for the different types of critical nouns.

The model serves as an existence proof that a unified memory-based model of preliminary referent assignment can provide a good fit with human data, and can suffice to process and effectively discriminate between introductory and anaphoric nouns. When the noun is encountered, the spreading activation and retrieval processes are 'blind' to whether a referent is generic or specific and in the latter case it is also blind as to whether the specific referent is from the current discourse or some other prior discourse/experience. However, the effects of the prior discourse context (recent referent usage) provide some activation head-start for specific discourse referents (vs. generic referents), without the need for a special-purpose search that is restricted to discourse referents.

However, like human readers, the model may sometimes mistake an anaphoric

noun to be introducing a new referent. The model, like the cloze study participants, almost never made the reverse type of mistake (i.e., a misclassify-as-anaphor error), that is, to associate one of the specific discourse referents to a noun intended to introduce a new referent. Thus the performance of the model provides is congruent with both the quantitative and qualitative patterns in human performance.

Besides modeling the local memory-based process of preliminary referent retrieval, the model simulates the effects of prior context on the activations of all the generic and specific referents in memory. Because the design of the model was informed by key empirical findings beyond those in the present research, the model's operation is consistent with such prior findings, and provides a procedural basis for them. In particular, when a referent is mentioned in a text, it will become active rapidly after reading the noun (in virtue of spread of activation from the noun, and a usage boost if it is retrieved). Further, at the time a referent is mentioned, the referent's 'companion' referents that were mentioned together with it in earlier sentences will also rapidly receive some activation (Dell et al., 1983). Finally, the referents mentioned in a sentence will also be active at sentence end (Balogh, Zurif, Prather, Swinney, & Finkel, 1998; Dell et al., 1983; McKoon & Ratcliff, 1980). In the present model, the companion effect occurs due to spread of activation through the discourse-dependent associations, which are forged among the referents in a sentence at sentence end. Such sentence-end wrap-up processes cause a re-activation of the sentence's referents (though the nature of the integration processes and coherence checks are not known). Finally, although the distance between the (last) mention of the target referent and the critical

anaphor was not systematically manipulated in the present stimuli, the usage component of activation operates according to equation (E1) so the greater the time ($t_{j=N}$) since the last mention of the referent, the lower its pre (and post) anaphor activation will be, and the greater the chance that a generic competitor referent may be retrieved instead. Thus, the more text between the last mention of the target referent and the intended anaphor, the greater the likelihood of preliminary error in referent retrieval (consistent with Levin, Guzman & Klin, 2000), which in turn will in turn lead to increased processing time for the anaphoric sentence (e.g., Clark, 1979) if re-analysis is subsequently triggered.

Finally, this memory-based model also offers a possible functional rationale for the regressions that readers often make to the critical anaphors. Regressions may serve to re-engage the same memory-based retrieval process in play during preliminary assignment, after the intended referent has been primed (spreading activation) from reading the post-anaphor part of the sentence. **Contrast with Other Models**

In this section, I contrast the capabilities and characteristics of the MEMBRAL model with two recent cognitive models of language processing in the literature: Budiu and Anderson (2004) and Lemaire and Bianco (2003). These models bear some high-level similarities with the current model. First, like the present model, these are working computational models, and they provide a unified procedural treatment of literal and metaphoric content. However, they differ from each other and from the present model in the nature of the processing and of the problem space (i.e., the nominal task and memory representations). Further, they both make use of LSA (though in

fairly distinct ways), and the present model followed from their lead in that regard. Finally, like the present model, the Budiu and Anderson model is also implemented within an ACT-R architecture (though not the python one). Because of these important, though limited, similarities it is important to clearly distinguish these models from the present model⁸.

Although these other models have been applied to the processing of texts involving both literal and metaphoric noun anaphors, they do not address the issue of referent assignment per se. That is, there are no individual referents (specific or generic) represented in either model, and thus, in 'processing' anaphoric nouns these prior models never actually associate a particular referent with the noun. This may seem surprising, since co-reference is at the core of anaphoricity, and referent assignment generally is an essential aspect of processing a noun regardless of its anaphoric status. So, what do these models do instead?

Lemaire and Bianco's (2003) Model

The Lemaire and Bianco (2003) model, henceforth the L&B model, is strongly based on a model by Kintsch (2000), and is deeply entrenched in the LSA framework (Ladauer & Dumais, 1997). Thus it pays to briefly recap LSA here. Recall that in LSA, each English word (lexical concept) is represented as a point in a multi-dimensional space (see <http://lsa.colorado.edu/>). Note that this LSA semantic 'space' has been constructed based on a vast English corpus, and is independent of any

⁸ I thank an anonymous reviewer of Pyke, West & LeFevre (2007a) for making this recommendation.

particular model and its stimuli. Hypothetically, the lexical concept STAR might be represented as the point (0,.3,.4,.7,.5,.3,,4, .8, .2). The individual digits and dimensions do not represent features in any symbolic sense, and have no independent meanings – in this sense they can be considered sub symbolic. The relative proximity of lexical concepts in this space depends, inter alia, on their frequency of co-occurrence in usage (see Landauer, Foltz, & Laham, 1998). The end result is that concepts that are near each other in LSA space tend to be cognitively associated.

Similarity is quantified by the cosine of angle between the two points (vectors), and a distance of .2 is considered "close" by L&B. For example, the points for the concepts GALAXY, ASTRONOMERS, CONSTELLATION, and BRIGHTEST are all within the .2 similarity radius of the point representing the concept STAR. In LSA space, the different senses of a word are not represented with separate points, so that a word like "pig" would correspond to a single point with associations with (similarity distances) <PIG, BARN>=0.33 and, <PIG, RUDE>=0.17, corresponding to its literal and metaphoric senses, respectively. In general, similarity values tend to be smaller for metaphoric associations (Budiou & Anderson, 2004). Thus, similarity or association strength between pairs of lexical concepts can be estimated using LSA, which is the role of LSA in MEMBRAL – that is, to quantify the spread of activation from each content word in the story (processed serially), to each referent in memory. In MEMBRAL, the activations levels of referents are also influenced by the passage of time (decay). The role of LSA in the L&B model, however, is very different than in MEMBRAL.

In the L&B model, the whole prior context up until the critical word (e.g., noun

anaphor) is represented as a single point in LSA space. In so doing, word order and syntax information are lost, so that the point for "Mary gave John a rose" would be the same as the point for "John gave Mary a rose", since these sentences are composed of the same set of words. In contrast to MEMBRAL, the L&B model does not conduct any serial processing of the context preceding the critical noun. Instead, this LSA-point representation of the context is provided to the model. Such a representational format affords no means to individuate the different referents present in the context. Rather than finding a referent for the current noun, the goal of the L&B model is to construct a 'meaning' for the noun that is influenced by this LSA-point representation of the prior context. Operationally, the 'meaning' constructed for the noun is (the centroid of) a set of 5 lexical concepts points that are 'near' to both the LSA point for the noun and the LSA point representing the prior context.

Thus, starting from the position in LSA space that represents the lexical concept for the critical noun, (e.g., STAR) the process searches the nearby LSA space for neighbouring points (concepts), and checks if any are also neighbours (i.e. within .2) of the point representing the context. Recall that the positions of concepts in LSA space are based on a vast English corpus (see <http://lsa.colorado.edu/>), and that this space was generated prior to and independent of any particular model and any particular stimuli. The L&B process successively expands the search radius around the critical noun point (STAR) until it finds 5 concepts that are within .2 of the point representing the context (note, however, that these concepts may be further than .2 from the critical noun's point, depending on how many times the search radius needed to be expanded to meet

the quota of 5 concepts within .2 of the context point). The number of times the search radius has to be expanded is hypothesized to reflect the processing time for that particular noun (though the time units in the model are abstract cycles, rather than milliseconds). Note that the 5 concepts that the model selects to comprise the noun's 'meaning' are not (necessarily) themselves explicitly present in the context.

The same process applies to both literal and metaphoric nouns. When the critical noun was metaphoric, their simulation indicated that the search radius had to be expanded many more times -- in order to come across 5 concepts within .2 of the context point -- than when the critical noun was literal. As such their model is consistent with evidence that sentences with metaphoric anaphors tend to take longer to process than sentences with literal anaphors (e.g., Gibbs, 1990). However their simulation is inconsistent with my present findings that delays for metaphoric sentences are due to subsequent re-analysis, rather than to delays in initial processing. Furthermore, it is entirely unclear how to evaluate the 'meaning' (a mid-point of 5 points in LSA space) generated by the L&B model. The operation and representations can shed no light on how (or whether) co-reference and anaphoric status are established. Further, in LSA space each lexical concept corresponds to a unique point, but this 'lexical space' does not readily allow for different specific referents of the same type to be discretely represented (e.g., [R:my dog] vs. [R:my neighbour's dog]), nor for specific referents of a given type to be discrete from the corresponding generic prototype, [G:dog].

