Understanding the Role of Referential Processing in Sentence Complexity

by

Tessa Cartwright Warren

B.A., Cognitive Psychology
Yale University, 1996

Submitted to the Department of Brain and Cognitive Sciences in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September 2001

Signature of Author: ________________
Department of Brain and Cognitive Sciences
August 7, 2001

Certified by: ________________
Edward A. F. Gibson
Associate Professor of Cognitive Science
Thesis Supervisor

Accepted by: ________________
Earl Miller
Associate Professor of Neuroscience
Chairman, Department Graduate Committee

Understanding the Role of Referential Processing in Sentence Complexity

by

Tessa Cartwright Warren

Submitted to the Department of Brain and Cognitive Sciences
on August 7, 2001 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in Cognitive Science

Abstract

Language comprehension requires syntactic, semantic and pragmatic processing. The work presented in this thesis clarifies the role that the resource demands of syntactic and referential processing play in sentence complexity. Results are interpreted within the framework of the Dependency Locality Theory (Gibson, 1998), which provides a hypothesis about how computational resources constrain the process of sentence comprehension. These new results support and further develop the DLT's discourse-based distance metric for computing locality.

The experiments presented here were designed to investigate the referential processing load imposed by relating noun phrase (NP) anaphors to their antecedents and to discover the ramifications of increased referential processing load on behavioral measures of language comprehension. Four questionnaire experiments tested the intuitive complexity of doubly nested sentences containing NPs that were differently referentially accessible. These experiments demonstrated that sentences with structural dependencies crossing less accessible referents are judged more difficult than sentences with structural dependencies crossing more accessible referents. They also showed that referential accessibility manipulations had a negligible effect on intuitive complexity in positions that did not interrupt long distance structural dependencies. Five self-paced word-by-word reading experiments elucidated the time course of the complexity ramifications of increased referential processing. Each of these experiments showed that when less accessible referents interrupted long distance structural dependencies, reading times slowed more at the completion of the structural dependency than at the referent itself.

From the results of these experiments it is argued that performing referential processing during an incomplete structural dependency makes accessing the representation of the beginning of the dependency more difficult at the dependency's completion. This finding is important to the development of the DLT, expanding it to take both referential and syntactic processing into account when predicting complexity effects. This work also provides new evidence about the relative processing loads incurred by multiple referential processes, new evidence concerning the mechanisms underlying referent accessibility and new evidence about the allocation of resources to different subprocesses of the human language comprehension system.

Thesis Supervisor: Edward A. F. Gibson

Title: Associate Professor of Cognitive Science

Acknowledgments

My experience in graduate school was very enjoyable, due in no small part to the contributions of the following people. I would like to thank:

Ted Gibson, from whom I have learned more about science and writing and teaching and research than I once thought there was to learn.

My thesis committee: Barbara Grosz, Neal Pearlmutter, Steve Pinker and Ken Wexler, for comments, ideas, suggestions and support.

Daniel Grodner, because grad school would have been miserable without him.

Duane Watson, because he lowers stress levels wherever he goes.

Past and present members of Tedlab: Edson Miyamoto, San Tunstall, Kara Ko, Florian Wolf, Doug Rohde, Timothy Desmet, Evan Chen, Alina Sheyman as well as UROPs Laura Dean, Natasha Olenchanski, Matthew Cain and Lauren Tsai, for running subjects on my experiments, helping write materials, talking about ideas and generally making Tedlab a fun and vibrant place to work and study.

The MIT women's ultimate team and my various Boston ultimate teams, because working towards goals with such empowered women provided important balance in my life.

Dr. and Mrs. Gerald Burnett , who generously endowed a fellowship which supported my fifth year of study.

Jonah Warren, because inspiration and wisdom can come from unexpected sources, even little brothers.

Nana and Popi, whose love, laughter and pride in me made being near Lynn one of MIT's biggest assets.

David Offner, whose unflagging encouragement, love, belief in me and help with dinner made grad school and the thesis writing process fly by.

Mom and Dad, because in addition to being sources of emotional support, recipes, help with semantics homework, movie reviews, advice about academia, life guidance and unconditional love, they are my favorite people.

This research was supported by funds from the NSF RTG grant to the Brain and Cognitive Science and Linguistics departments at MIT, by a NSF graduate fellowship and by a Burnett Fellowship.

Table of Contents

Introduction

0.1 Introduction

The study of sentence interpretation is relatively new compared to other investigations of language. Philosophers have been studying sentence meaning since the time of Aristotle, but only in the past thirty years have scientists begun to investigate the way language users dynamically build sentences and meanings out of strings of words. Work in sentence interpretation has mirrored work in formal linguistics in that it considers processing at multiple levels of language. Psycholinguists investigating syntactic processing study the way language comprehenders assign structure to strings of words by combining words into phrases, phrases into clauses and clauses into sentences. Psycholinguists investigating semantic processing study the way language comprehenders assign meanings to strings of words. Psycholinguists studying pragmatics study the way language comprehenders relate individuals and events mentioned in linguistic input to context and to the contents of their memories.

The methods that psycholinguists have at their disposal to study these processes are limited. Like any other mental process, sentence comprehension is opaque. The input to the comprehension system is a string of visually presented words or auditorially presented sounds. The output is complex conceptual information that can be used in reasoning, stored in memory, or related to further input. The process by which strings of words are transformed into conceptual structures cannot be directly observed, making the behavioral repercussions of small changes to the input the main source of information about the components of the human sentence processing mechanism (HSPM). Psycholinguists have many ways of measuring the behavioral repercussions of changes to the HSPM's input. Reading times and eye-movement patterns reflect increases in processing load at particular words or sentence regions. Word-identification latencies can indicate the accessibility of particular words and concepts. Measures of brain activity show characteristic areas and time courses of activation for different levels of processing. These kinds of data provide indirect information about the internal workings of the HSPM.

There are few theories of on-line interpretation that are broad enough to encompass syntactic, semantic and discourse processes. The few works that attempt to address sentence processing comprehensively generally provide frameworks rather than fully developed theories, because so much about the HSPM remains unknown. In a 1999 monograph, Frazier sketches the outline of a partially modular theory that includes syntactic, semantic and pragmatic subprocesses. MacDonald, Pearlmutter and Seidenberg (1994) describe a system of on-line interpretation based on the development of lexical, syntactic and discourse representations through the interaction of multiple constraints. These frameworks are exemplars of two kinds of research programs currently dominant in the study of sentence processing. These research programs diverge in two ways: they assume different underlying architectures for the HSPM and they assume different bases for the output of the HSPM. Frazier's approach assumes that the HSPM's architecture is modular, such that syntactic information is processed first, and semantic and pragmatic processing build upon the output of the syntactic processes. In her model, comprehension processes operate according to principles that are tightly bound to theories from formal linguistics. MacDonald's approach assumes a non-modular architecture, where multiple constraints such as word frequency and contextual plausibility interact to determine the output. MacDonald hypothesizes that these constraints do not necessarily have their roots in linguistic principles, but rather in the statistical properties of the previous input to the HSPM.

Another approach to understanding the HSPM is to focus on explaining the computational resource demands created by component processes within the HSPM. This approach ties predictions about the HSPM's output to patterns of complexity discovered in previous behavioral evidence. For example, if sentences with particular kinds of structures tend to cause difficulty for the HSPM, those structures are studied for similarities. When similarities are found, there is an attempt to explain them in terms of some process that is already known to place a high demand on computational resources. Further predictions deriving from that process are then tested.

Sentence (1) provides a concrete example of a sentence that theories based on formal syntax, experience and resource usage all predict will be difficult to process, though for different reasons.

1. The horse raced past the barn fell.

Sentence (1) is a temporarily ambiguous sentence leading to a garden path effect, which means that the first reading that most readers follow is the incorrect one. In (1), readers generally report confusion at "fell," because they initially process "horse" as an agent subject and "raced" as a main verb, but then have no way to connect "fell" into the structure that they have built so far. At this point the reader has been lead down the proverbial garden path and must go back and reanalyze the beginning of the sentence. The correct analysis has "raced" as beginning a reduced relative clause meaning the same thing as "The horse that was raced past the barn".

The garden path theory (e.g. Frazier, 1987), a syntax-based parsing theory, uses the principle of Minimal Attachment to explain the tendency for "raced" in (1) to be read as a main verb rather than as a verb in a reduced relative clause. Minimal Attachment requires the minimization of the number of syntactic nodes introduced at every step during syntactic structure building. Since attaching a verb as a main verb requires fewer new syntactic nodes than attaching a verb in an embedded clause, the main verb attachment is preferred. An experience-based theory, like the one proposed by MacDonald and Christiansen (2001) would explain the same tendency as resulting from the fact that readers see a much higher percentage of these sorts of ambiguous structures resolved as main verbs than as reduced relatives. Readers are hypothesized to follow readings that are more frequent in their experience, and so they attach the verb as a main verb. A resource-based theory such as the Dependency Locality Theory (DLT) (Gibson, 1998) would explain this tendency in a third way. According to the DLT, this sentence is difficult because the potential resource usage of one reading is higher than the potential resource usage of the other reading. Gibson and colleagues noted that structures with fewer shorter dependencies were easier to read than structures with more long dependencies, and reduced relative clauses introduce more long dependencies than do main verbs. If there is a general system-wide tendency to try to minimize complexity and resource usage, then the main verb reading will be preferred over the reduced relative reading because there are fewer long dependencies in the main verb reading.

As shown above, resource-based theories can make predictions about ambiguity preferences, but their main purpose is to predict points of complexity in the processing of sentences. Most complexity theories assume a resource limitation on the HSPM and predict that sentences whose resource requirements pass that threshold will be unprocessable. One example of such a theory is Kimball's (1973) principle of two sentences. Kimball hypothesized that the maximum number of uncombined syntactic subjects that could be processed was two. According to this theory, doubly nested sentences are unprocessable because they require three syntactic subjects to be held in memory before any verbs are processed. Though most early complexity theories were concerned with predicting the threshold between processable and unprocessable sentences, newer complexity theories have focused on gradations in complexity at individual words in a sentence.

One positive ramification of developing complexity theories is that the kinds of structures that are used to test complexity theories are often different from those used to test ambiguity resolution theories or to test the HSPM for a modular or non-modular architecture. This is important in the field of sentence processing, where much of the recent behavioral data gathering has been focused on a limited set of structures that may help to differentiate between the different possible underlying architectures that have been proposed for the HSPM. With so much still unknown about language processing, it seems that the strategy of gathering as much data from as many varied structures as possible will be the most productive. The more structures are studied, the higher the chance that informative patterns and similarities will emerge in the behavioral data, leading to better and more comprehensive theories.

0.2 A complexity theory: the contributions of this thesis

Predicting patterns in the behavioral effects of language comprehension on the basis of the resource usage of comprehension processes is not a new idea. Chomsky and Miller (1965), Kimball (1973), and others have proposed theories locating the source of doubly nested sentences' complexity in the resource limitations of syntactic building processes (see Gibson, 1998 for a review). More recently, Gibson and colleagues have undertaken a research program aimed at developing a comprehensive theory of sentence complexity, based on characterizing the computational resource demands of processes within the HSPM (Gibson, 1998, 2000; Grodner, Gibson & Tunstall, 2001, among others). Gibson and colleagues attribute patterns of processing load to the memory demands of basic language comprehension processes. The resource demands that Gibson and colleagues have investigated most closely so far are those associated with the process of linking new words into linguistic representations. Their results have demonstrated that the difficulty of adding a new word to a linguistic representation is strongly affected by the distance between the new word and the location in the partially-built linguistic structure to which the new word must be integrated (Gibson, 1998; Grodner, Watson & Gibson, 2000). Gibson and colleagues call the linguistic complexity theory they are developing the dependency locality theory (DLT).