Thus, although the L&B model does provide a unified mechanistic treatment of

literal and metaphoric content, it does not fulfill the previously outlined criteria regarding a suitable problem space or representational units to capture the task of referent assignment, nor does process the context incrementally or make predictions in real-time units. As such, there is effectively no procedural similarity between the L&B model and the present model. The present model makes use of LSA only to estimate the pairwise association strengths between concepts in order to predict the spread of activation between them in memory. However, I remain skeptical about the psychological reality, in the L&B model, of representing whole paragraphs as single points in LSA space because this representation abstracts away from word order and temporal effects and does not permit referents in the context to be individuated, and consequently associated with an incoming anaphor. In all, the L&B model does not constitute a model of referent assignment, and consequently, I argue, does not address the fundamental nature of anaphor processing. Next, the Budiu and Anderson (2004) model is discussed.

Budiu and Anderson's (2004) Sentence Comprehension Model

Like the present model, the Budiu and Anderson (2004) model, henceforth the B&A model, was implemented in an ACT-R cognitive architecture, albeit in the traditional one rather than in the python one. However, the B&A model does not associate an input noun with a referent in memory. Rather, the nominal task of the B&A model is to 'comprehend' a sentence as a whole. The model's memory is initially seeded with a repertoire of possible sentences (propositions), and the task of

'comprehending' an input sentence roughly amounts to retrieving/recognizing that sentence from among the model's repertoire of pre-stored sentences. Clearly this recognition type of sentence comprehension process is incompatible with current linguistic theories about compositionality, which emphasize readers' ability to process entirely novel sentences. However, although it is cognitively implausible when billed as a means of comprehending the current sentence, the B&A model can be re-cast to account for the means by which readers retrieve related (rather than synonymous) propositions from memory. I will first outline the general operation of the model and then discuss and critique its application to sentences containing noun anaphors.

As stated, the model attempts to comprehend an input sentence by retrieving the matching sentence from its pre-stored repertoire of sentences in memory. To qualify as the correct 'interpretation' of the input sentence, however, a repertoire sentence need not be identical to the input sentence, so the repertoire can be somewhat short of exhaustive. For example, B&A's system can associate the active version of a sentence with the passive version of the same sentence. For example, input sentence (I1), could be associated with repertoire sentence (R1).

(I1) The college students were taught by professors.

(R1) Professors taught the college students.

Furthermore, the words/concepts in the input sentence and the repertoire sentence selected as its 'interpretation' need not be identical. As in the present model, LSA values are used to estimate the association strengths between concepts. Thus, input sentence (I2), like (I1) could also be associated with repertoire sentence (R1), despite

the fact that (R1) contains the concept “professors”, whereas (I2) contains the term “lecturers”.

(I2) Lecturers taught the college students.

Similarly, B&A suggest, due to some similarity (association) between chicken and coward, their model would sometimes be capable of 'comprehending' a metaphoric sentence (I3), by associating it to interpretation (R3), given that this latter proposition should be present in the pseudo exhaustive repertoire.

(I3) The chicken...didn't go on the roller coaster.

(R3) The coward....didn't go on the roller coaster.

How does the B&A model operate? Like the proposed model, the B&A model operates incrementally in that as each word is serially encountered some processing is conducted – however, the processing does not involve referent assignment, rather the retrieval of the 'best' repertoire sentence is re-attempted after each input word/phrase. In particular, when the first phrase of the input sentence is encountered, the system attempts to retrieve a sentence from its repertoire that contains a similar phrase. What, mechanistically, determines which repertoire sentence is selected? Operationally, the procedure is much like the memory-based procedure for retrieving a referent in the proposed model, however here the representational units are repertoire sentences rather than referents. Activation spreads from the current word/phrase ("The chicken") to all the possible sentences in the repertoire. The amount of activation spread to a particular repertoire sentence will be determined by the LSA similarities between the current input word ("chicken") and each of the concepts in the repertoire sentence (e.g., coward and

roller coaster). After this spread of activation, the most active repertoire sentence is retrieved (but only if it exceeds a threshold value, which apparently must be re-adjusted for each stimulus set that the simulation is applied to, in order to furnish a fit with human data). The retrieved repertoire sentence is the model's current hypothesized/anticipated interpretation of the sentence being read. As each additional word/phrase of the input sentence is 'read', the retrieval of a repertoire sentence is re-attempted, in light of the (activation spread by) new word/phrase⁹. And a different repertoire sentence may be retrieved.

In light of the procedural detail, it seems that the ability of the B&A model to retrieve the correct repertoire sentence for a metaphoric input sentence like (I3) above is a model artifact that is conditional on the absence of key competitors from the memory repertoire. Even if we accept that the 'interpretation' for an input sentence is the matching (most active) sentence from an, in principle, exhaustive repertoire, in practice, the necessarily truncated repertoire implemented for a simulation must faithfully include the interpretation candidates (repertoire competitors) that would become especially active in light of that input sentence. In particular, the repertoire should, in principle, also contain the linguistically possible literal proposition (R4), which, given the procedural operation of the model, would invariably receive more spreading activation from the input (I3) sentence than the intended metaphoric interpretation (R3).

(R4) The chicken(literal)....didn't go on the roller coaster.

⁹ Because the B&A model was implemented in traditional ACT-R the prior words (up to 3 phrases worth) in the input sentence are buffered, and also simultaneously spread activation to all repertoire sentences along with the current word.

Thus, I suggest that the B&A model will only successfully retrieve the intended interpretation for a metaphoric sentence from the repertoire, if the possible literal interpretation is conspicuously excluded from the model's memory repertoire. In this way, even within their framework, which characterizes sentence interpretation as recognizing the sentence from a pre-stored repertoire, the model has under-represented the problem space facing a reader. In contrast, in the proposed model, memory will contain the generic chicken referent to compete with the referent corresponding to the metaphoric interpretation.

How has the B&A model been applied to (predict the processing time for) sentences with noun anaphors? Budiu and Anderson ran a simulation on the anaphoric stimuli of Onishi and Murphy (1993), which are similar to those in the present research. For each story, the context sentences (and only context sentences) were a priori included in memory to serve as the repertoire. Right away, this impoverished repertoire is theoretically incompatible with B&A's nominal role of the repertoire as a relatively exhaustive repository of 'interpretations' for any possible input sentence. Such a starting point for memory is also incompatible with the fact that memory contains propositions (and concepts, referents) beyond those in the current discourse. Furthermore, because the context sentences are all provided simultaneously to the model's memory, the activations of these representations in memory do not reflect any temporal effects that would be associated with a serial processing of the context.

However, the model does then operate on the critical anaphoric sentence incrementally, and attempts to 'comprehend' it by retrieving a repertoire (context)

sentence from memory that matches it. At face value, this endeavour will be futile because the critical anaphor sentence will not be identical to, or synonymous with, any of the context (repertoire) sentences, which constitute the entire memory of the model. However, mechanistically, at each juncture (word/phrase) of the input sentence, activation spreads from all the words read so far to each of the repertoire (context) sentences in memory, and the retrieval at that juncture will be 'successful' as long as the activation level of one of the repertoire sentences reaches a minimum threshold¹⁰. Alternately, when the current word is processed, none of the repertoire candidates may reach the minimum threshold of activation despite the activation spread from the current (and preceding words). Whenever the net spread of activation from the portion of the anaphoric sentence 'read' so far is insufficient to raise the activation of any context sentence above threshold, the process is explicitly programmed to temporarily stall (the "bug" production) before it resumes operation to read the next word, and attempt repertoire retrieval again. Any such stalls will contribute to increased processing time for the critical sentence. And such stalls occur more frequently when processing metaphoric sentences, because a metaphoric noun is set to spread less activation to the context sentences than a literal noun. In this B&A simulation, the spreading activation contribution from an anaphoric noun to the repertoire sentence(s) containing the antecedent is set to .27 for all metaphoric critical nouns and to .45 for all literal critical nouns. In contrast, MEMBRAL uses the individual association strengths for each

¹⁰ This threshold was set to (-3.5) for the simulation Budiu and Anderson ran on the anaphoric stimuli of Onishi and Murphy (1993), however, this threshold was adjusted depending on the stimulus set.

particular noun within each anaphor type, to enable me to predict differences in retrieval accuracy within each anaphor type (e.g., literal and metaphoric).

Thus in the B&A model there is greater likelihood of a stall penalty for sentences beginning with metaphoric anaphors. As subsequent words in the input sentence are read, they will successively add clout to the net spread of activation from the input sentence to each repertoire sentence. By sentence end, the model may or may not have been able to retrieve any particular repertoire sentence, but B&A only report simulation data on their model's net time to 'process' the anaphoric sentence. Due to the level at which B&A set their retrieval threshold to induce their stall penalties, processing times will be longer for sentences with metaphoric anaphors than literal anaphors.