This thesis presents a series of experiments that have played an important role in the development of the DLT. The fundamental hypothesis of the DLT is that processing load increases as the distance between syntactically dependent elements in a sentence increases. But this hypothesis is incomplete without a metric for computing distance. The experiments in this thesis suggest, test and support the claim that the metric for computing distance in the DLT should include a cost for referential processing. The more referential processing that is necessary between the endpoints of a syntactic dependency, the more difficult that dependency seems to be to build. This simple hypothesis is important and exciting for a number of reasons. First, as already stated, it has been important in the development of a new complexity theory. It has expanded a complexity theory based primarily on syntax and incorporated a measure for referential processing into that theory. The finding that distance is best captured by a combination of structure building costs and referential access costs gives this complexity theory a wider scope than past complexity theories, because it considers resource usage from multiple comprehension processes in its prediction of complexity patterns. Second, it provides a way to test whether the resource stores for different linguistic subprocesses overlap or are separate. If the resource demands of referential processes affect the amount of resources available for later syntactic processes, that would indicate that the two processes share resources at some level in the system. This is important in understanding the extent of the interaction of subprocesses in the HSPM. Third, this hypothesis allows the prediction of the reading time effects of referential costs at a finer grained level than has traditionally been investigated. Most of the past work in referential processing has measured reading time effects across whole sentences, because the exact location of reading time effects could not be predicted or was not relevant. But this hypothesis explicitly predicts that the ramifications of referential processing resource usage will be most apparent at the endpoints of syntactic dependencies that cross the referential processing.

The exposition of this thesis uses the DLT as a framework within which to explain and interpret experiments. The following section introduces the DLT as presented in Gibson (2000). But because this thesis work has been part of the development of the DLT, the referent of the name DLT will change slightly over the course of the thesis. This will primarily happen between the first and second chapters, as some of the experiments in the second chapter test the modified version of the DLT supported by the results in the first chapter.

0.3 The Dependency Locality Theory

The DLT predicts reading times and intuitive complexities based on the resource usage of basic memory and sentence comprehension processes. According to the DLT, sentence comprehension involves at least two components of computational resource use: 1) structural integration: connecting an input word into the current structure, and 2) structural storage: keeping track of the incomplete structural dependencies in the current structure. The DLT is distance-based; the cost associated with performing a structural integration increases with the distance between the elements being integrated.[1] In order to computationally motivate this hypothesis, Gibson proposed that the process of sentence comprehension is activation-based. Integrating a newly input maximal projection, XP, headed by h_2, with a previous syntactic category headed by h_1 (as in Figure 1) involves retrieving aspects of h_1 from memory.

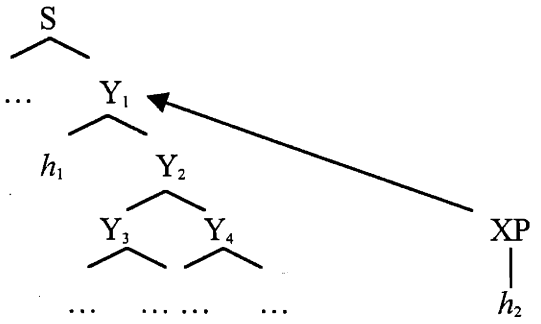

Fig. 1 Structural integration of a maximal projection XP of a newly input head h_2 to an attachment site headed by a head h_1 in the structure for the input so far.

In an activation-based framework, this process involves re-activating h_1 to a target threshold of activation. Because of the limited quantity of activation in the system, h_1's activation will decay as intervening words are processed and integrated into the structure for the input. Thus, the difficulty of the structural integration depends upon the complexity of all aspects of the integrations that have taken place in the interim since h_1 was last highly activated.

As an initial simplifying hypothesis, Gibson assumed that the distance metric makes a binary distinction between integrating across newly created discourse referents and integrating across already existing referents, such that there is one unit of cost to integrate across new referents and there is no cost to integrate across old referents:

2. DLT linguistic integration cost (Gibson, 2000): The structural integration cost associated with connecting the syntactic structure for a newly input head h_2 to the projection of a head h_1 that is part of the current structure for the input is dependent on the complexity of the computations that took place between h_1 and h_2. For simplicity, 1 unit of cost will be counted for each new discourse referent in the intervening region.

[1] Gibson (1998) also presents a version of this theory (the syntactic prediction locality theory) in which storage cost increases over distance. However, there are empirical and conceptual problems with this alternative (Gibson & Ko, 1998). Consequently, I will focus on the version of the theory in which only integration costs increase with distance. Gibson (2000) refers to this version as the dependency locality theory, and I will follow this convention here.

Thus, integrating a new word *w* across linguistic material indicating a new discourse referent is more costly than integrating *w* across linguistic material referring back to a pre-existing discourse referent. This formulation of integration cost is based on the assumption that building new discourse structure (e.g., a representation for a discourse referent, such as a person or object in the real world; Kamp, 1981; Heim, 1982) requires more resources than accessing previously constructed discourse structure. This assumption has considerable support in the discourse processing literature, as shown in experiments on unambiguous sentences by Haviland and Clark (1974), Haliday and Hasan (1976), Garrod and Sanford (1977, 1994) and Murphy (1984) among others. Evidence for this hypothesis from the processing of ambiguous structures is provided by Crain and Steedman (1985) and Altmann and Steedman (1988), who argue that when the processor is faced with ambiguity, it follows the reading requiring less new discourse structure.

The comprehension difficulty at a word in a sentence (e.g., as measured by reading times) is assumed to be determined by a combination of integration cost and storage cost, together with other factors that have been shown to be important in on-line sentence comprehension (see Gibson & Pearlmutter, 1998; Tanenhaus & Trueswell, 1995 for summaries of relevant results), such as lexical frequency, contextual plausibility, and reanalysis difficulty. Reading-time data in support of the distance-based integration hypothesis with respect to unambiguous sentence materials is provided by Gibson and Ko (1998) (reported in Gibson, 1998, 2000) and Grodner, Watson and Gibson (2000). Reading-time data in support of the distance-based hypothesis with respect to ambiguous sentence materials is provided by Altmann, VanNice, Garnham and Henstra (1998); Gibson, Pearlmutter, Canseco-Gonzalez, Hickok (1996), Gibson, Pearlmutter and Torrens (1999); and Pearlmutter and Gibson (2001).

Finally, it is assumed that the overall intuitive complexity of a sentence depends to a large degree on the maximum intuitive complexity incurred at any state during its processing. Doubly nested sentences are good benchmarks for theories of intuitive complexity, because they are very complex but their complexity lessens when their most embedded subject is a pronoun (Bever, 1974; Kac, 1981). This can be see in examples (3) and (4), where (3) is a doubly nested sentence that is very difficult to process and (4) is a doubly nested sentence that is somewhat easier to process.

3. The nanny [who the agency [which the neighbors recommended] sent] was adored by all the children.
4. The nanny [who the agency [which you recommended] sent] was adored by all the children.

When the most embedded subject is changed from a definite description (e.g. “the neighbors” in (3)) to a pronoun (e.g. “you” in (4)), the sentence becomes easier to understand. The DLT explains this phenomenon as being a result of the interaction of resource demands of discourse and structural processes, as described in the formulation of integration cost. For sentences such as (3) and (4) the most important factor contributing to on-line complexity is the integration cost at the verbs, where the maximal integration costs occur for these sentences. There are two integration steps which take place when the most embedded verb "recommended" is processed in (3): integrating the verb "recommended" to its preceding subject "the neighbors"; and integrating the object position of the verb "recommended" to the filler "who", which is co-indexed with the NP "the agency". For the subject-verb integration, only the event referent corresponding to the verb "recommended" has been introduced since the attaching position ("the neighbors") was last processed. For the filler-object-position integration, two new discourse referents have been processed since the attaching position was last activated: the verb "recommended", and the NP "the neighbors". Two similar integration steps take place at the verb "sent": a subject-verb integration crossing three new discourse referents, and a filler-object-position integration crossing four new discourse referents. There is only one integration when the last verbal region "was adored" is processed - a subject-verb integration - and this integration crosses five new discourse referents.

Each of the integration steps that crosses the most embedded subject is less costly in (4), where "the neighbors" is replaced with the indexical pronoun "you". First- and second- person pronouns require little processing, because "I", "you" and "we" are anchored in the discourse (Enc, 1983). An NP is anchored in the discourse if both the speaker and the listener know its intended referent. Even in a null context, first and second person pronouns are assumed to be part of the deictic frame, because in every discourse there is a speaker/writer and a listener/reader (Chafe, 1987). Thus, integrating the object position of the verb "recommended" to the filler "who" crosses only one new discourse referent in (4), rather than two for the corresponding integration in (3). Furthermore, the integration steps at the verb "sent" in (4) cross only two and three new discourse referents, rather than three and four new discourse referents in (3). And the subject-verb integration at "was adored" in (4) crosses four new discourse referents rather than five, as in (3). The decrease in maximal integration cost from seven cost units in (3) to five cost units in (4) predicts a corresponding decrease in intuitive complexity between the two sentences.

The first experiments in this thesis test the DLT’s prediction that doubly nested sentences with pronouns in the most embedded subject position are easier to process than doubly nested sentences with other kinds of noun phrases (NPs) in the most embedded subject position. Later experiments test the DLT's account of this phenomenon in a much more extensive way .

In order to claim that a pattern of complexity is due to the resource demands of particular comprehension processes it is important to show that: 1) there is evidence for the existence of those comprehension processes and 2) there is evidence that they drain processing resources. To that end, the remainder of this introduction will review evidence from the literature suggesting that representations of the linguistic structure and the discourse model are built and used during language comprehension and more specifically, that introducing referents to or accessing referents from the discourse representation are resource-demanding processes. Claims about the computational resource demands incurred during the process of linking newly input words into a partial structure for a sentence will not be addressed in this introduction, for exactly such claims were made and supported in Gibson (1998).

0.4 Literature Review

The following survey of the discourse processing literature reviews evidence about the existence and use of linguistic and conceptual representations during processing. This leads into a discussion of the construction of discourse representations, especially evidence concerning the addition of information to the conceptual model through inferences. Inferences play an important role in Crain and Steedman (1985) and Altmann and Steedman's (1988) referential theory, a theory of syntactic ambiguity resolution based on minimizing additions to discourse models. After this treatment of the construction of discourse models, the questions of how information is accessed from a discourse model and what the resource costs of these accessing processes are, are addressed in a discussion of research into the accessibility of different types of referents in discourse. The review ends with a discussion of the time courses of referential processing and structural processing.

0.4.1 The Existence of Linguistic and Conceptual Representations

One of the fundamental assumptions underlying the formulation of integration cost in the DLT and the arguments in this thesis is that during language comprehension, comprehenders create and use both a linguistic and a conceptual representation of the input. The DLT assumes a linguistic representation of the input, because it predicts a cost for creating syntactic structure from the combination of existing structure and the structure associated with newly input words.

The version of the DLT from Gibson (1998) does not necessitate an additional conceptual representation, but the versions of the DLT argued for in this thesis assume a conceptual representation because gradations in referent status are hypothesized to affect complexity.

There is an extensive literature focusing on the kinds of representations used during language comprehension. Sentences can be interpreted with reference to the physical context, the linguistic context, or to a conceptual representation that has been constructed based on the previous discourse (Garnham, 1987). Much of the study of representation use has involved anaphors, which are words that take their reference from antecedents in context. Linguists have proposed that there are two types of anaphors: surface anaphors and deep anaphors (Hankamer & Sag, 1976, Sag & Hankamer, 1984). Surface anaphors are hypothesized to have antecedents in a purely linguistic representation, while deep anaphors take their antecedents from a conceptual representation. For example, in (5) there is an elided verb phrase in the second sentence, "did", that takes "take the oats down to the bin" as its antecedent.

5. Someone had to take the oats down to the bin. So Sally did.
6. ?? The oats had to be taken down to the bin. So Sally did.
7. The oats had to be taken down to the bin. So Sally did it.

Here, "did" is a surface anaphor, because the verb phrase antecedent for this structure must be in the linguistic record. If the verb phrase is not found in the linguistic representation, as in (6), the discourse that results is not coherent. But in a structure with a deep anaphor, there is no need for an antecedent in the linguistic structure. (7) is a coherent discourse because "it" is a deep anaphor and takes the concept take-the-oats-to-the-bin as an antecedent. Psycholinguists have looked for evidence of separate linguistic and conceptual representations in processing by comparing the processing of deep and surface anaphors (Murphy, 1985; Tanenhaus & Carlson, 1990; Garnham & Oakhill, 1989, 1990). Studies by Tanenhaus and Carlson (1990) looked at the effects of syntactic parallelism on the two types of anaphors. Their results showed that surface anaphors were easier to process when the linguistic antecedent was syntactically parallel than when it was not, and that syntactic parallelism did not affect the processing of deep anaphors to nearly the same degree. They concluded that deep and surface anaphors do draw on separate representations. But this is hardly a settled question. Garnham and Oakhill (1989, 1990) remark on the fact that participants in Tanenhaus and Carlson's (1990) study were able to interpret the majority of surface anaphors with non-parallel antecedents even though they judged more of them nonsensical. Garnham and Oakhill argue that the requirement of parallelism for surface anaphors is due to prescriptive grammar and does not reflect a true difference in the way anaphors are interpreted.