However the cognitive reality of the B&A model is open to question in that it is unclear in what sense the model's operations can be said to have 'processed' a sentence (anaphoric or otherwise) during said processing time. Certainly co-reference relations are never explicitly established by their model. Instead, the final 'interpretation' for the input sentence as a whole is nominally a (final) retrieved repertoire sentence as a whole. Even then, it is often possible that no repertoire sentence may ever be selected, because the anaphoric sentence may not be very similar to any of the context sentences. In such cases, the B&A model will just chug through the input sentence word-by-word, retrieving nothing and thereby incurring programmed stall penalties, but ultimately will not form an 'interpretation' at all for the input sentence to show for this processing time.

In general, the explicit stalls in the B&A model (which occur after each word when no repertoire sentence reached retrieval threshold) are somewhat suspect. It is unclear

what independent functional justification such stalls have, other than to selectively induce delay, for delay's sake, in the processing of input sentences commencing with metaphoric anaphors, to thereby be consistent with Onishi and Murphy's (1993) human data on net sentence processing times. But the time course of the stalls is not compatible with current data. In particular, a stall in the model is especially likely at the time a metaphoric noun is processed, but the present research indicates that readers do not in fact locally stall when reading metaphoric anaphors.

Instead, for human readers of sentences with metaphoric anaphors, delays are contributed by (successful) reanalysis, not successive retrieval failures. The idea of retrieval failure (which in turn instigates the programmed stalls) in the B&A model is also suspect. The cloze data in the present research indicates that readers are able to form an interpretation (though possibly not the intended one) for a metaphorically-intended noun prior to reading additional words in the sentence. In reality, memory is populated with a plethora of propositions (and concepts, referents). So for any possible noun (e.g., "mop"), memory will contain propositions, and generic and specific referents for that concept, which the noun can bring to mind (regardless of whether it may also bring to mind a context proposition). A model which results in a failure to retrieve any 'interpretation' or information from memory on encountering a noun is implausible. When human readers process an anaphoric sentence incrementally, they do not fail to find any initial interpretation for the noun, but rather, they seem to succeed in finding the wrong interpretation or referent. Thus the memory system's problem space in the B&A simulation is not properly populated to model the formation of the

right types of wrong interpretations.

In summary, there are fundamental differences between the proposed model and the B&A model. In terms of the model criteria, the B&A model is implemented in ACT-R and thus has some incremental and real-time processing (however, in the B&A model, this incremental processing is not applied to the context sentences, only to the critical anaphoric sentence). However, the B&A model does not have referents as representational units, and does not ultimately associate a noun to a referent. Nor is the memory system populated to enable the model to make common types of incorrect interpretations.

Synopsis of Alternative Models

The fact that the L&B and B&A models have been applied to 'process' the same types of stimuli as the proposed model (i.e., texts with literal and metaphoric noun anaphors) belies the fact that the problem spaces (i.e., processes, products, and goals) of the three models are extremely different. The MEMBRAL model predicts which referent will be retrieved for a noun during preliminary referent retrieval (i.e., it provides a procedural account of the referent retrieval process and makes predictions about its precision). In contrast, these prior models were designed to predict the *duration* of preliminary processing (L&B) and/or sentence processing (B&A).

In terms of net sentence processing time, the performance of these prior models is superficially consistent with human data: longer net processing times for sentences with metaphoric anaphors than those with literal anaphors. However, even as such, they predict that (some) delay will occur during the processing of the critical noun, which is

inconsistent with eye-tracking evidence (Experiments 2 and 4). More importantly, it is not clear that the nature of the processing they propose is cognitively plausible, nor is it clear what exactly has been accomplished at the end-state of their processing, and how it constitutes an interpretation (accurate or otherwise) of the anaphor. Neither assigns a referent to the noun, nor explicitly addresses the classification problem (i.e., is the noun introducing a new referent or is it anaphoric). Rather than assigning a referent to the noun, the L&B model attempts to construct a 'meaning' for the noun that is operationalized as an abstract point in a multi-dimensional LSA space. B&A attempt to find a 'meaning' for the anaphoric sentence as a whole by recognizing it from a set of pre-stored sentences in memory. This above comparison emphasizes the importance of having cognitively-motivated criteria for the evaluation of a model (e.g., appropriate representational units) that extend beyond a fit with a single measure of human performance, especially one that underdetermines the nature of the process as much as net sentence reading time.

Limitations of MEMBRAL

As with any computational model, the MEMBRAL model also has some limitations. These limitations can be broken into two types: i) limitations in scope; and ii) implementation-level limitations. These limitations are outlined, in turn, below.

Limitations in scope

The MEMBRAL model is by no means being billed as a complete model of language comprehension, and is not yet equipped to perform some functions outside the purview of its primary task scope (i.e., preliminary referent retrieval for nouns), for

example: referent assignment for names and pronouns; syntactic-level processes, and global coherence checks and other post hoc pragmatic processes that operate at the proposition level. The introduction of such additional functionality may be a matter for future work, as will be discussed in the final chapter.

Implementation-level limitations

There are some limitations surrounding the use of the LSA estimates for semantic association strengths.

i) Cognitive reality: LSA values are based on analyses of language corpora, and may not always accurately reflect cognitive association strengths, nor do they capture any individual differences (which was simulated with random noise in MEMBRAL). That said, LSA was nonetheless the 'best game in town' as an independent source of association strength estimates, given such strengths cannot be directly observed, and that obtaining estimates using a subjective rating poll was logistically prohibitive, and would also be of questionable validity. Recall that MEMBRAL uses these LSA estimates to quantify the spread of activation between two concepts, and that each content word in each story spreads activation when it is encountered, to all the concepts/referents in the story, and their generic counterparts. Thus, with approximately 15 content words per story, there are over 100 pairs of concepts per story, and therefore over 4200 concept pairs in the stimulus set of 42 stories! In view of such considerations, the use of LSA estimates seemed justifiable, and furthermore resulted in performance that provided a good fit with the human data.

ii) Learning: The association strengths are currently static in the MEMBRAL

model, and represent the participants' pre-experimental associations. However, in practice, the strength of the cognitive association among concepts (and among words, and between words and concepts) may change over time according to the agent's experience. In particular, repeated exposure to a metaphor (e.g., The surgeon was a butcher) should strengthen the association pathways between the concepts, and between the word "butcher" and the bad surgeon concept (see also Budiu & Anderson, 2006). In the present model, stronger association strengths would provide the mechanistic basis for how familiarity would facilitate the processing of metaphoric anaphors. The model could be adapted to include the effects of the prior frequency of exposure to a metaphor on association strengths.

Besides the limitations regarding the LSA association strength estimates, there are also some limitations imposed by the ACT-R architecture. In the MEMBRAL model, referent selection is arbitrated by activation levels, and only one representation will be retrieved (i.e., the most active). However, some researchers have suggested that two (or more) referents might achieve comparable activation levels (e.g., Cristea & Dima, 2001), which might necessitate a subsequent non-memory-based selection process, and/or might prompt a postponement in assignment altogether until more of the sentence is available. For example, "The toy" in (23) is technically ambiguous when the reader encounters it, because there are two toys mentioned in (24).

(23) The teacher gave a toy₁ to Billy and a toy₂ to Sally.

(24) The toy_{1,2?}.....that Billy got was red.

The ACT-R architecture, however, does not support multiple simultaneous

retrievals. In the present example, ignoring the effects of noise, the referent with the highest activation after spread of activation from "toy" in (24) will be Sally's toy (toy₂). Although both toy referents (toy₁ and toy₂, and also the generic/prototype toy referent) receive an equal activation boost from the noun "toy", the toy₂ referent was more recently used which gives it an activation advantage (Equation E1). So in the present model, the memory-based arbitration will result in the selection of only one referent, which precludes the presence of an extra arbitration phase during preliminary processing. Text stimuli with such closely matched competitors may be relatively rare in practice, and it remains an empirical question whether an extra arbitration phase might sometimes play a role in preliminary assignment. The answer to this question will determine the extent to which this potential limitation of the present model (and of ACT-R more generally) is germane. Because MEMBRAL is implemented in the python instantiation of ACT-R (Pyke, West & LeFevre, 2007a, 2007b; Stewart & West, in press), my colleagues and I are in a position to alleviate this limitation at the architectural level if need be. An additional limitation of the present model is that it does not recognize that relative clauses, like "that Billy got" in (24) are part of the noun phrase. MEMBRAL presently lacks a syntactic parser, so preliminary referent assignment is invoked on encountering the noun "toy", and will not be re-invoked upon reading the entire noun phrase "The toy that Billy got". With the addition of a syntactic parser, this limitation could be easily remedied. The present model makes the plausible prediction that preliminary assignment will automatically be invoked upon reaching the noun (which often signals the end of the noun phrase). The same preliminary process

could simply be re-invoked at phrase end if a relative clause is present. The procedural operation would be the same, except that referents in memory would receive activation spread from all the words in the phrase prior to retrieval.

In summary, MEMBRAL has some limitations in scope (e.g., syntactic parsing), and some limitations inherent in the LSA estimates of association strengths and in the ACT-R architecture. Many of the limitations, particularly in scope, can be readily remedied in future iterations of the model, and despite such limitations, the model has provided good predictions of the frequency and types of preliminary assignment errors, and the time course of preliminary assignment.