The investigation of verb phrase anaphors has not resolved the question of whether there are separate linguistic and conceptual representations accessed by different types of anaphors. So researchers have turned to noun phrase anaphors for more evidence. Gernsbacher (1991) provided psycholinguistic evidence that pronouns are conceptual/deep anaphors. In her example: "I need a plate. Where do you keep them?", there is no plural linguistic antecedent for the plural pronoun. The pronoun refers to a set of plates inferred to be in the default context. Gernsbacher (1991) demonstrated that inferences like that from "a plate" to a set of plates were automatically and easily made, and in fact, plural pronouns in these types of constructions were easier to process than singular pronouns.

There is also evidence suggesting that a linguistic context is used during the processing of pronouns. Garnham et al. (1995) studied anaphor resolution in languages where pronouns can carry either the syntactic gender marking of their antecedent or the natural gender of their antecedent. They reported that syntactic gender matches speeded pronoun resolution. They argued that comprehenders use both a linguistic representation and a conceptual representation to interpret both deep and surface anaphors. This is consistent with linguistic theories that argue that syntax can rule out potential antecedents for pronouns (for example through the binding principles), but that the actual selection of an antecedent must take place at a pragmatic rather than syntactic level (e.g. Reinhart, 1983). Although theories about the validity of distinctions between deep and surface anaphors differ, this research has led to a general consensus in the field that both conceptual and linguistic representations are used and updated during processing (Garnham & Oakhill, 1989, 1990; Lucas, Tanenhaus & Carlson, 1990; Frazier, 1999).

0.4.2 Characterizations of Linguistic and Conceptual Representations

These lines of research investigating anaphors indicate that people use both linguistic and conceptual representations during language comprehension. But what are these representations like? And what kinds of processes are necessary to create and update them? Many researchers who discuss linguistic representations take the linguistic representation to be Logical Form (see e.g., Frazier, 1999; Garnham & Oakhill, 1989). According to linguistic theory, Logical Form is the linguistic representation which is hypothesized to be used at the interface between the linguistic system and the conceptual-intentional system (Chomsky, 1995). It is a representation containing the syntactic and logico-semantic relations between words in a sentence.

Researchers investigating the conceptual representation have tended to focus on its relation to other memory processes rather than its structure. Kintsch (1998) hypothesizes that conceptual representations are segments of long-term memory that are activated by the working memory representations resulting from reading. When activated by working memory, these long-term memories effectively become completely available to the working memory system. Kintsch's conceptual representation is a selection of information from long term memory relating to the discourse that is currently active in the working memory system. The mental models hypothesis, a theory of mental representations used in reasoning and language, is slightly less explicit about exact memory mechanisms, but seems to assume a system like the one Kintsch proposes. Garnham (1997) describes mental models as representations that can include information from perception, language and general world knowledge. The constructs used in building a mental model are taken to be the same elements that make up an ontology of things that are in the world. Borrowing from the ontology proposed in various formal semantic theories (see e.g. Heim & Kratzer, 1998; Kamp & Ryele, 1993), they include entities, events, states, processes and properties. One way to add elements to a mental model is through the process of interpreting language. Noun phrases generally indicate entities, verb phrases indicate events or states and adjectives indicate properties. The addition of elements to a mental model requires resources, as discussed in the following section.

0.4.3 Component processes of representation building

0.4.3.1 Introducing a new referent

There is evidence that the process of mapping from words to referents and creating new memory representations requires computational resources. Murphy (1984) reports a series of experiments showing that introducing a new referent to discourse requires more processing resources than accessing a referent that is already available in the discourse. In one experiment, Murphy introduced a referent in a context sentence and then tested reading times over a target sentence that either referred back to that referent or introduced a new token of that type of referent. He accomplished this status manipulation by changing the form of an NP in the target sentence. For example, when the NP in (8b) is "the truck", the truck referred to is the same truck that was introduced in (8a). When the NP in (8b) is "a truck", the truck referred to is different from the truck that was introduced in (8a).

8a. Though driving 55, Steve was passed by a truck.
8b. Later, George was passed by {a, the} truck too.

Murphy measured reading times over the second sentence and found that the sentence was read more slowly when it introduced a new referent, "a truck," than when it referred to an already existing referent, "the truck." These results suggest that introducing a new referent is a costly process.

0.4.3.2 Inferring a new referent

It seems clear that discourse representations represent explicit linguistic and perceptual input, but it is less clear what sorts of inferences from general knowledge are incorporated into the representation. Researchers are engaged in a long-standing debate (see McKoon & Ratcliff, 1992) about what kinds of inferences are made during the building of a discourse model. Inferences are important to understand from the perspective of a complexity theory because making inferences often involves building new discourse structure, which can drain computational resources. Some of the earliest work investigating inferences during reading was done by Haviland and Clark (1974). Haviland and Clark's experiment was designed to determine whether the processor accesses a linguistic or conceptual representation when searching for an antecedent for a definite description, but the lasting impact of this work has come from its demonstration that inferring new entities and relations into a conceptual discourse representation can be a resource draining process. Haviland and Clark gave subjects pairs of sentences like the following, where in one condition, (9a,c), the referent of the definite in the second sentence was explicitly introduced in the linguistic record of the first, and in the other condition, (9b,c), the referent of the definite in the second sentence could be inferred into the discourse model established by the first sentence:

9a. Mary unpacked some beer.
9c. The beer was warm.

9b. Mary unpacked some picnic supplies.
9c. The beer was warm.

Haviland and Clark found that subjects took significantly longer to read (9c) when it followed (9b) than when it followed (9a). They assumed that reading a definite description, i.e. "the beer," triggered a search of the linguistic representation for an appropriate antecedent. Since there was no explicit linguistic antecedent in the discourse, they argued that the reader must create an antecedent for the description in the conceptual representation and relate the newly formed antecedent to the prior discourse. They called this process a "bridging inference". According to Haviland and Clark, it is the bridging inference introducing an antecedent for "the beer" that causes the slow down in (9c) when it follows (9b).

Garrod and Sanford (1982) took issue with Haviland and Clark (1974) and noted that most definite descriptions do not have explicit linguistic antecedents in prior discourse, yet are processed easily. They performed a study using the same method as Haviland and Clark, but in their experiment the condition without an explicit linguistic antecedent for the definite description, i.e. (10b,c), had a closer conceptual relationship to the referent of the definite in the target sentence.

10a. Keith took his car to London.
10c. The car kept overheating.

10b. Keith drove to London.
10c. The car kept overheating.

So, for example, in (10b), "the car" is not explicitly introduced, but it does naturally fill the role of the vehicle Keith was driving to London. Garrod and Sanford tested subjects' reading times on (10c) after they read either (10a) or (10b), and found no difference. From this they argued against Haviland and Clark's claim that first a purely linguistic representation is searched for a referent, and then, if that fails, a bridging inference mechanism is activated to create a new referent. According to Garrod and Sanford (1982), Haviland and Clark's results came about because the definite description triggered a search of the conceptual representation, but the relationship between the definite and the rest of the context was not close enough to have triggered an automatic inference introducing its referent. So, in Haviland and Clark's example, beer is not a default element in the schema for "some picnic supplies." These experiments bring up an important question, which is how to characterize the situations in which an inference will or will not be automatically made. Being able to predict the automaticity of an inference is important for a complexity theory, because automatic inferences are easier to perform and require fewer resources than non-automatic inferences.

Building on issues raised by Garrod and Sanford's (1982) work, Mauner, Tanenhaus and Carlson (1995) provide evidence that the implicit arguments of verbs are automatically encoded. But almost every other category of inference that has been investigated has been shown not to be automatic. Lucas, Tanenhaus and Carlson (1990) and McKoon and Ratcliff (1992) report experiments indicating that instrument inferences are not automatic. So, for example, when reading "He swept the floor every week on Saturday," subjects did not seem to enter "broom" into their discourse model. Garnham and Oakhill (1988) argue that entities from verbs corresponding to nouns are not automatically encoded. Participants in their experiment had more difficulty with "John dreams a lot but he never remembers them," than with "John has lots of dreams but he never remembers them." In the face of a general lack of evidence for automatic inferencing, Garnham and Oakhill (1988) hypothesized that comprehenders only introduce new elements into discourse models when absolutely necessary. They claimed that the only inferences that are automatically made are those stored in the lexicon and those that are necessary for comprehension. McKoon and Ratcliff (1992) expanded on this and proposed the minimalist hypothesis. The minimalist hypothesis asserts that inferences are only automatically made in situations when the information necessary for the inference is quickly and easily available or when the inference is required to establish local coherence between sentences. The minimalist hypothesis and Garnham and Oakhill's hypothesis both suffer from a lack of specific detail about what information is quickly and easily available and how to determine whether an inference is necessary for local coherence or comprehension.

McKoon and Ratcliff contrasted the minimalist hypothesis with a constructivist view, which asserts that readers make many inferences during language comprehension. The constructivist view claims that many on-line inferences are made as "readers attempt to construct a meaningful referential situational model that addresses the reader's goals, that is coherent, and that explains why actions, events and states are mentioned in the text" (Graesser, Singer & Trabasso, 1994, p. 372). Chafe (1987) and Garrod and Sanford (1982, 1994) agree with some of the claims of the constructivists and argue that linguistic input can cause entire schemas or scripts from the comprehender's long term memory to be retrieved. Subsequent linguistic input is then interpreted with respect to the concepts in those schemas. Experimental support for the idea that linguistic input can activate entire schemas comes from Gernsbacher (1991). She argues that the sorts of discourse inferences she finds evidence for are exactly those based in world knowledge and schemas. For example, she argues that the reason "I need a plate. Where do you keep them?" is coherent is that the reader assumes that the request is taking place in a typical American home, where a set of plates will generally be kept in one location. If "plate" had been replaced with "iron," then the singular pronoun would have been more appropriate because the typical home has only one iron. But if the situation were a department store, then using a plural pronoun would be better, as in "I need an iron. Where do you keep them?" because the department store schema contains multiple irons. In each of these examples, the referent of the pronoun is dependent on the conceptual schema associated with the location at which the utterance is interpreted.

There has been a considerable amount of research and effort directed at attempting to determine what kinds of inferences are automatically made while building a discourse model and what kinds of inferences require extra cognitive effort. But the prediction of which inferences are automatically made is still very much an open problem. For the purposes of building a complexity theory, the evidence presented here suggests that most inferences besides those involving the implicit arguments of verbs require at least some processing resources during their computation.

0.4.4 Referential Theory

Though the question of which inferences are made in a discourse model is still unresolved, it is generally accepted that most inferences require processing resources. Crain and Steedman (1985) and Altmann and Steedman (1988) proposed a theory of syntactic ambiguity resolution, the referential theory, built upon the idea that introducing elements into a discourse model is a costly process and thus is avoided whenever possible. Referential theory is based on a general discourse principle which helps guide ambiguity resolution. The Principle of Parsimony asserts that a reading carrying fewer unsupported presuppositions will be favored over one carrying more, if the plausibilities of the two readings are equal and none of the new presuppositions are in conflict with the current state of the discourse model. The unsatisfied presuppositions of the accepted reading will be added to the discourse representation upon processing. The intuition behind the principle of parsimony is that at points of ambiguity, the processor will choose the reading that minimizes the amount of new information added to the discourse model.

As evidence for the principle of parsimony, Crain and Steedman presented the results of an experiment on the complement/relative clause ambiguity. In the sentence "The psychologist told the wife that he was having trouble with to leave her husband," there is a temporary ambiguity before the verb "to leave." In this sentence, the ambiguity is resolved as a relative clause (RC). The RC modifies "the wife", creating a complex NP. This complex NP introduces a presupposition that there are multiple tokens of the simple NP in the context; in this example, the presupposition is that the psychologist interacted with more than one wife. If the principle of parsimony is correct, then in a context with more than one wife, readers should expect the ambiguous clause to be resolved as a relative clause. But the clause could also have been resolved as a complement clause (CC), as in "The psychologist told the wife that he was having trouble with her husband." When the ambiguous clause is resolved as a CC, it introduces no presuppositions of multiple wives, and in fact would be infelicitous in a context with more than one wife. If readers follow the principle of parsimony, then providing a reader with a context supporting a different set of presuppositions should change the reader's initial interpretation of the ambiguous clause. Crain and Steedman showed that by providing subjects with a context including one token of the simple NP subject of the ambiguous clause, they could induce processing difficulty in the sentences that were resolved as relative clauses, and by providing a context with two tokens, they could induce processing difficulty in sentences resolved as complement clauses.

Complement -inducing Context

A psychologist was counseling a married couple. One member of the pair was fighting with him but the other one was nice to him.

Relative-inducing Context

A psychologist was counseling two married couples. One of the couples was fighting with him but the other one was nice to him.

Complement target sentence

The psychologist told the wife that he was having trouble with her husband.

Relative target sentence

The psychologist told the wife that he was having trouble with to leave her husband.

When the target followed the appropriate context, subjects had no trouble processing it. When the target followed the inappropriate context, subjects had difficulty after the ambiguous region. This evidence suggests that information in the discourse representation can affect ambiguity resolution, such that alternatives requiring more additions to the representation are disfavored.

From the evidence presented so far, it seems clear that there is a conceptual discourse representation that is important in sentence processing. This discourse representation is both an aid to and a product of language understanding. Adding explicitly introduced new referents to this representation is a resource-demanding process. Inferring non-explicitly introduced entities and relations usually requires computational resources as well, though some inferences seem to be automatic (Garrod & Sanford, 1982; Mauner et al., 1990) and the conditions under which inferences are or are not made are still not well characterized. The evidence presented thus far has related to the existence and construction of discourse representations. The next section will address questions of how information from discourse representations is retrieved and what cost ramifications retrieval processes have.

0.4.5 Information retrieval from conceptual representations

Investigations into retrieval from discourse representations have focused on what makes some entities in a model more accessible than others. Researchers in linguistics, psychology and computational linguistics have identified recency of mention, grammatical role, thematic role, focus status, and relation to the intentions or global discourse structure as the basic determinants of accessibility (see Almor, 1999; Ariel, 1990; Arnold, 2001; Morrow & Greenspan, 1989; Grosz, Joshi & Weinstein, 1995 among others). In linguistics, much of the research in this area has involved investigating the distribution of anaphors in natural language corpora (Ariel, 1988, 1990; Arnold, 1998; Gundel, Hedberg & Zacharski, 1993). Linguists have made use of the fact that a referent can be referred to with many different types of noun phrases or referential forms, which can be categorized by the amount of information each form incorporates about its referent (e.g. Prince, 1981; Marslen-Wilson et al., 1982). These categories form a hierarchy that begins with zero anaphors and pronouns, which provide little information about their referent besides number and gender, and extends through demonstratives to full definite and indefinite noun phrases, which provide a relatively large amount of information about their referents. Linguists studying accessibility identify all of the anaphors in a natural text and determine their antecedents. They then correlate the referential forms of the anaphors with the recency of mention, grammatical role, and focus-status of their antecedents. Entities that have been more recently mentioned, focused or referred to in a subject position tend to be referred to with referring forms that contain less information. Assuming that language producers tailor the amount of information they provide in an NP to the difficulty of accessing the referent to which the NP refers, these findings begin to give a hint as to how the representation might be structured.

Psycholinguists have addressed the question of referent accessibility using reading time, question-answering and priming experiments. The evidence from these studies is vital to the hypotheses in this thesis, because unlike the linguistic studies, many of these studies directly test the processing load imposed by accessing differently accessible referents from the discourse model. The previously discussed experiments from Murphy (1984), Haviland and Clark (1974), and Garrod and Sanford (1982) all touch on this issue, but they compare the processing of referents from the discourse model with referents which may or may not be in the discourse model yet. Experiments comparing the accessibility of referents that are already part of the discourse model have shown many factors to be important in accessibility. Sanford (1989) discusses a number of studies showing that main characters in a discourse are more easily accessed than secondary characters in a discourse. Sanford, Moar and Garrod (1988) found that referents introduced by a proper name are more accessible than referents introduced using a definite description. Gordon, Grosz and Gilliom (1993) showed that referents referred to in the subject position of a previous sentence were more accessible than referents referred to in object position. Further experiments in the same paper showed that first-mentioned referents were more accessible than referents mentioned later in a sentence. Chambers and Smyth (1998) found that in certain syntactic environments, referents in parallel positions were more accessible than referents in non-parallel positions. McKoon, Ratcliff, Ward and Sproat (1993) presented evidence that referents and modifiers in syntactically salient positions were more accessible than referents and modifiers in non-salient positions. They tested the relative accessibility of direct objects and indirect objects as well as elements in a main clause and a modifying clause. McKoon, Ward, Ratcliff and Sproat (1993) looked for effects of syntactic and pragmatic factors on accessibility within a single experiment. They found that referents introduced as a direct object were more accessible than referents introduced in a verbal compound and that referents that were more closely related to the topic of a passage were more accessible than referents that were not closely related to the topic of a passage. All of these studies, and others like them, show that the process of accessing referents from a discourse model requires different amounts of computational resources depending on the syntactic, semantic and pragmatic status of the referent.

0.4.6 The time course of structural and discourse processes

If the resource demands of discourse processes affect the resources available for linguistic structure building, as this thesis argues, it must be the case that discourse processes and structure building processes are carried out in similar time scales. The discourse processes that are hypothesized to contribute to integration cost must occur before the end of the processes that integrate the right endpoints of the long distance structural dependencies that they are predicted to affect.

There is evidence supporting immediate anaphor resolution, as well as evidence suggesting that anaphor resolution can sometimes be delayed for a few words (see Garrod &

Sanford, 1994 for an overview). Studies using priming techniques have shown that in many cases, an antecedent is more activated immediately after an anaphor than before (e.g. Gernsbacher, 1989). Self-paced reading studies also suggest that there are immediate automatic checks for co-reference and attempts at reference resolution upon reading noun phrases (see Sanford, 1989 for a review). Garrod et al. (1994) provide evidence that only focused referents are immediately resolved and that evidence of non-focused referent resolution only appears in second-pass reading times, but if accessing non-focused referents is more difficult than accessing focused referents then this the resolution of non-focused referents is predicted to be slower. Van Berkum, Brown and Hagoort (1999) used event-related brain potentials to address the question of whether syntactic and discourse processes can happen during the same time course. Their data suggested that both syntactic and discourse information were used immediately during ambiguity resolution. Though reference resolution is not always immediate in cases with ambiguous anaphors or non-focused referents, evidence suggests that for the most part it is initiated and solved within a matter of hundreds of milliseconds.

0.5 Summary and Overview

The picture of language processing that has emerged over the course of this literature survey is of a system that builds both a linguistic and a conceptual representation incrementally during reading. The discourse processing system relates new linguistic input to the conceptual representation, accessing and building frequently, drawing on information from perception, language and memory to do so. The processes of adding new referents to and retrieving old referents from a discourse representation have been shown to cause increases in reading times and other complexity measures. These processes have also been shown to occur over the same time course as the building of representations for sentence structure, and in fact there is evidence that structure-building processes can be very quickly affected by the state of the discourse representation. As a result, these referential processes are good candidates for inclusion in a comprehensive complexity theory like the one that Gibson and colleagues are developing.

The experiments in this thesis explore and characterize the kinds of discourse processes that cause processing difficulty in the locations predicted by the DLT. Chapter 1 presents five experiments testing the effects of changing the referential form of a subject in a subject-modifying object-extracted relative clause from an NP indicating an easily accessible referent to an NP indicating a less accessible referent. Chapter 2 presents four experiments that use context to directly manipulate the accessibility of a referent in discourse and test the effects of those manipulations on the difficulty of completing dependencies crossing that referent. Ramifications of the results presented in chapters 1 and 2 are discussed in the conclusion.

Chapter 1

1.1 Introduction

The experiments in this chapter test and develop the complexity theory introduced in Gibson (1998). Gibson's DLT incorporates measures of syntactic and referential processing load in its calculations of complexity. The work in this thesis tests Gibson's measures of processing load and further characterizes the effects of referential processing on complexity. According to the DLT, manipulations of referential processing load will have different effects depending on the syntactic environment in which the manipulations take place. The experiments in this thesis made use of two structures that provide similar semantic environments but different syntactic environments in which to test the effects of referential processing. The following paragraphs discuss how the DLT accounts for processing differences between these two structures, and why manipulations in referential processing load are predicted to affect the complexities of the two structures differently.

Every complexity theory must account for the fact that nested (or center-embedded) syntactic structures are more difficult to process than non-nested structures. A syntactic category A is said to be nested within another category B if B contains A, a constituent to the left of A and a constituent to the right of A. Increasing the number of nestings soon makes sentences unprocessable (Chomsky, 1957, 1965; Yngve, 1960; Chomsky & Miller, 1963; Miller & Chomsky, 1963; Miller & Isard, 1964; see Gibson, 1998, for a review of relevant literature). For example, the sentences in (1) are increasingly complex:

1a. The nanny was adored by all the children.
1b. The nanny [who the agency sent] was adored by all the children.
1c. The nanny [who the agency [which the neighbors used] sent] was adored by all the children.

The simple sentence in (1a) contains no nested clauses, and is easy to process. In (1b), the relative clause (RC) "who the agency sent" is nested between the subject noun phrase (NP) "the nanny" and the verbal region "was adored", giving rise to a more complex sentence. The sentence in (1c) is doubly nested, consisting of the nested structure in (1b) with an additional nesting: the RC "which the neighbors used" between the embedded subject "the agency" and the verb "sent". Correspondingly, (1c) is extremely difficult to understand.

Note that there is no temporary ambiguity in (1c), so the processing difficulty associated with this sentence is not related to ambiguity confusions. Second, note that the difficulty in understanding (1c) is not due to lexical frequency or plausibility, because sentence (2) contains the same words and expresses the same ideas as (1c), and yet (2) is much easier to understand:

2. The neighbors used the agency [which sent the nanny] [who was adored by all the children].

The RCs in (2) are not nested as they are in (1c), so (2) is not difficult to understand.

The DLT (Gibson, 1998) accounts for the complexity difference between nested and non-nested structures with a measure incorporating a cost for building syntactic structures. Gibson hypothesized that building long dependencies was more costly than building shorter dependencies and supported the hypothesis with a survey of processing results across a series of different syntactic structures. According to the DLT, nested structures are more difficult to process than non-nested structures because nested structures have more long dependencies than non-nested structures. For example, in (1c) there are four long dependencies: 1) the subject-verb dependency between "nanny" and "was adored", 2) the subject-verb dependency between "agency" and "sent", 3) the filler-gap dependency between the first "who" and the object position of "sent" and 4) the filler-gap dependency between the "which" and the object position of "used". There is an additional subject-verb dependency between "neighbors" and "used," but it is not long distance, as no words intervene. In (2), on the other hand, essentially the same dependencies have to be created, but in each case the endpoints of the dependencies are in consecutive positions. Since none of these dependencies are long, the DLT predicts that structures like (2) will be less complex than structures like (1c).

But the finding reported in the introduction, that intuitive complexity seems to be less for doubly nested sentences with a pronoun in the innermost subject position, c.f. (3a-c), cannot be explained with a purely syntactic mechanism.

3a. The reporter who everyone that I met trusts said the president won't resign yet. (Bever, 1974)
3b. Isn't it true that example sentences that people that you know produce are more likely to be accepted? (DeRoeck et al, 1982)
3c. A book that some Italian I've never heard of wrote will be published soon by MIT press. (Frank, 1992)

Bever (1970, 1974) was the first to note that examples like these were acceptable. He attributed their acceptability to the syntactic non-similarity of the three kinds of subject NPs in the structure. However, the discussion of similarity was not precise enough to explain why a pronoun and not some other dissimilar NP, such as a proper name or an indefinite NP, must occur in the most embedded subject position in order to make the structure acceptable. Kac (1981) was the first to notice the generalization that these structures were acceptable with pronouns in the most embedded position.