Summary

In this chapter I have described the operation and performance of my MEMBRAL model of the preliminary referent retrieval process. The MEMBRAL model provides the first (to the author's knowledge) detailed, real-time operationalization of the memory-based view of referent retrieval for nouns. By simulating the processing of the context word-by-word in real-time, the model estimates the (pre and post-noun) accessibilities of referents in memory. It thus predicts when a particular anaphor will be initially misinterpreted as a new referent. The proposed model confirms that an automatic memory-based account is sufficient to account for a high rate of success at preliminary referent retrieval for anaphoric nouns. This model could be adapted to have a broader practical application – to assess how comprehensible (each referring

expression in) a given text is. Such a tool would also be useful in psycholinguistics research to assess the degree to which referent accessibility levels are controlled/comparable across stimuli.

CHAPTER 8: Summary and Conclusions

In the present research I investigated the process of preliminary referent retrieval for nouns. The research program consisted of four empirical studies (two eye-tracked reading studies and two cloze studies) and culminated in a working computational model of the process (MEMBRAL), which simulated the processing of texts in real-time, and provided estimates of the pre- and post-noun activation levels of referents in memory. In this final chapter, I provide a brief overview of the theoretical and empirical contributions contained in this research, and I outline some directions for future work.

Synopsis of Contributions

This discussion of the contributions of this research project have been partitioned into the following categories: theoretical, empirical, and applied.

Theoretical Contributions

In this research I identified and addressed a previously neglected and underestimated problem in noun processing: The Anaphoric Status Classification Problem. In particular: how are non-anaphoric definite nouns (vs. anaphoric definite nouns) processed; how, mechanistically, do readers explicitly or implicitly differentiate between them; and how accurate are such preliminary classifications. These issues had not received explicit attention in the traditional framework for several reasons (empirical, theoretical and methodological). First, the empirical evidence that definite noun phrases are equally likely to be introductory as to be anaphoric is relatively recent (Poesio & Veria, 1999; Gundel et al., 2001). Second, the terminology and research

framework adopted from research on pronouns obscured the classification issue for referent assignment to nouns. Finally, many prior psycholinguistic accounts of noun processing took the form of verbal descriptions rather than computational models, which allowed some processing aspects to remain largely unnoticed or underspecified.

By exposing some problematic presuppositions inherent in the traditional terminology and assumptions, the present research provided a framework that highlighted the classification problem, and, in particular, a key type of preliminary interpretation error in which readers interpret an anaphorically-intended noun as a new referent (a misclassify-as-new error). This new framework then paved the way for the development of a unified memory-based model (MEMBRAL) in which classification is implicit, and is determined by which type of referent (particular or prototype) becomes most active after spreading of activation from the critical noun.

The MEMBRAL computational model itself constitutes a theoretical contribution. It serves as an existence proof that general memory mechanisms (un-compartmentalized spreading activation/resonance) are sufficient to account for a high rate of accuracy in preliminary referent assignment for anaphorically-intended nouns. As such, parsimony argues against the existence of special-purpose process to accomplish this function. Further, MEMBRAL demonstrates that such general memory mechanisms suffice to retrieve referents for either anaphoric or non-anaphoric nouns, and in so doing, to discriminate between these two possible uses (though not always with perfect accuracy). Thus, a unified process suffices to process both anaphoric and non-anaphoric nouns, and no explicit classification process or a priori information about a

noun's anaphoric status is required. This is an important attribute of the model because definite nouns can be either anaphoric or introductory. Thus, MEMBRAL expands on prior memory-based (resonance) accounts in the noun anaphor literature in several significant ways.

i) It emphasizes that memory is populated with prototype as well as particular referents, and explicitly addresses the classification problem by providing a unified process model applicable to non-anaphoric as well as anaphoric nouns.

ii) It provides a quantitative, real-time operationalization of the influence of prior context on referent activations (both particular referents and prototypes), and thereby provides a procedural basis for the influence of some known effects on noun anaphor processing (e.g., distance).

iii) It provides a quantitative operationalization of the effects of the general memory mechanisms (i.e. spreading activation/resonance) that are automatically triggered when the noun is encountered; thus it predicts the relative activation levels of referents in memory, both within and outside the current discourse.

iv) MEMBRAL's level of operational detail enables specific predictions about which referent will be retrieved for a given noun in a given context; and consequently can predict the likelihood of all three types of errors (misclassify-as-new, misclassify-as-anaphoric, and discourse-distractor errors). It thereby provides a way to quantitatively assess the effectiveness of a particular choice of anaphor word in a given context.

v) It suggests a possible functional rationale for regressing to a critical anaphor

(overt re-analysis), in particular, such regression may serve to re-engage the same automatic memory mechanisms that are invoked during first-pass processing, under circumstances where the intended referent may have been primed by post-anaphor words.

In summary, the MEMBRAL model has extended the traditional memory-based treatment in terms of both breadth and detail.

Empirical Contributions

At the empirical level, the present research filled some notable evidentiary gaps in the literature. I applied three paradigms (eye-tracked reading, anaphor-status-marking, and cloze tests), which had never previously been applied for comparing the processing of non-anaphoric nouns versus literal noun anaphors versus metaphoric noun anaphors. This new eye-tracked reading evidence revealed that:

- i) readers did not spend longer during first-pass reading of non-anaphoric definite nouns versus anaphoric definite nouns, which is inconsistent with the serial step for instantiating new referents under a special-purpose account; and
- ii) in contrast to prior evidence from the word-by-word self-paced paradigm (e.g, Budiu & Anderson, 2002; Janus & Bever, 1985), the present research indicated that the extra time spent on sentences with metaphoric anaphors (vs. literal anaphors) was due to reanalysis processes rather than due to extended time on the noun.

I also augmented the eye-tracked reading investigation by developing and applying an anaphor-status-marking paradigm in which the anaphoric status of the noun was indicated by font colour, to determine if readers could leverage anaphor status

information on encounter, and thus increase the accuracy and duration of preliminary assignment of referents to nouns (Experiment 4). Additionally, I adapted a cloze paradigm (Experiments 1 and 3) to gather evidence about a reader's interpretation of the noun prior to exposure to the remainder of the sentence. In contrast to time course data, which do not provide direct evidence about the nature of the preliminary noun interpretation, the cloze data confirmed unequivocally that readers do sometimes make new referent interpretations when they encounter an anaphorically-intended noun, even when it is literal. Further, the cloze test in Experiment 3 constituted the first (to the author's knowledge) empirical exploration of Gundel et al.'s (2001) hypothesis that the demonstrative determiner "that" should provide an immediate indication of anaphoricity, whereas "the" does not. While not conclusive, the data indicate that the demonstrative determiner "that" did not substantially reduce the likelihood of misclassification-as-new errors, which suggests that these two determiners may not have distinct impacts on preliminary referent retrieval. Below, I propose future work to contrast the influence of the definite article with the indefinite article.

Finally, the simulation results and their correlations with the human performance data also constitute an empirical contribution. In all, the empirical evidence gathered in this research lends support to the memory-based account over the special-purpose account of preliminary referent retrieval.

Applied Contributions

MEMBRAL may also have a practical application as a tool for assessing the comprehensibility of a text. It can be used to predict both the likelihood and nature of

misinterpretation for each noun in a given context.

Future Work

All three facets of the research (theoretical, empirical and applied) offer interesting directions for future work. For example, there are several ways in which the capabilities of the computational model could be expanded. Because the model is a production system consisting of if-then rules that are implemented according to the ACT-R architecture with its cognitively-motivated constraints and principles, it should be relatively straightforward to incorporate the capabilities from other ACT-R models such as pronoun processing (e.g., Emond, 1997) and syntactic processing (e.g., Budiu & Anderson, 2004).

Additional empirical work would also be useful to further explore the process of noun phrase referent assignment and to further test and expand the model. For example, the present experiments focused on definite noun phrases (those commencing with "the"), and it would be useful to gather additional evidence about the possible impact of the indefinite article ("a" vs. "the") on the accuracy and time course of preliminary referent retrieval. In the current version of MEMBRAL, the pre-nominal article does not have an impact on preliminary referent retrieval. The noun rather than the article affects the activation levels of referents in memory (via spreading activation/resonance), and thereby determines which referent gets selected -- and that referent may be either generic or specific, whichever referent is most active. However, in contrast to "the" which does not convey reliable information about the anaphoric status of a noun (Poesio & Veria, 1998), the indefinite article "a" seems reliably correlated

with a non-anaphoric status. How will the indefinite article (i.e., this non-anaphoric status information) impact preliminary or subsequent processing? Dell et al., (1983) reported that the indefinite article does not immediately inhibit (i.e., lower the activation level of) discourse referents. However, it is possible, in principle, that a different special-purpose referent 'search' process may be immediately invoked for nouns preceded by the indefinite article (vs. definite article), which is restricted to generic referents, and precludes retrieving a specific referent from the discourse. Consider the following example:

(25) A robin stood on a big rock.

(26a) The bird/girl looked at the robin.

(26b) A bird/girl looked at the robin.

In (26a), the preliminary interpretation of "The bird" will usually prompt the retrieval of the robin referent mentioned in (25), which should lead to a coherence problem (and consequently a delay and/or regression) when the reader reaches "robin" at sentence end. In contrast, if there is a special-purpose process for the indefinite article (as in Figure 1.1 in Chapter 1), it should preclude a preliminary interpretation that "A bird" in (26b) refers to the robin in (25), and there should be no delay at sentence end on "robin" (whether the sentence begins with "A bird" vs. "A girl").

However, the memory mechanisms (spreading activation from the noun) responsible for referent retrieval in MEMBRAL are assumed to be automatic (non-optional), and they suffice to reliably retrieve the correct prototype referent for non-anaphoric nouns preceded by "the", so these points argue against the need for a separate

special-purpose process for non-anaphoric nouns preceded by "a".

Concluding Remarks

Our ability to associate referents to referring expressions is fundamental to linguistic communication. When readers encounter a noun, the evidence in the present research suggests that preliminary referent retrieval is rapid though sometimes erroneous. In particular, the present research (see also Levine et al., 2000) indicates that readers may misclassify an anaphoric noun as a new referent during preliminary processing (misclassify-as-new error), because the anaphoric status of a noun is not explicitly marked, and because readers do not seem to have an anaphoric bias in preliminary interpretation. Consequently, the findings in the present research support a memory-based, rather than special-purpose account of preliminary referent retrieval for definite nouns. In particular, the MEMBRAL model developed in the present research illustrates that a unified memory-based system can process definite nouns intended as either anaphoric or introductory, and can provide a good fit with human performance in preliminary referent assignment. Although preliminary assignment may not always be accurate, readers may often subsequently detect and correct misclassify-as-new errors based on content in the remainder of the sentence.

The present focus has been on the role of the rapid automatic retrieval process of preliminary reference assignment, rather than on the nature of the subsequent reanalysis processes. However, I note that reanalysis of the referent assignment may often prompt a regression to the noun, and that reanalysis is apparently hampered when such regressions are not possible, as in the word-by-word self-paced paradigm (Budiu &

Anderson, 2002). Such regressions to the noun may facilitate the re-use of the automatic resonance mechanism, under 'hindsight' conditions where the relative activations of the candidates in memory have been primed by the concepts present in the remainder of the sentence.

Whenever the final referent assignment for an anaphor is achieved by means of reanalysis, however, there may be lasting impacts that persist beyond on-line processing delays. For example, when readers perform syntactic reanalysis of garden path sentences, the final representation of the discourse content may be inaccurate due to lingering remnants of the preliminary misinterpretation (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, Christianson, & Hollingworth, 2001). Degraded integrity of the final discourse representation may also occur for cases of referential reanalysis, and may have contributed to the poorer question answering performance for the metaphoric trials in the present research, even in cases when reanalysis was performed. If some anaphors are initially instantiated as new referents in the discourse model, despite the correction of this misinformation, vestiges of these referents could persist in the model and compete as antecedents for subsequent anaphoric expressions (Johnson & Seifert, 1998). Further research is necessary to explore such potential long-term 'costs' of using anaphors that are difficult to resolve during initial analysis.

REFERENCES

- Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106, 748-765.
- Altmann, G. T. M. (1994). Regression-contingent analyses of eye-movements during sentence processing: Reply to Rayner and Sereno. *Memory & Cognition*, 22, 286-290.
- Altmann, G. T. M., Garnham, A., & Dennis, Y. (1992). Avoiding the garden path - Eye-movements in context. *Journal of Memory and Language*, 31, 685-712.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. L. (2004). An integrated theory of the mind. *Psychological Review*, 111, 1036-1060.
- Balogh, J., Zurif, E., Prather, P., Swinney, D., & Finkel, L. (1998). Gap-filling and end-of-sentence effects in real-time language processing: Implications for modeling sentence comprehension in aphasia. *Brain and Language*, 61, 169-182.
- Baum, S. R. & Titone, D. (2005). Acoustic correlates to idiomatic interpretations in brain-damaged populations. *Brain and Language*, 95, 223-224.
- Bean, D. L. & Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, (pp. 373-380).
- Brown, G. (1983). Prosodic structure and the given/new distinction. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 67-77). Berlin: Springer Verlag.
- Budiu, R. & Anderson, J. R. (2002). Comprehending anaphoric metaphors. *Memory &*

Cognition, 30, 158-165.

- Budiu, R. & Anderson, J. R. (2004). Interpretation-based processing: a unified theory of semantic sentence comprehension. *Cognitive Science*, 28, 1-44.
- Budiu, R. & Anderson, J. R. (2006). Word learning in context: Metaphors and neologisms. In Michel Vishnevsky (Ed.), *Focus on Cognitive Psychology Research*, Nova Science Publishers, NY: Hauppauge.
- Cacciari, C. & Glucksberg, S. (1994). Understanding figurative language. In M.A.Gernsbacher (Ed.), *Handbook of Psycholinguistics*, (pp. 447-478). San Diego, CA: Academic Press.
- Caramazza, A., Grober, E., & Garvey, C. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour*, 16, 601-609.
- Centre for Lexical Information (2001). *CELEX* [On-line]. Max Plank Institute for Psycholinguistics, Nijmegen. Available: <http://www.ru.nl/celex/>
- Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42, 368-407.
- Clark, H. H. & Haviland, S. E. (1982). Comprehension and the given-new contract. In R. Freedle (Ed.), *Discourse Production and Comprehension* (2nd ed.), (pp. 1-39). Norwood, NJ: Ablex Publishing Co.
- Clark, H. H. & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, 7, 35-41.

- Cooper, R. M. (1974). "The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing." *Cognitive Psychology* 6(1), 84-107.
- Corbett, A. T. (1984). Prenominal adjectives and the disambiguation of anaphoric nouns. *Journal of Verbal Learning and Verbal Behavior*, 23, 683-695.
- Cristea, D. & Dima, G. E., (2001). An Integrating Framework for Anaphora Resolution. *Information Science and Technology*, 4, 271-291. Bucharest: Romanian Academy Publishing House.
- Dell, G. S., McKoon, G., & Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. *Journal of Verbal Learning and Verbal Behavior*, 22, 121-132.
- Duffy, S. A. & Rayner, K. (1990). Eye-movements and anaphor resolution: Effects of antecedent typicality and distance. *Language and Speech*, 33, 103-119.
- Emond, B. (1997). *Modeling natural language comprehension and anaphora resolution with ACT-R*. Paper presented at the Fourth Annual ACT-R Workshop. Pittsburgh, PA.
- Erlich, K. & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behaviour*, 22, 75-87.
- Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30, 3-20.

- Fraurud, K. (1990). Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7, 395-434.
- Frisson, S. & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology-Learning Memory and Cognition*, 25, 1366-1383.
- Frisson, S. & Pickering, M. J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. *Metaphor and Symbol*, 16, 149-171.
- Garnham, A. (1984). Effects of Specificity on the Interpretation of Anaphoric Noun Phrases. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, 36, 1-12.
- Garnham, A. (1989). Integrating information in text comprehension: The interpretation of anaphoric noun phrases. In G. N. Carlson & Tanenhaus M. (Eds.), *Linguistic Structure in Language Processing* (pp. 359-399). Norwell, MA: Kluwer Academic Publishers.
- Garnham, A., Oakhill, J., & Cain, K. (1997). The interpretation of anaphoric noun phrases: Time course, and effects of overspecificity. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, 50, 149-162.
- Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the online resolution of sentences in a discourse. *Journal of Memory and Language*, 33, 39-68.

- Garrod, S. & Sanford, A. (1977). Interpreting anaphoric relations - Integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior*, 16, 77-90.
- Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition*, 32, 99-156.
- Gerrig, R. J. & Healy, A. F. (1983). Dual processes in metaphor understanding: Comprehension and appreciation. *Journal of Experimental Psychology-Learning Memory and Cognition*, 9, 667-675.
- Gerrig, R., & O'Brien, E. (2005). The scope of memory-based text processing. *Discourse Processes*, 39(2&3), 225-242.
- Gibbs, R. W. (1990). Comprehending figurative referential descriptions. *Journal of Experimental Psychology-Learning Memory and Cognition*, 16, 56-66.
- Gildea, P. & Glucksberg, S. (1983). On understanding metaphor - the role of context. *Journal of Verbal Learning and Verbal Behavior*, 22, 577-590.
- Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis (Psycholinguistics). *Cognitive Linguistics*, 8, 183-206.
- Giora, R. (2002). Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, 34, 487-506.
- Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, 7, 92-96.
- Grice, H. (1975). Logic and conversation. In *Speech Acts* (pp. 41-57). New York, NY: Seminar Press.