This phenomenon inspired the inclusion of a measure of referential processing in the DLT's calculation of integration cost. Because the pronouns in these lower complexity doubly nested sentences are always indexical pronouns, which have referents that are part of the deictic frame and so can be considered old to discourse (Chafe, 1987), it was hypothesized that dependencies crossing new referents are more difficult to create than dependencies crossing old referents. This hypothesis provided a metric for computing the length of dependencies. An example of how integration cost captures the complexity difference between doubly nested sentences with either a pronoun or a definite description in the most embedded subject position was provided in the introduction to this thesis.

According to the DLT, the reason that changing the status of the referent in the most embedded subject position in a doubly nested sentence has a strong effect on intuitive complexity is that that subject position is crossed by four long distance dependencies. If an NP indicating an old referent were substituted for one of the NPs in (2), which has a right branching structure, the DLT would predict no changes in intuitive complexity, because none of the NPs in (2) interrupt long distance dependencies. The DLT predicts that changes in referential processing load will have effects on complexity that are dependent on the syntactic environments in which those changes take place.

This chapter reports five experiments designed to identify which properties of pronouns are important in causing this complexity contrast in nested sentences. Experiment 1 establishes the complexity contrast experimentally, using a questionnaire. Experiment 2 tests two versions of the DLT's integration cost: the version proposed in Gibson (1998) with a binary cost distinction between new and old referents, and a version incorporating continuous costs based on Gundel et al.'s (1993) Givenness Hierarchy. Experiment 3 rotates NPs with low referential processing costs through each of the subject positions in doubly nested sentences to show that intuitive complexity is least when low cost NPs intervene between the most long distance dependencies. Experiments 4 and 5 test the DLT's predictions for singly nested sentences using the self-paced reading method.

1.2 Experiment 1

Experiment 1 tested the claim that doubly nested sentences with pronouns in the most embedded subject position are easier to process than doubly nested sentences with other types of NPs in that position. The DLT's integration cost predicts that doubly nested sentences will be easier to process if their most embedded subject is an old discourse referent than if it is a new discourse referent. Experiment 1 tested four types of NP as the most embedded subject: 1^{st} and 2^{nd} person pronouns, short names, 3^{rd} person pronouns and definite descriptions[2]. 1^{st} and 2^{nd} person pronouns are indexical and refer to the communicator and the communicatee, and so are old to any discourse model. Names and definite descriptions introduce new referents in a null context. 3^{rd} pronouns are usually used to refer to old discourse referents, but in this experiment they were infelicitously used in a null context and as a result introduced new referents. If Gibson's (1998) formulation of integration cost is correct, the conditions with the 1^{st} and 2^{nd} person pronouns should be easier to process than the other conditions. This experiment also addressed the possibility that sentences with pronouns are easier to process simply because pronouns are shorter than other NPs. The short name condition provided a comparison with which to test this possibility.

1.2.1 Method

Participants

Forty native English speakers from the Yale, MIT and SUNY Stony Brook communities were recruited to fill out a questionnaire that took approximately 20 minutes to complete.

Materials

Twenty doubly nested experimental items were tested, each with four conditions. Conditions differed with respect to the subject of the most deeply embedded clause. This subject was varied between a 1^{st} and 2^{nd} person pronoun, a 3^{rd} person pronoun, a short proper name and definite description. For example:

1^{st} /2^{nd} pronoun

[2] Throughout the thesis, experimental conditions containing phrases of the form "the *noun*" will sometimes be referred to as "definite" or "full NP." These names are intended to highlight the property of the definite description that is most relevant in the current experiment and are not intended to be fully correct labels.

4a. The student who the professor who I collaborated with had advised copied the article.

Non-referring 3rd pronoun
4b. The student who the professor who they collaborated with had advised copied the article.

Short proper name
4c. The student who the professor who Jen collaborated with had advised copied the article.

Definite description
4d. The student who the professor who the scientist collaborated with had advised copied the article.

The non-referring 3rd pronoun conditions were constructed so that the pronoun could not be interpreted as referring to one of the other individuals introduced in the sentence. This was accomplished in one of two ways. In some items gender plausibility was manipulated so that the gender of the pronoun was not the typical gender of either of the other two NPs, e.g. having the pronoun "she" after the NPs "umpire" and "baseball player." In other items, number performed the same function, so that the pronoun was plural after two singular NPs. In these cases, it was not possible for the plural pronoun to refer to the set consisting of the conjunction of the two NPs, because doing so would result in a c-command violation.

The short proper name condition provided a comparison that kept length similar to the 1st/2nd pronoun condition. Fourteen of the proper names were two or three letters long, five names were four letters long and one was five.

Each questionnaire was made up of 20 experimental items and 80 fillers. The fillers were similar in length to the experimental items and were also complex. The four conditions were counterbalanced across lists, so each subject saw one version of each item and five versions of each condition. The lists were pseudo-randomized so that no two experimental items occurred back to back and the order of the questionnaire pages was varied for each participant. A complete list of items is included in Appendix A.

Procedure

Participants were asked to rate the complexity of sentences on a scale of one to five, one being "easy to understand" and five being "hard to understand". The questionnaire began with a page of instructions asking participants to make their judgments based on their first impressions without reading sentences more than once. In the instructions, participants were given six practice items with a brief discussion of the sort of ratings each of the practice items might be assigned. The first two example sentences were relatively comprehensible, while the final four were more difficult to understand. None of the example sentences had the same doubly nested structure as the experimental items, but one of the difficult sentences was triply nested: "The man who the woman who the cat which the dog chased bit on the ankle met at the party talked to her yesterday in the late afternoon". Ratings in the 1 or 2 range were suggested for the easier sentences, and ratings in the 3, 4 or 5 range were suggested for the more difficult sentences, but participants were advised that individuals often differ on which sentences they find easier or harder to understand.

1.2.2 Results

The mean ratings for each condition are presented in graphical form in Figure 1.1.

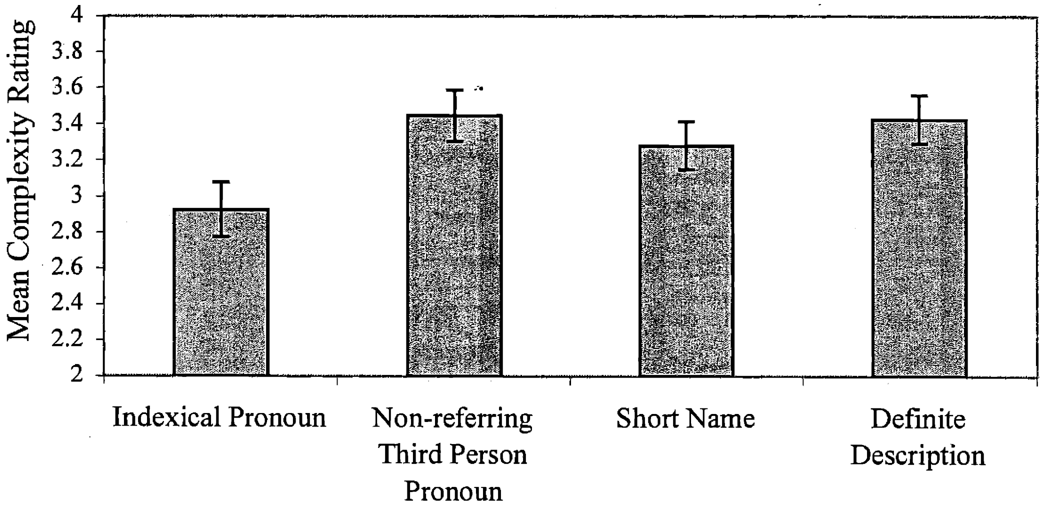

Figure 1.1 Mean complexity ratings for Experiment 1

When tested with repeated measures F-tests and adjusted for multiple tests using the Bonferroni correction, the 1st/2nd pronoun condition was significantly less complex than each other condition as follows: the 3rd condition, (F1(1,39)= 19.99, MSe= .27, p<.005; F2(1,19)= 17.38, MSe= .16, p<.01), the short name condition (F1(1,39)=18.01, MSe= .14, p<.005 ; F2(1,19)= 13.65, MSe= .09, p<.01) and the definite condition (F1(1,39)= 32.29, MSe= .15, p<.005; F2(1,19)= 33.22, MSe= .08, p<.005). There were no differences among the 3rd pronoun, name and definite NP conditions, Fs < 4, ps > .2 with Bonferroni corrections. Furthermore, the results were unchanged in an analysis over the 1st/2nd pronoun and short name conditions of items in which the length of the pronouns and names differed by no more than one letter. Thus word length is likely not the cause of the observed differences between the 1st/2nd pronoun condition and the other conditions.

1.2.3 Discussion

The results of Experiment 1 show that doubly nested sentences with 1st or 2nd person pronouns as their innermost subject are less complex than the same sentences with other NPs as their innermost subjects. Multiple hypotheses can account for this result. One such hypothesis is the DLT. According to the DLT, the distance of an integration is measured in terms of new discourse referents. The DLT predicts that the 1st/2nd pronoun condition will be less complex than the other conditions, because 1) the most embedded subject position interrupts multiple long distance dependencies and 2) the 1st/2nd condition does not require the building of a new discourse referent, while the other conditions do.

An alternative hypothesis that Experiment 1 did not rule out is that the 1st/2nd pronoun condition may have been more plausible than the other three conditions. The conditions differ slightly in meaning as a result of the different NPs in the innermost subject position, and it is possible that subjects found it more plausible for "I", "you" or "we" to perform the most embedded action than for any of the other NPs to perform it.

A third possibility is that differences in the accommodation (or bridging) that was required among the conditions may have contributed to the observed complexity differences. In discourse, new referents are usually introduced with indefinite NPs, while definite descriptions and pronouns are used to refer back to already established referents (Heim, 1982). First names are also usually used in situations where the identity of the referent is already known (Ariel, 1990). Using a name, definite description or pronoun to introduce a new referent initiates a referent access process that attempts to find a referent for the NP from the discourse model or memory. If no appropriate referent is found, then a new referent must be introduced and related to the rest of discourse (Haviland & Clark, 1974). This process is called accommodation or bridging. The 1st/2nd pronoun condition did not require accommodation or bridging, whereas the other conditions did. Thus, it is possible that the processing load associated with accommodating referents for names, 3rd pronouns and definite descriptions increased their complexity.

Experiment 2 addresses the above alternative explanations for the data from Experiment 1. It also investigates another formulation of the DLT's integration cost metric. In Gibson (1998), the integration cost metric relied on a binary distinction between new and old referents. It assumed that all referents can be categorized as either new or old and all new referents impose a cost while all old referents do not. An alternative hypothesis supported by the literature is that there is a continuum of accessibility or prominence in discourse. Evidence suggests that old referents which have appeared in prominent syntactic positions are more accessible than old referents which have appeared in less prominent positions (Arnold, 1998; Ariel, 1990; Gordon et al., 1993; Gundel et al., 1993). The effort required to introduce a new referent into discourse is also variable, and with strong contextual support sometimes new referents can be processed as quiçkly as old referents (Garrod & Sanford, 1982; Gernsbacher, 1991; Tanenhaus & Carlson, 1990). Experiment 2 begins to differentiate between these two versions of the DLT's integration cost metric: the binary version and a continuous version based on the accessibility of the intervening discourse referents.

1.3 Experiment 2

Like Experiment 1, Experiment 2 tested the complexity of doubly nested sentences with different types of NPs as the most embedded subject. In addition, Experiment 2 included right branching versions of the doubly nested sentences to control for plausibility differences between conditions. The right branching sentences had the same meaning and words as the doubly nested sentences, and so allowed the effects of syntactic structure to be isolated. A referent-based distance metric predicts that there will be differences among both the right branching and nested conditions, but that the differences will be greater among the nested conditions than among the right branching conditions, resulting in an interaction between structure (nested vs. right branching) and NP type. This prediction results from the greater number of long distance integrations in nested structures than in right branching structures. In Experiment 2, the types of the subject NPs were chosen from different levels of the Givenness Hierarchy (Gundel et al., 1993). The Givenness Hierarchy hypothesizes a link between the type of an NP and the degree to which its antecedent is accessible in discourse, and it provides a continuum along which to test the predictions of the binary and continuous distance metrics.

Numerous researchers have proposed categorizing the types of referring expressions by the cognitive or discourse status of their referents (e.g. Prince, 1981; Garrod & Sanford, 1982; Garrod et al., 1994; Ariel, 1988; Gundel et al., 1993). Gundel et al. define the cognitive status of a referent as the cognitive location of a representation for that referent. For example, a referent can be represented in the currently active piece of the discourse model, in an inactive part of the discourse model, in long term memory, or it can have no prior representation at all. Gundel et al. (1993) presented a cross-linguistic study categorizing all the NPs in a set of naturally occurring discourses according to the cognitive status of their referents. The results of the study indicated that the type of an NP was a reliable indicator of its cognitive status. Gundel et al. (1993) claimed that by using a particular linguistic expression speakers indicate the cognitive status of the referent they are referring to. A referent for an NP which is low on the hierarchy (peripheral[3]) is thus signaled to be newly introduced or found in long term memory, while a referent for an NP high on the hierarchy (central) must be highly activated in the current discourse model.

Gundel et al.'s (1993) Givenness Hierarchy:

Central   Peripheral
in focus < activated < familiar < uniquely identifiable < referential[4]
{*it*}   {*this, that*}   {*that* N}   {*the* N}   {*a* N}

Gundel et al.'s Givenness Hierarchy places different types of referring expressions and their cognitive statuses into a hierarchy where each more central status logically implies all of the more peripheral statuses. In situations without a high-level discourse shift, speakers usually use expressions from the most central appropriate levels of the hierarchy. This follows from Grice's (1975) Maxim of Quantity, because using a peripheral form when a more central form would have been appropriate can sometimes be misleading. For example, if a woman tells her sister a story about a common acquaintance and refers to the acquaintance as "a friend" rather than by using his/her name, she is breaking a conversational maxim. By using terms indicating more centrality, speakers imply that the conditions of all more peripheral levels of the hierarchy are satisfied (Gundel et al., 1993).

In certain situations, speakers use forms from the more peripheral end of the hierarchy to refer to a referent with a more central cognitive status. So, for example, in a conversation about a doctor, a speaker might refer to the doctor using a pronoun "she," but then in the next sentence use a full NP such as "the steady-handed surgeon," even though the doctor is already highly

[3] I am introducing the terms "central" and "peripheral" to refer to opposite ends of the Givenness hierarchy. I find these terms more descriptive than "high" and "low". The choice of "central" and "peripheral" is intended to refer to the subset relation between the cognitive statuses on the hierarchy and also to the current activation of the referent in discourse. Sometimes I will refer to particular types of NPs as being central or peripheral even though the terms are more natural for describing cognitive statuses than for describing types of NPs. When I say that an indefinite is peripheral I mean that indefinites usually introduce referents that are found towards the peripheral end of the scale.

[4] Gundel et al.'s (1993) Givenness Hierarchy includes one cognitive status that is not considered here. Gundel et al. make a distinction between referential and type-identifiable indefinites. Type-identifiable indefinites assert the existence of a referent of some type, but do not actually introduce a referent to discourse. I assume that during normal processing, indefinites usually instantiate referents. Hence, the most peripheral cognitive status that I consider is referential.

activated. Sometimes this is done in order to highlight a particular property of a referent. Other times it can indicate contrast or avoid ambiguity (Chafe, 1987). Speakers do not, however, use more central referential expressions to refer to referents with more peripheral cognitive statuses; e.g. a speaker will never use "it" to refer to a completely new referent. This holds true for adults, but it seems to be a late developing pragmatic constraint, as several researchers have shown that young children often use central expressions for referents which have not yet been mentioned in the current discourse (e.g. Maratsos, 1976).

The most peripheral cognitive status that will be considered is *referential,* which instantiates a specific yet possibly unfamiliar individual. For example, an indefinite NP such as "a student" is referential if the speaker has a particular student in mind in the sentence "A student cheated on the test." Definite NPs require their referents to be *uniquely identifiable*. This can be either because the unique referent is clear from the current discourse, or because the NP is modified in such a way as to restrict the set of possible referents in the world to one. In this second case, the referent of the definite NP does not necessarily have to be an entity that is old to the discourse, but could be a new highly specified entity. A *familiar* entity is one that the hearer can uniquely identify because it has been either been previously mentioned in the current discourse or because there is a representation of the entity already in long term memory. *Activated* referents are those that are currently in the discourse model and thus in working memory. 1st and 2nd person pronouns always have activated status, because the speech participants are part of the immediate linguistic context. Finally, the most restrictive cognitive status is the property of being *in focus*. Referents that are in focus are not only in working memory, but are the current center of attention. They are most likely to be realized syntactically as subjects and lexically as pronouns or zero anaphors (Brennan, 1995; Gundel et al., 1993).

In order to use the Givenness Hierarchy in a theory of linguistic complexity, one must establish a theory about the mechanics of referential processing and the way givenness affects referent access. Garrod & Sanford (1982) proposed a theory of referential processing asserting that there are two costs associated with processing NPs in discourse: 1) the cost of introducing a new discourse referent, and 2) the cost of accessing a referent from the discourse model. One natural way to relate givenness and referential access is to assume that givenness is an indicator of the activation level of a referent. This assumption is compatible with Gundel et al.'s hypothesis that givenness indicates the cognitive location of a representation for a referent, as long as referents in different cognitive locations are associated with different activations. In fact, Gundel et al. name one of the most accessible cognitive statuses "activated," indicating that they consider accessibility, activation and cognitive location to be closely linked. If the ease of accessing a representation for a referent is related to the representation's activation level and representations from more central levels are more activated than representations from more peripheral levels, then it should be easier to access a referent for an NP from the more central levels of the Givenness Hierarchy than one from the peripheral levels.

If accessing or introducing referents for NPs from the more peripheral levels of the Givenness Hierarchy requires more processing than accessing referents for NPs from more central levels, then, assuming a referent-based distance metric like that in the DLT, integrations which cross more peripheral NPs should be more complex than integrations which cross more central NPs. Experiment 2 tested this prediction by having participants judge the complexity of doubly nested sentences where the most embedded subject was either an indefinite description (referential), a definite description (uniquely identifiable), a first name (familiar), a famous individual (familiar), a 3rd person pronoun with a referent (activated) or a 1st or 2nd person pronoun (activated/ in focus). Gundel et al. (1993) did not include proper names in their hierarchy, but Ariel (1990, p. 40) argued that because first names are less rigid than full names, "they must refer to relatively highly accessible entities, ones which are more accessible than those referred to by definite descriptions or full names." Consequently we have labeled first names with Gundel at al.'s "familiar" category.

Non-nested conditions were also included with the doubly nested conditions in Experiment 2. Both the binary and continuous discourse-based distance metrics predict an interaction in complexity between the nested conditions and the non-nested conditions, because the subject NP being varied interrupts more integrations in the nested conditions than in the non-nested conditions. The binary version of this metric (new vs. old) predicts a complexity difference between the pronoun conditions and all other conditions, with the familiar and more peripheral levels being rated more complex than the activated and in focus levels. In contrast, the continuous distance metric predicts more gradations among the conditions. In particular, it predicts that the difficulty of introducing or finding a referent for an NP will be correlated with its position in the hierarchy. If this is the case, then as NPs are chosen from more central positions on the hierarchy, one should see a gradual decline in complexity because the referent representations have higher activations and are easier to access. This effect should be greatest for the nested conditions.

1.3.1 Predictions for individual conditions

Definite and Indefinite descriptions

The definite and indefinite conditions were designed to test: 1) the continuous distance metric theory and 2) the possibility that accommodation failure and its resultant infelicity caused the high complexity in the definite condition in Experiment 1. Because indefinite descriptions carry no presuppositions of uniqueness or familiarity, they require no accommodation or bridging in a null context. If accommodation failure was responsible for the complexity of the definite description condition in Experiment 1, then the indefinite condition should be less complex than the definite condition in Experiment 2. According to the continuous distance metric theory, building and accessing referent representations causes increased processing load. The indefinite and definite conditions test this theory because indefinites always introduce new discourse referents and do not require accessing a previously introduced referent. Definites, on the other hand, sometimes refer to relatively inactive referents and sometimes refer to new referents that are closely related to elements already in the discourse model. If definites sometimes refer to inactive referents, then upon reading a definite it should be necessary to attempt to access a referent for it from the inactive portions of the discourse model or from long term memory. Since the referent's representation will be inactive, if it is in the discourse model at all, the access process will be difficult. If the only resource-draining process is referent access, then definites should require more resources to process than indefinites, making the definite conditions more complex. If both construction and access require resources, then the complexities of the two conditions may be closer. It is debatable whether constructing referents for definites and indefinites requires the same amount of resources (for a suggestion that it does not, see Webber, 1979). But if it does, then the definite conditions should be more complex than the indefinite conditions because of the additional cost of attempting to access a representation for the referent in the definite condition.

It is also possible that the indefinite condition will be rated more complex than the definite condition. This might happen because all the other NPs in the experimental sentences are definite descriptions. The only indefinite in the indefinite condition is the subject of the most embedded clause. In this experiment the embedded clauses were RCs presented without commas, which makes them more likely to be interpreted as restrictive. Importantly, restrictive RCs usually contain background information, which is already known to the hearer/reader. Thus indefinite descriptions should be unnatural in restrictive RCs, because they introduce new entities. The indefinite condition might therefore be rated as more complex than the definite condition, because of this violation of backgrounding.

A second reason why the indefinite condition might be rated more complex than the definite condition is provided by Webber (1979). Webber hypothesizes that more information from predicates external to the NP may be needed to build a referent from an indefinite NP than from a definite NP. For example after processing the NP "the doctor who I spoke to today," the hearer can build a unique referent for the doctor. In contrast, information from the main predicate of the sentence may be necessary to instantiate a referent for an indefinite. For example, in "A doctor who I spoke to today prescribed rest," there could be multiple doctors fitting that description. The only way to uniquely identify the particular doctor in question is to label it as the doctor whom the speaker spoke to today, who prescribed rest and who was mentioned in the previous utterance. Webber's analysis therefore predicts that processing indefinites requires more resources than processing definites, with the consequence that the indefinite condition should be rated as more complex than the definite condition.

Famous names

The evidence gathered in Experiment 1 did not distinguish between the possibility that referents which are new to the local discourse increase complexity and the possibility that referents which are new to the knowledge base of the listener increase complexity. The famous name condition differentiates between these two possibilities, because in a null context, the individual referred to by the famous name is new to the local discourse, but is old to the listener's knowledge base.

If the binary distance metric is correct and if being new to the current discourse creates complexity, then the famous name condition should pattern with the definite and indefinite conditions. If it is the property of being new to the listener's knowledge base that causes complexity, then this condition should pattern with the pronoun conditions. Alternatively, if the continuous distance metric is correct, then this condition should be easier than the definite and indefinite conditions, but harder than the pronoun conditions. Accessing a name's referent will be more difficult than accessing a pronoun's referent because pronominal referents are more activated. Accessing a name's referent will be easier than attempting to access and then to build a definite's referent, because in this experiment the referent access process initiated by the definite necessarily failed and it was necessary to build a new representation for the referent.

First names

The binary distance metric predicts that the first name condition will pattern with the definite and indefinite conditions, because first names introduce new discourse entities in this experiment. If there is a cost for accommodation in this experiment, then the continuous metric theory also predicts that the first name condition should be as complex as the definite condition. In both the first name and definite conditions, a reference form that is usually reserved for accessible referents is being used to introduce a new referent and that should cause complexity. If, on the other hand, there is no cost for accommodation in this experiment, then the continuous metric will make a different prediction. According to accessibility theorists, first names are generally used to refer to more central referents than full names (Ariel, 1990). In the famous name condition, many of the names were full names. Therefore, if the continuous distance metric is correct and if there is no cost for accommodation in this experiment, the first name conditions should be judged easier than the famous name conditions.