- Gundel, J. (2003). Information Structure and Referential Givenness/Newness: How much belongs in the grammar? In S.Muller (Ed.), *Proceedings of the HPSG03 Conference*. Michigan State University, East Lansing: CSLI Publications.
- Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69, 274-307.
- Gundel, J., Hedberg, N., & Zacharski, R. (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *Journal of English Language and Linguistics*, 5, 273-295.
- Grosz, Joshi & Weinstein (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 2, 203-225.
- Heim, I. R. (1982). *The semantics of definite and indefinite noun phrases*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, 93, 411-428.
- Janus, R. A. & Bever, T. G. (1985). Processing of metaphoric language: An investigation of the 3-stage model of metaphor comprehension. *Journal of Psycholinguistic Research*, 14, 473-487.
- Johnson, H. M. & Seifert, C. M. (1998). Updating accounts following a correction of misinformation. *Journal of Experimental Psychology-Learning Memory and Cognition*, 24, 1483-1494.
- Just, M. A. & Carpenter, P. A. (1987). *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon, Inc.

- Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.
- Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7, 257-266.
- Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Latent Semantic Analysis at University of Colorado at Boulder. (2003). <http://lsa.colorado.edu/> [On-line]. Available: <http://lsa.colorado.edu/>
- Lemaire, B. & Bianco, M. (2003). Context effects on metaphor comprehension: Experiment and simulation. In F. Detje, D. Dorner, & H. Schaub (Eds.), *Proceedings of the 5th International Conference on Cognitive Modelling (ICCM)* (pp. 153-158). Bamberg, Germany.
- Levine, W. H., Guzman, A. E., & Klin, C. M. (2000). When anaphor resolution fails. *Journal of Memory and Language*, 43, 594-617.

- McElree, B. & Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. *Psychonomic Bulletin & Review*, 6, 486-494.
- McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99, 440-466.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, 30, 551-561.
- Myers, J. L. & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26, 131-157.
- Noveck, I., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol*, 16, 109-121.
- O'Brien, E. J. (1987). Antecedent search processes and the structure of text. *Journal of Experimental Psychology-Learning Memory and Cognition*, 13, 278-290.
- O'Brien, E. J. & Albrecht, J. E. (1991). The Role of Context in Accessing Antecedents in Text. *Journal of Experimental Psychology-Learning Memory and Cognition*, 17, 94-102.
- O'Brien, E. J., Duffy, S. A., & Myers, J. L. (1986). Anaphoric inference during reading. *Journal of Experimental Psychology-Learning Memory and Cognition*, 12, 346-352.

- O'Brien, E. J., Plewes, P. S., & Albrecht, J. E. (1990). Antecedent Retrieval-Processes. *Journal of Experimental Psychology-Learning Memory and Cognition*, 16, 241-249.
- O'Brien, E. J., Raney, G. E., Albrecht, J. E., & Rayner, K. (1997). Processes involved in the resolution of explicit anaphors. *Discourse Processes*, 23, 1-24.
- Onishi, K. H. & Murphy, G. L. (1993). Metaphoric reference: When metaphors are not understood as easily as literal expressions. *Memory & Cognition*, 21, 763-772.
- Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting Metaphors and Idioms - Some Effects of Context on Comprehension. *Journal of Verbal Learning and Verbal Behavior*, 17, 465-477.
- Poesio, M. & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24, 183-216.
- Prince, E. (1992). The ZPG letter: Subjects, definiteness, and information status. In William C. Mann & Sandra A. Thompson (Eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text* (pp. 295-325). Philadelphia: John Benjamins.
- Pyke, A., West, R. L. & LeFevre, J. (2007a). On-line reference assignment for Anaphoric and non-anaphoric nouns: A unified, memory-based model in ACT-R. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*.
- Pyke, A., West, R. L. & LeFevre, J. (2007b). How Readers Retrieve Referents for Nouns in Real Time: A Memory-based Model of Context Effects on Referent

- Accessibility. In *Proceedings of the 8th International Conference on Cognitive Modelling (ICCM)*. Taylor & Francis/Psychology Press.
- Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review*, 85, 59-108.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.
- Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology-Human Perception and Performance*, 30, 720-732.
- Rayner, K. & Sereno, S. C. (1994). Regressive eye-movements and sentence parsing - on the use of regression-contingent analyses. *Memory & Cognition*, 22, 281-285.
- Rosch, E. (1978): Principles of categorization. In Rosch, E. & Lloyd, B. B. (Eds), *Cognition and Categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.
- Sanford, A. J. & Garrod, S. C. (1989). What, when, and how?: Questions of immediacy in anaphoric reference resolution. *Language and Cognitive Processes*, 4, 235-262.
- Searle, J. (1979). Metaphor. In A.Ortony (Ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109-147.
- Sperber, D. & Wilson, D. (1995). *Relevance: Communication and Cognition*. (2nd ed.) Oxford: Blackwell.

- Stewart, M. T. & Heredia, R. R. (2002). Comprehending spoken metaphoric reference: A real-time analysis. *Experimental Psychology*, 49, 34-44.
- Stewart, T. C., West, R.L. (in press). Deconstructing and Reconstructing ACT-R: Exploring the Architectural Space. *Cognitive Systems Research*.
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48, 542-562.
- van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. San Diego, CA: Academic Press.
- van Gompel, R. P. G., Liversedge, S., & Pearson, J. (2005). Antecedent typicality effects in the processing of noun phrase anaphors. In M.Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 119-135). New York, NY: Psychology Press.
- Vieira, R. & Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, 26, 539-593.
- West, R. L., Stewart, T. C., Pyke, A., Emond, B. (2006, July). *Modeling Emotion in ACT-R*. Paper presented at the 13th Annual ACT-R Workshop.
- Wiley, J., Mason, R. A., & Myers, J. L. (2001). Accessibility of potential referents following categorical anaphors. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 1238-1249.

APPENDIX A: Stimulus Materials

Notes

The following stimuli were adapted for each of the four experiments.

For each story, there are 3 context sentences, a fourth critical sentence, and one or two possible TRUE/FALSE questions.

In Experiments 1, 2, the critical sentence began with "The", however, in Experiment 3, half the anaphoric trials had "The" as the first word of the fourth sentence, and half of the anaphoric trials were presented with "That" as the first word of the critical sentence.

The first noun in the fourth sentence is the metaphoric version of the anaphor, and the second noun is a literal anaphor for the same antecedent, and the third noun is non-anaphoric.

In the cloze experiments (1 and 3), the fourth sentence was only shown up until and including the noun anaphor, and the questions were not presented.

Stimuli for Experiments 1-3

1. There was a gorgeous redhead at the bar.
 George went to go speak to her.
 He tripped on the dance floor on his way over.
 The girl/fox/D.J. didn't notice.
 question: The redhead saw George trip.
2. Lisa went in for an operation on her knee.
 An incompetent doctor performed the procedure.
 Afterwards, the knee was even worse than before.
 The butcher/surgeon/nurse didn't care.
 question: The physician didn't show remorse.
 question: The operation was on Lisa's knee.

3. Miss Bell really wanted her class to like her.
So she didn't always enforce the school's rules.
One day, a student lit a cigarette in her class.
The teacher/creampuff/parents did nothing.
question: Miss Bell reprimanded the student.
4. John expected an easy divorce from his naïve wife.
But her attorney was formidable and aggressive.
John felt the division of assets was unfair.
The shark/lawyer/judge didn't care.
question: The attorney wasn't concerned about John.
5. Susan smiled at her talkative two year old, Mary.
One had to be careful not to swear around Mary.
Mary would remember and repeat any word she heard.
The parrot/toddler/guar said a bad word at the pool.
question: There was a pirate in the story.
question: Mary was four.
6. Sally asked her father if she could play outside.
He said she had to ask her mom.
Sally had chores to do and it looked like rain.
The boss/mother/forecast said that it would be OK, though.
question: Her mother gave permission to play.
7. Bev paid for Derek's lunch in the cafeteria.
He often conveniently forgot his wallet.
Bev was annoyed at having to buy his lunch again.
The parasite/borrower/cashier didn't notice.
question: Derek knew Bev was annoyed.
8. Jen cast a hostile glance towards the principal.
He was confused by the angry look.
Then he realized her ex-boyfriend was behind him.
The dagger/glare/comment was meant for her ex-boyfriend.
question: Jen was angry at the principal.
9. Josh was a bachelor with a very messy home.
He never used his vacuum or his mop.
His mother refused to visit him.
The place/dump/yard was a disgrace.
10. Cathy walked along the old train tracks.
She jumped from plank to plank.

She looked towards the horizon.
The ribbons/rails/day seems to stretch on forever.