1st and 2nd person pronouns and 3rd person pronouns

In the 3rd pronoun conditions, each sentence was prefixed with the phrase "according to *name*". The name introduced a referent for the pronoun that appeared later in the sentence and eliminated the infelicity in the 3rd pronoun condition from Experiment 1. The binary distance metric predicts that the 3rd pronoun condition will pattern with the 1st/2nd pronoun conditions, because in this experiment the 3rd pronoun refers to a previously introduced referent. The continuous distance metric predicts that the 3rd pronoun condition will be as complex and possibly more complex than the 1st/2nd pronoun condition. 1st/2nd pronouns have the special property of being in a class of deictics which always are part of the indexical frame and thus always have referents (Chafe, 1987). Not only do 1st and 2nd person pronouns always refer to individuals who are old to the discourse model, but they usually have a unique referent in that model. 3rd person pronouns are more likely to be able to refer to multiple referents in a discourse, and the computation of semantic or syntactic restrictions is often necessary to determine which individual is the intended referent. This extra processing may cause 3rd person pronouns to be more difficult to process than 1st and 2nd person pronouns. 1st and 2nd person pronouns may therefore be easier to access than all other types of NPs tested in this experiment.

Summary of predictions

To sum up, the binary version of the DLT's distance metric predicts that the two pronoun conditions- and possibly the famous name condition- will be easier than the other conditions. The continuous version of the distance metric predicts that the complexity of the sentences will correlate with the amount of processing required to interpret the referent of their most embedded subject NP. If the amount of processing required to interpret a referent is a combination of the difficulty involved in accessing and/or building a representation of that referent, then definites and indefinites should require the most processing. Indefinites require no access process, because they always introduce new referents, but they may be rated as more complex than definites in these items either because it is odd to introduce new discourse structures in restrictive RCs or because referents are more complicated to build from indefinites. Definites may not be as difficult to build, but they initiate a referent access process that attempts to locate a referent for the definite among the relatively inactive referents in the discourse model and memory. Famous names should require less processing, because their referents are locatable in memory. First names, as long as there is no penalty for accommodation, usually refer to even more accessible entities than full names, so the referent access processes that they initiate should quit after being unable to locate a suitable referent among the most activated referents in memory. As a result, the first name condition should be less complex than the famous name condition. Pronouns should require the least processing, because their referents are always "in focus" or "activated" and very accessible. Thus, according to the continuous discourse-accessibility-based theory, sentences with indefinite and definite NPs should be rated most complex, famous names and first names less complex, with the two pronoun conditions least complex. These effects should be stronger in the nested conditions than in the right branching conditions.

1.3.2 Method

Participants

Sixty members of the MIT community were paid participants in this study. Participants took approximately 20-25 minutes to complete the survey and were paid $4.00 to do so. Most participants also took part in an unrelated self-paced reading experiment during the same trip to the lab. The total amount of time the participants spent in the lab was approximately an hour.

Materials

Experiment 2 tested 36 items in a 2 x 6 design crossing structure (nested vs. right branching) with subject NP type (indefinite description, definite description, famous name, first name, 3rd person pronoun or 1st or 2nd person pronoun). In this experiment, 3rd person pronouns were tested in a more natural way than they were in Experiment 1. The phrase "according to *name*" was added to the beginning of the test sentences in the 3rd pronoun conditions, so that the name could serve as the antecedent of the subsequent pronoun. A sample item from the experiment is included below:

Nested
5a. The salesman who the woman who {I, the company, a company, Microsoft, Bob} hired dealt with was very polite.

Right branching
5b. {I, the company, a company, Microsoft, Bob } hired the woman who dealt with the salesman who was very polite.

Nested, 3rd pronoun
5c. According to Bob, the salesman who the woman who he hired dealt with was very polite.

Right branching, 3rd pronoun
5d. According to Bob, he hired the woman who dealt with the salesman who was very polite.

The questionnaire in Experiment 2 was made up of 36 target items, 30 experimental items from an unrelated experiment and 54 additional fillers. The 12 conditions of each target item were counterbalanced across lists. Participants read only one condition per item and three versions of each condition across the questionnaire. The lists were pseudo-randomized so that at least one filler intervened between each pair of experimental items and the order of the questionnaire pages was varied for each participant. A complete list of items is included in Appendix B.

Procedure

The procedure in Experiment 2 was the same as in Experiment 1.

1.3.3 Results

The means for all conditions are presented in Figure 1.2.

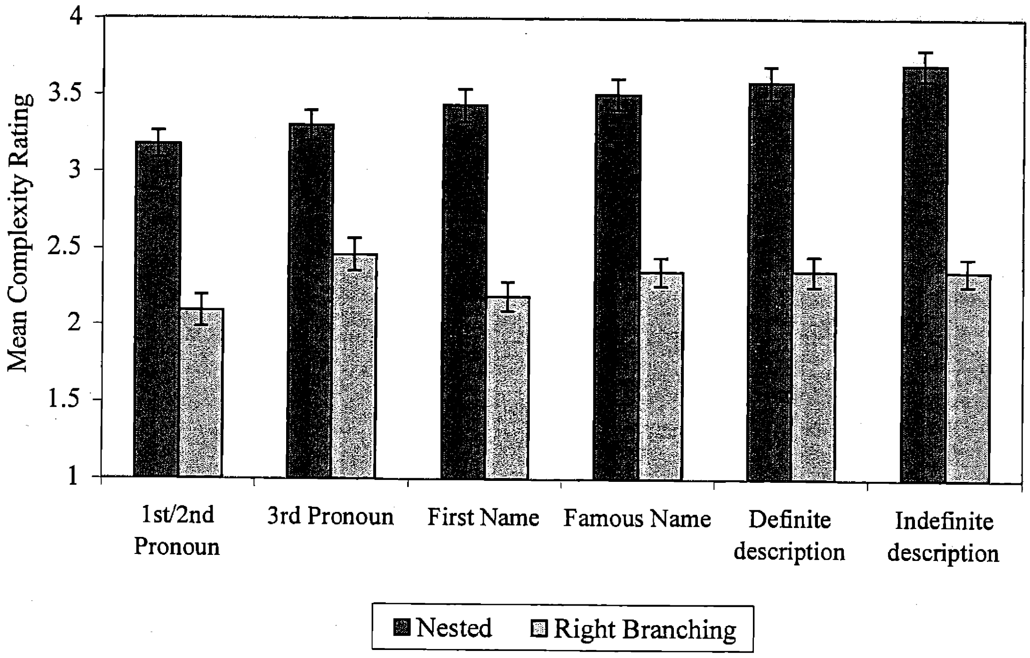

Figure 1.2. Mean complexity ratings for Experiment 2

A 2 x 6 repeated measures Analysis of Variance comparison showed a main effect of structure type, with the right branching structures rated significantly easier than the nested structures (F1(1,59)= 264, MSe= .91, p<.001; F2(1,35)= 333, MSe= .43, p<.001). There was also a main effect of NP type (F1(5,295)= 6.61, MSe= .33, p<.001; F2(5,175)= 5.44, MSe= .24, p<.001), and an interaction between structure and NP type (F1(5,295)= 3.29, MSe= .30, p<.01; F2(5,175)= 2.48, MSe= .24, p<.05). In order to discover the reason for this interaction, post hoc contrast tests were carried out between the 1st/2nd pronoun condition and each of the rest of the conditions. The reason for comparing each condition to the 1st/2nd pronoun condition was that Experiment 1 had shown that the 1st/2nd pronoun condition was the only condition that differed from the other conditions. None of these individual contrasts revealed any significant interactions.

In another probe of this interaction between structure and NP type, additional repeated measures ANOVAs were performed. In an ANOVA across just the embedded sentences, looking for effects of discourse type and using a Bonferroni correction for multiple tests, there was a significant effect both by participants (F1(5,295)=6.12, MSe= .35, p<.005) and by items (F2(5,175) =6.31, MSe= .21, p<.005). The same ANOVA performed on the right branching conditions showed a significant effect by participants (F1(5,295)= 3.68, MSe= .28, p<.05), but the effect did not reach significance in the items analysis (F2(5,175)= 2.24, MSe= .28, p>.1) and so may be an artifact. These two ANOVAs suggest that the overall interaction may be due to the variation in the embedded sentences' complexities and a lack of variation in the right branching sentences' complexities.

The last tests performed were repeated measures contrasts, to test whether there was a significant relationship between the NP type's status on the Givenness Hierarchy and the complexity ratings. A polynomial contrast test indicated that a monotonically increasing trend accounted for a highly significant amount of the variance in the nested conditions ($t(59)= 5.67$, $p<.001$, $r_{contrast} =.59$; $t(35)= 6.77$, $p<.001$, $r_{contrast} = .75$ with Bonferroni corrections), but did not account for a significant amount of the variance in the right branching conditions ($t(59)= 1.55$, $p >.2$, $r_{contrast} = .20$, $t(35)= 1.49$, $p >.2$, $r_{contrast} = .24$ with Bonferroni corrections). Thus, in the nested sentences, as the NP type of the innermost subject became more peripheral, complexity increased. There was no corresponding relationship between an NP's status in the hierarchy and complexity ratings in the right branching sentences. A repeated measures contrast test encompassing all of the conditions in the experiment, assigning low, even weights to the right branching conditions and higher, increasing weights to the nested conditions was also done. This test was important because it directly compared the data to the pattern predicted by the DLT with a continuous distance metric. The predicted pattern accounted for a significant amount of the variance in the data, ($t(59)= 16.53$, $p<.001$, $r_{contrast} = .91$; $t(35)=18.38$, $p<.001$, $r_{contrast} = .95$ with Bonferroni corrections).

1.3.4 Discussion

As predicted by both versions of the DLT, the effects of changing the type of the innermost NP of the nested sentences (the outermost NP of the right branching sentences) were stronger in the nested conditions than in the right branching conditions. The results of Experiment 2 support a continuous version of the DLT's referent-based integration cost over a binary version. The data did not show two distinct clusters of conditions, as predicted by a binary distance metric, but instead suggested that complexity is sensitive to gradations of status between new and old. The complexity pattern in the nested conditions matched the predictions of the Givenness Hierarchy almost exactly.

Interestingly, the indefinite conditions were numerically, if non-significantly, more difficult than the definite conditions. This complexity pattern is consistent with the hypothesis that indefinites are slightly peculiar in restrictive RCs. This pattern is also consistent with Webber's (1979) hypothesis that building a referent for an indefinite is more difficult than building a referent for a definite. The pattern of results between the name, definite and indefinite conditions suggests that there was no significant effect of accommodation failure in this task. In particular, indefinites, which require no accommodation, were numerically the most complex condition. This lack of accommodation failure effects could be a result of the experimental task. Accommodation relies on factors such as relevance and communicative intentions, which play a role in normal referential and discourse processing (Sperber & Wilson, 1986), but were not important in this experiment where each sentence defined a separate discourse. In a reading task with many sentences, no contexts and many definite descriptions, it is not surprising to find no effects of accommodation.

The data from Experiment 2 support a referential processing theory based on Gundel et al.'s (1993) Givenness Hierarchy. The increasing monotonic relationship between complexity and the varied NP's position on the hierarchy indicated that each NP initiated a referent access process and that the amount of resources this process consumed was related to the level of activation associated with the NP's position in the hierarchy. Data from the first name and definite conditions support the hypothesis that the Givenness Hierarchy is an indicator of referent activation and that NPs from the more central end of the hierarchy refer to only very active referents while NPs from the more peripheral end of the hierarchy can refer to either active or inactive referents. In the first name and definite conditions, all attempts at referent access should have failed because there were no referents for the NPs in question. Since names must refer to entities that are highly activated, while definites can refer to entities that are highly activated or have lesser activation, the access process for names was expected to be easier than for definites even when no referent is returned. The fact that the definite NP conditions were more difficult to process than the first name conditions is consistent with the theory that attempting to access a referent for a definite consumed more resources because the access process for the definite had to consider referents that were less active.

1.4 Experiment 3

Consistent with the continuous integration metric version of Gibson's (1998) DLT, the results of Experiments 1 and 2 suggest that nested structures are more complex when the discourse structure for the NP in the most embedded position is harder to build or access. But these results can also be accounted for by Bever's (1970, 1974) complexity hypothesis. The relevant prediction of Bever's hypothesis is that nested sentences with similar subject NPs are more difficult to process than ones with dissimilar subject NPs. Thus, doubly nested sentences with a pronoun in one of the subject positions are predicted to be less complex than doubly nested sentences with three definite descriptions, because pronouns are sufficiently different from definite descriptions as to make the series of three NPs dissimilar. It is possible that manipulating the innermost subject NP in Experiments 1 and 2 changed the degree of similarity between the three NPs. If this was the case, and if the similarity of the NP-types to definite descriptions varied in the same way as the NP-status given in the Givenness Hierarchy, then Bever's hypothesis might account for the results of these experiments.