11. In the country, the night sky looked different.
It was filled with tiny, twinkling points of light.
Jill noticed one that was especially bright.
The diamond/star/moon was beautiful.
question: The star was beautiful.
12. After graduation, Gail wanted a research job.
Her professor wrote her a glowing recommendation.
She picked up the envelope from his office.
The ticket/letter/degree would get her the job.
question: The recommendation wasn't necessary.
13. The spinster had absolutely no regard for sentiment.
An admiral had been trying to seduce her in vain.
He sent her a lonely love poem every day.
The icicle/woman/postman threw them out.
question: The spinster cried over the poems
14. Stan was hired by a family friend who was a builder.
Stan would earn two thousand dollars for the job.
Stan was happy about the opportunity.
The bread/money/tools would come in handy.
15. Marie broke her resolution and lit a cigarette.
She knew it was unhealthy, but smoked a pack a day.
Her family was upset that Marie often smoked inside.
The chimney/smoker/complaints didn't stop.
question: Marie quit smoking.
16. Leroy was the best student in his grade one class.
He never got tired of learning about nature.
He had a remarkable memory for the facts.
The sponge/boy/teacher summarized the lesson.
17. Bob was a lazy man addicted to TV.
Every day he'd lie on the sofa all day long.
His friends were surprised he didn't get bored.
The potato/fellow/channel never changed.
question: Bob adjusted his habits.
question: Bob sat in an arm chair all day.

18. Ann carelessly dropped Lou's laptop computer.
Lou lost his temper and expressed his anger.
But it might have been worse.
The tantrum/storm/warrantee soon ended.
question: Lou's anger didn't last long.
19. John was timid and avoided all risk.
His sister signed him up for scuba diving.
But it never happened.
The coward/chicken/instructor didn't show up.
question: John went scuba diving.
20. Jill's teenage brother Mark had terrible manners.
At the table, he belched and ate with his hands.
Jill chastised Mark in vain.
The pig/slob/host was amused.
question: The person who belched was embarrassed.
21. It was Friday, and Melissa had a test.
It covered the whole first half of the term.
She spent two nerve wracking hours writing it.
The midterm/nightmare/week was finally finished.
question: Melissa wrote an exam.
22. The daycare teacher was playing the piano.
The class was singing a song.
Cookies were served afterwards.
The troops/children/visitors were happy.
question: The children had enjoyed the event.
23. It was the first pregnant woman Luke had seen.
His dad told him she had a baby in her tummy.
Luke stared at it as the woman came toward him.
The barrel/belly/dress was huge.
question: Luke didn't think the woman's tummy was very big.
24. Old Mr. McCann got a gold watch when he retired.
Lyn would miss him but hoped to get his position.
He had been with the car company for fifty years.
The fossil/man/wife was eager for his retirement.
question: Lyn was looking forward to retirement.
25. Frank adored his month-old Siamese.
He pampered her, and bought her a silk pillow.

Frank almost also got a dog in his apartment.
 The kitten/princess/landlord didn't like dogs.
 question: Frank's Siamese was fond of dogs.

26. Sally's boyfriend was an inmate in prison.
 He was dishonest and her mom disliked him.
 Sally continued going to the jail anyway.
 The weasel/convict/guard was handsome.
 question: The inmate was physically attractive to Sally.
27. Brad had lost eighty pounds by his birthday.
 His size XXL top looked huge and hung off him.
 On his scale, he vowed not to gain any back.
 The tent/shirt/cake would go into the trash.
 question: Brad was going to get rid of his old shirt.
 question: Brad had lost eighty pounds.
28. Greg had a sudden insight about a project at work.
 He shared the insight with his supervisor.
 She was very impressed and enthusiastic.
 The idea/gem/future was exciting.
 question: The supervisor had heard the idea before.
29. Kevin didn't enjoy visits from his wife's mother.
 She constantly watched and interrogated him.
 He went to the local pub to avoid her.
 The bloodhound/mother-in-law/neighbour found him at the park.
30. Earl left home without much common sense.
 He put an unopened can of soup in his microwave.
 His mom explained why this was a bad idea.
 The turkey/fool/family didn't listen.
 question: Earl put metal in the microwave.
31. The bank employee responded slowly to the request.
 Mark almost lost his composure. He waited a long time for the money.
 The snail/teller/heist was finally finished.
 question: The teller fulfilled Mark's request.
32. Julie looked at a red sweater in a trendy store.
 The male clerk started pressuring her to buy it.
 She didn't like the atmosphere.
 The bulldozer/salesperson/music was annoying.
 question: The clerk annoyed Julie.

33. The small plane flew over the forest.
The pilot followed the eastward trail.
It wound back and forth through the trees.
The snake/path/sun led towards the valley.
question: The valley was east.
34. Jack bought a used car.
It was inexpensive but also nine years old.
He planned to drive to Hollywood to find work.
The vehicle/lemon/radio didn't survive the trip.
question: The car wasn't working properly.
35. Han ate a burger as he worked overtime at his desk.
The incentive was a free trip if the project was done.
Though he was tired, it would be worth
The carrot/reward/team was motivating.
question: Han was motivated by food.
36. It was Fred's first day at Carleton.
He used the underground passages.
He took a wrong turn and was late for biology.
The maze/tunnels/lecture was(were) crowded.
question: There were several other people in the passages.

Stimuli for Experiment 4

Slight adjustments were made to the above stories for use in Experiment 4, because sometimes the intended antecedent was a proper name rather than a noun. In Experiment 4, a noun antecedent was required so that the same noun could be used in fourth sentence as an anaphor (antecedent-match condition). In the following stimuli, the first noun in the fourth, critical sentence is the metaphoric anaphor, the second noun is the literal category anaphor, and the third noun is the antecedent-match anaphor.

1. Bob saw a sexy redhead at the bar.
He cut across the dance floor to try to meet her.

To his embarrassment, on the way to her he fell.
 The fox/girl/redhead didn't notice.
 question: The girl saw Bob trip.

2. Lisa went in for an operation on her knee.
 An incompetent doctor performed the procedure.
 Lisa told him that the knee was worse than ever.
 The butcher/doctor/surgeon didn't care.
 question: The physician showed no remorse.
3. The live-in babysitter wanted the kids to like her.
 So she never disciplined them or enforced the rules.
 The kids were rude and sometimes even yelled at her.
 The creampuff/nanny/babysitter didn't scold them.
 question: The babysitter scolded them.
4. Amy considered suing her neighbour but felt guilty.
 She told an attorney about the fence paint dispute.
 He listened to her with an attitude of cold analysis.
 The shark/lawyer/attorney told her to sue.
 question: The attorney suggested that Amy drop the case.
5. Sharon smiled at her talkative toddler, Mary.
 One had to be careful not to swear around Mary.
 Mary would remember and repeat every word she heard.
 The parrot/child/librarian/toddler learned a bad word from the radio.
 question: A bird was listening to the radio.
6. Mark passed a sunbather asleep on the beach.
 She was getting a bad sunburn, and was red all over.
 Mark decided to wake her up before it got worse.
 The lobster/woman/sunbather was grateful to him.
 question: The woman was glad he woke her up.
7. Cindy lent her boss money in the cafeteria.
 He almost always forgot to pay her back.
 Cindy frowned while giving the money to him.
 The parasite/borrower/boss bought a cola with the money.
 question: Cindy's boss bought a cola.
8. A driver gave Jen an angry stare as he drove past.
 But his cold look hadn't been intended for her.
 The taxi on her right had cut him off.
 The dagger/glare/look had been directed towards the taxi.

question: The driver was angry at Jen.

9. Josh was a bachelor with a very messy house.
His mother refused to visit him.
He never tidied or used his vacuum or his mop.
The junkyard/place/house was a disgrace.
question: Josh's house was near a junkyard.
question: Josh wasn't married.
10. Cathy walked along the old train tracks.
She jumped from plank to plank.
She looked towards the horizon.
The ribbons/tracks/rails seemed to run on forever.
question: Cathy could see the end of the tracks.
question: Cathy walked along a dirt road.
11. In the country, the night sky captivated Jill.
It was filled with tiny, twinkling points of light.
She saw a little one that was especially bright.
The diamond/star/light twinkled high in the north.
12. Gail wanted a research job at Oxford.
Her professor wrote her a glowing recommendation.
She picked up the recommendation letter from his office.
The ticket/reference/letter would get her the job.
question: The recommendation wasn't important.
question: Gail wanted a research job at Harvard.
13. The spinster had absolutely no regard for sentiment.
An Admiral had been trying to seduce her in vain.
He sent a lonely love poem to her every day.
The icicle/lady/spinster just threw the poems away.
question: The spinster cried over the daily poems.
14. Stan was hired by a family friend who was a builder.
Stan would be paid two thousand dollars for the job.
Stan knew what he wanted to spend the money on.
The bread/cash/money would help pay for a motorcycle.
question: Stan was going to work for a baker.
15. Marie was a smoker.
She knew it was unhealthy, but she was hooked.
Her friends and family begged Marie to at least cut back.
The chimney/addict/smoker went through two packs a day.

question: Marie smoked less than three packs a day.

16. Leroy was the best student in his grade one class.
He never got tired of learning about history.
Remarkably, he remembered each and every fact.
The sponge/boy/student scored perfect on every quiz.
question: Leroy did well on the quizzes.
question: Leory was the best student in his class.
17. George was a lazy man addicted to TV.
Every day he'd lie on the sofa all day long.
His friends were surprised he didn't get bored.
The potato/fellow/man hardly ever went outside.
question: George often went out.
question: George sat in an arm chair.
18. Ann carelessly dropped Lou's laptop computer.
Lou lost his temper and threw a fit.
Ann had never seen him with so filled with anger.
The storm/tantrum/fit was over within a few minutes.
question: Lou's rage didn't last for long.
19. Fay's brother was a girl who loved adventure and excitement.
Her brother was timid and avoided all risk.
Their parents invited him to come camping.
The chicken/coward/brother refused to go.
question: John tried scuba diving.
question: His sister signed him up for sky diving
20. Joan's son Mark had terrible manners.
At the table he belched and ate with his hands.
Joan scolded Mark in vain.
The pig/slob/son just continued being rude.
question: Mark stopped being rude.
21. Rick was afraid to meet his girlfriend's parents.
He arrived at noon for lunch at their house.
During the meal, her dad criticized all his goals.
The nightmare/event/lunch was finally over at two o'clock.
question: Rick fell asleep.
22. The daycare teacher was playing the piano.
The class was singing a song.
Cookies were served afterwards.

The troops/children/class enjoyed their snack.

question: The class enjoyed the cookies.

question: Cake was served after the song.

23. It was the first pregnant woman Luke had seen.
His dad told him she had a baby in her tummy.
Luke stared at it as the woman came toward him.
The barrel/belly/tummy looked quite huge to him.
question: Luke didn't find the woman's tummy big.
question: Luke was with his Mom.
24. The company's elderly accountant was retiring.
Nobody was sure of his age, but he looked about 80.
He'd handled the company's finances for 50 years.
The dinosaur/gentleman/accountant was eager for his retirement.
question: The accountant didn't want to retire.
25. Frank adored his tiny month old Siamese cat.
He pampered her, and bought her a velvet pillow.
He almost got a dog to play with her, but didn't.
The princess/kitten/cat didn't like dogs.
question: Frank's Siamese was fond of dogs.
question: Frank bought his cat a pillow.
26. Sally's boyfriend was an inmate in prison.
He was dishonest and her mom disliked him.
Sally continued going to the jail anyway.
The weasel/convict/boyfriend was handsome.
question: The inmate was physically attractive.
question: Sally's mother disliked Sally's boyfriend.
27. Brad's favourite top was too big for him now.
He had lost eighty pounds.
The top was XXL, and looked huge and hung off him.
The tent/shirt/top was going into the trash.
question: Brad was going to get rid of his old shirt.
question: Brad had lost eighty pounds.
28. Len had an insight for improving profits at work.
He shared the valuable insight with his supervisor.
She was very impressed and enthusiastic about it.
The gem/idea/insight would cut costs in half.
question: The idea was not met with enthusiasm.
question: The supervisor was male.

29. Kevin saw his nosy neighbour approaching.
If she saw him outside, she'd pry for gossip.
He tried hiding in the garage to avoid her.
The bloodhound/busybody/neighbour found him and said "hello".
question: Kevin successfully avoided his neighbour.
30. Tim's roommate careless and had no common sense.
He put an unopened can of beans in the microwave.
Tim tried to warn him not to press start.
The turkey/fool/roommate didn't listen.
question: Tim's roommate heeded his advice.
31. The bank teller responded slowly to the request.
Shaun almost lost his composure.
He waited a long time for her to do the paperwork.
The snail/clerk/teller finally finished counting.
question: The teller eventually finished the task.
32. Julie looked at a red sweater in a trendy store.
The male clerk started pressuring her to buy it.
She left because she didn't like the atmosphere.
The bulldozer/salesperson/clerk wouldn't leave her alone.
question: The clerk's pressure did not appeal to Julie.
question: The sweater was red.
33. The small plane flew over the forest.
The pilot followed the stream eastward.
It wound back and forth through the trees.
The snake/river/stream led towards the valley.
question: The valley was east.
34. Jack bought a used car.
It was inexpensive but also nine years old.
He planned to drive it to Hollywood.
The lemon/vehicle/car broke down on the way.
question: The car ran smoothly on the trip.
35. Han worked overtime at his desk.
The incentive was a free trip if the project was done.
Though he was tired, the trip kept him motivated.
The carrot/reward/trip was attainable.

question: It was possible to complete the project on time.
 question: Han ate a burger.

36. It was Fred's first day at Carleton.
 He used the underground passages.
 He took a wrong turn on his way to his English lecture.
 The maze/tunnels/passages often confused new students.
 question: Fred was confused by the passages.
37. Young Louise loved visiting her rich relatives.
 They had a large stone house with twenty rooms.
 She always spent the summer holidays there.
 The castle/mansion/house was hard for her to leave.
 question: The house was big.
38. George lay in his hammock with his eyes closed.
 He listened to the wind in the leaves.
 It was a soothing and rhythmic rustling.
 The music/sound/rustling lulled him to sleep.
 question: George was listening to an Ipod.
39. Mrs. Lou's little terrier was white and shaggy.
 It kept her company since her husband passed away.
 Everyday she'd take it for a walk in the park.
 The mop/dog/terrier barked at the birds.
 question: Mrs. Lou's dog didn't bark.
40. The tallest guy in the school was over 6'6" tall.
 On the basketball team, he was the captain.
 The coach was only 5'2".
 The giraffe/player/captain had to bend down to hear the coach.
 question: The tall guy was captain of the volleyball team.
41. Sandra was cleaning out her walk closet.
 She had hundreds of slacks that she no longer wore.
 She tossed them into a huge pile.
 The mountain/heap/pile almost reached the ceiling.
 question: Sandra made a pile of her old shoes.
42. Marnie's roommates threw her a surprise party.
 She didn't suspect anything as she went downstairs.
 Everyone cheered and threw white confetti into the air.
 The snow/paper/confetti fluttered around her.
 question: The party was outside.

APPENDIX B: Descriptive Statistics for Critical Nouns

Table B1 details the lengths (i.e., number of letters), corpus frequencies, and LSA anaphor-referent similarities for the different types of critical nouns. For the non-anaphoric critical nouns, the LSA similarity was calculated between the noun and the context as a whole, to gauge the overall ‘fit’ of the noun with story. For the anaphors, LSA similarities were calculated between the noun and the set of descriptive terms for the *intended* referent in the context. The mean similarity between an anaphor term and the other referents in the discourse was only 0.16, 0.17, 0.4 for the antecedent-match anaphors, literal category anaphors, and metaphoric anaphors, respectively.

Table B1: Mean Statistics for Critical Nouns (std. dev.)

Critical Noun Type	Characters	Log	LSA
		Frequency ϕ	Similarity
Anaphors (Mean)	5.96	1.31	0.34
Antecedent Match	6.12 (1.97)	1.36 (.85)	0.49 (.22)
Category	5.76 (1.78)	1.54 (.82)	0.35 (.15)
Metaphoric	6.05 (1.74)	1.03 (.59)	0.19 (.10)
Non-Anaphors (Mean)	5.83	1.33	
Group 1	6.07 (2.69)	1.20 (.85)	
Group 2	5.57 (2.00)	1.49 (.62)	
Grand Mean	5.90	1.32	0.32

ϕ Note: Frequencies were determined using the CELEX database.

APPENDIX C: Reading Rates for Context Sentences in Experiment 4

Context sentences (i.e., the first 3 sentences in each story) had no colour marking in either block. In this experiment (vs. Experiment 2), timing data were also recorded for the context sentences, in part to check for effects of blocking independent of the marking manipulation. A two-tailed correlational analysis of reading times indicated that context sentences were read more quickly in Block 2 than in Block 1 (2367 ms vs. 2887 ms), $r(5098) = -.232, p = .000$. Furthermore, within each block, context sentences were read more quickly as the trial number increased, $r(5098) = -.143, p = .000$. Within a story, participants tended to read later sentences in the story faster than earlier sentences in the story, $r(5098) = -.219, p = .000$, which is a finding also reported by McKoon and Ratcliff (1980). Finally, sentence reading times were positively correlated with various measures of sentence length: i) the number of non-space characters in the sentence, $r(5098) = .334$; ii) the number of words in the sentence, $r(5098) = .278, p = .000$; and iii) number of words longer than 3 characters in the sentence, $r(5098) = .272, p = .000$. The number of non-space characters proved to be most predictive of sentence reading time, and no additional correlation remained between context sentence reading time and number of words once the number of characters was partialled out, $r(5097) = .011, p = .441$. Interestingly, although participants took generally longer to read sentences with more characters, participants also tended to increase their per character reading pace when reading long sentences, so reading time per character was inversely correlated with sentence length, $r(5098) = -.145, p = .000$.

These patterns and 'practice' effects found in the context sentences were expected

to generalize to the critical sentences. In particular, the reading pace was expected to increase overall in Block 2, and in later trials within each block, in addition to any effect of noun status marking. To control for such effects between and within blocks, sentence length and trial number were included as covariates in the time course analyses on the critical sentences.