Experiment 3 was designed to help distinguish the two theories. A prediction made by the DLT, but not by Bever's subject-similarity hypothesis, is that nested structures with a low-discourse-cost NP (such as a pronoun) in the most embedded subject position should be less complex than the same structure with the same NP in one of the other two subject positions. Experiment 3 tested this prediction. The materials for Experiment 3 consisted of doubly nested sentences whose subject NPs and verbs were the same across conditions, with one quantified NP (usually a quantified pronoun such as "everyone")[5] and two definite NPs rotated through the three subject positions in the sentence, as shown in the examples in (6):

6a. Doubly nested, Outer
Everyone who the journalist who the photographer met liked was at the party.

6b. Doubly nested, Middle
The photographer who everyone who the journalist met liked was at the party.

6c. Doubly nested, Inner
The journalist who the photographer who everyone met liked was at the party.

The low-discourse-cost NP in these materials was a quantified pronoun in _ of the items: "everyone", "everybody", or "no one". In the remaining _ of the items, the low-discourse-cost NP was "many people". All of the other subject positions in the sentences contained definite descriptions referring to people. The low-discourse-cost NPs were low cost because the noun and pronouns that were quantified indicated undifferentiated sets of humans. Unlike the definite descriptions in this experiment, which require searching for and building a referent, the low-discourse-cost NPs only require a check of the discourse to verify that humans are relevant.

In (6a) the pronoun is in the matrix subject position, in (6b) the pronoun is in the middle subject position, and in (6c) the pronoun is in the most embedded subject position. Each condition has the same three subject NPs, so similarity among the NPs is constant. Thus, Bever's

(1970,1974) similarity-based complexity hypothesis predicts no differences among the three conditions. Gibson's (1998) theory, on the other hand, predicts a difference because: 1) pronouns and definite NPs differ in their referential status, and 2) the three subject positions in doubly nested sentences interrupt different numbers of head-dependent relationships. In particular, the outer subject NP interrupts no dependencies, the middle subject NP interrupts two (the matrix subject-verb dependency and the outer relative pronoun-gap dependency), and the inner subject interrupts four (the two that the middle subject interrupts as well as the corresponding two from the middle clause). According to a discourse-referent-based distance metric, the condition with the quantified pronoun as the outermost subject (6a) should be most complex, because definite descriptions occupy the inner two subject positions, increasing the distance between heads and dependents. The condition with the pronoun as the middle subject (6b) should be less complex, and the condition with the pronoun as the innermost subject (6c) should be least complex.

1.4.1 Method

Participants

Sixty members of the MIT community were paid participants in this study. Participants took approximately 20-25 minutes to complete the survey and were paid $5.00 to do so. About half the participants also participated in an unrelated self-paced reading study during the same testing session, and spent approximately one hour in the lab. The other half of the participants participated only in this study.

Materials

Experiment 3 tested 18 items in a 2 x 3 design crossing structure (doubly nested vs. singly nested) with low-discourse-cost NP position (outer, middle or inner subject position). The singly nested conditions were included in order to provide a plausibility control for the doubly nested conditions. In the conditions with the low-discourse-cost NP in the innermost and outermost subject positions, singly nested versions were created by extracting the most embedded clause and conjoining it with the singly nested sentence that remained. In the conditions with the low-discourse-cost NP in the middle subject position, the outermost clause was extracted and conjoined to the remaining singly nested sentence. It was not possible to use fully right

[5] Quantified NPs (usually pronouns) were used in the materials for Experiment 3 rather than referring pronouns like "you" and "she", as in the materials in Experiments 1 and 2, because it is not generally

branching versions of the nested sentences, because the nested and right branching sentences had different meanings. To illustrate: the nested sentence "everyone who the journalist liked was at the party" does not have the same meaning as its right branching variant "the journalist liked everyone who was at the party." The clause with the low-discourse-cost NP remained nested in the singly nested conditions, so that meaning was as similar to the doubly nested versions as possible. The doubly nested versions of an item were given in (6) above. The singly nested versions of this item are included below:

Singly-nested, Outer
7a. The photographer met the journalist, and everyone who the journalist liked was at the party.

Singly-nested, Middle
7b. The photographer was at the party, and everyone who the journalist met liked the photographer.

Singly-nested, Inner
7c. Everyone met the photographer, and the journalist who the photographer liked was at the party.

The questionnaire in Experiment 3 was made up of 18 target items, 20 experimental items from an unrelated experiment and 62 additional fillers. The six target conditions were counterbalanced across lists. Participants saw one condition from each item and three items for each condition across the questionnaire. The lists were pseudo-randomized so that experimental sentences were always separated by at least one filler and the order of the questionnaire pages was varied for each participant. The complete set of items is included in Appendix C.

Procedure

The procedure in Experiment 3 was the same as in Experiments 1 and 2.

1.4.2 Results

The means for each condition are shown in Figure 1.3.

possible to modify referring pronouns with a relative clause, as is necessary in the manipulation here.

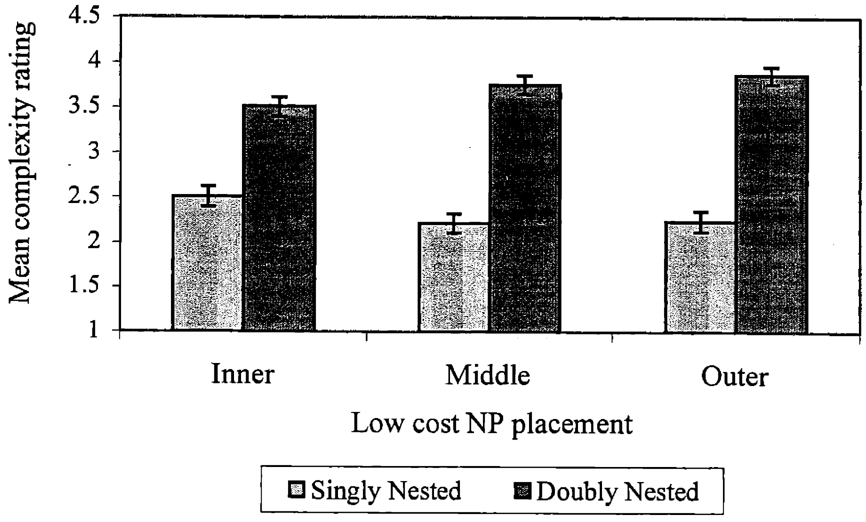

Figure 1.3. Mean complexity ratings for Experiment 3

A repeated measures 2 x 3 ANOVA revealed a significant interaction between the factors of structure and quantifier placement (F1(2,118)=11.32, MSe= .30, p<.001; F2(2,34)=10.82, MSe= .09, p <.001). There was also a significant main effect of structure, such that the doubly nested conditions were rated as more complex than the singly nested conditions (F1(1,59)=162.06, MSe= 1.07, p<.001; F2(1,17)= 516.78, MSe= .10, p<.001). No effect of quantifier position was apparent at this level (Fs < 1). A repeated measures ANOVA across the three doubly nested conditions showed a significant difference among the conditions (F1(2,118)= 7.69, MSe= .25, p<.001; F2(2,34) = 4.88, MSe= .12, p < .05). Planned individual comparisons revealed that the innermost subject condition was significantly easier to process than either the outermost subject condition (F1(1,59) = 13.7, MSe= .27, p<.001; F2(1,17)= 19.08, MSe= .06, p<.001) or the middle subject condition, although this effect was marginal in the items analysis (F1(1,59) = 6.86, MSe= .26, p <.05; F2(1,17)=3.53, MSe= .15, p =.08). Finally, the middle subject condition was numerically easier than the outermost subject condition, but this difference did not reach significance (Fs < 1.5, ps>.2).

1.4.3 Discussion

The results of Experiment 3 support the DLT's hypothesis that increased discourse processing costs makes integrations which cross new discourse structures more difficult. The overall pattern of the results is as predicted by the DLT: the condition with the low-discourse-cost

NP in the most embedded subject position was least complex, the condition with the low-discourse-cost NP in the middle subject position was more complex and the condition with the low-discourse-cost NP in the outer subject position was numerically the most complex. Furthermore, the results from the singly nested conditions demonstrate that plausibility differences are not the cause of the complexity difference among the doubly nested conditions, because the singly nested conditions had similar meanings and showed a different complexity pattern. Finally, these results are not predicted by Bever's (1970) similarity-based complexity account.

1.5 Experiment 4

The data from the first three experiments were gathered using complexity judgment questionnaires. Experiments 4 and 5 were designed to test similar but simpler structures using a self-paced word-by-word reading paradigm. According to the DLT, integration cost is one factor contributing to reading times at a word (Gibson, 1998, 2000; Gibson & Ko, 1998; Grodner, Watson & Gibson, 2000). Therefore, the DLT predicts that nesting complexity in English will usually be manifested in reading time differences at verbs, the points of long-distance integrations in nested structures.

Experiment 4 evaluated predictions made by both the binary and continuous versions of the discourse accessibility-based distance metric, without attempting to distinguish between the theories. The items in this experiments consisted of relative clause (RC) and complement clause (CC) structures, where the subject of the embedded clause was either a 1st or 2nd person pronoun or a definite description as in (8):

8a. RC
The woman who {you/ the boy} had accidentally pushed off the sidewalk got upset and decided to report the incident to the policeman standing nearby.

8b. CC
The woman knew that {you/ the boy} had accidentally pushed the girl but gave him/you a long lecture anyway.

Consider the DLT's predictions with respect to processing the relevant verbs in the sentences in (8). At the embedded verb, "pushed," in the CC conditions, 1 unit of integration cost is consumed because "pushed" is a new discourse referent and integrates to the most recently occurring referent, the subject NP "you" or "the boy". The same subject-verb integration occurs in the RC conditions and also has one unit of cost. There is an additional unit of integration cost in the RC pronoun condition, because there is an additional integration between the relative pronoun and the object position of the embedded verb that crosses the new referent indicated by the verb "pushed." The referent, "you," also intervenes in this integration, but it causes no or little additional cost because it is old to the discourse. There are two additional units of integration cost for the relative pronoun-gap integration in the RC full NP condition, because two new referents intervene between the endpoints of this integration: the new referent indicated by the verb "pushed", and the new referent indicated by the NP "the boy". Hence the DLT predicts that the RC conditions should be slower than the CC conditions at the verb "pushed", with the RC full NP condition slower than the RC pronoun condition and no difference between the CC conditions.

At the main verb in the RC conditions, ("got upset" in (8)), the DLT predicts an effect of NP-type, because the main verb must be integrated with its subject NP ("the woman" in (8)). This integration crosses the intervening RC, which has either a new referent (full NP) or old referent (pronoun) subject NP. Thus the DLT predicts that the RC full NP condition will be slower than the RC pronoun condition at the main verb of the sentence. Furthermore, the DLT predicts that this difference will be larger than any differences observed on either the preceding region (the auxiliary and adverb, "had accidentally" in (8)) or on the following NP, because integrations at these regions do not cross the embedded subject NP.

1.5.1 Comparing the predictions of the DLT with a purely discourse-based theory

In addition to testing these reading time predictions, Experiment 4 addresses the broader issue of whether the DLT measures the costs of constructing syntactic representations. As presented in Gibson (1998), DLT integration cost measures the difficulty of incorporating a new word into a syntactic structure. But it is possible that a resource-usage theory operating over only discourse models or representations such as those used in discourse representation theory (Kamp, 1981) may make many of the same predictions as the DLT. This is because the endpoints of long-distance dependencies where the DLT predicts high integration costs are also often points at which discourse models are updated. For example, when the verb "pushed" is encountered in (8a) syntactic processes must establish subject-verb and filler-gap dependencies, but discourse processes also must expand the discourse model to include new roles and relations introduced by "pushed" and fill as many of them as possible with referents. In order to discuss the predictions made by a purely discourse-based theory, a more elaborated theory of discourse processing is necessary. We will sketch one here and compare the predictions for Experiment 4 made by the DLT to those made by a purely discourse-based resource theory.

A simple activation-based discourse processing theory would predict that discourse referents are activated when they are introduced or accessed, and that activation decays as a result of processing other referents. Referents start with varying levels of activation depending on how they are introduced (see e.g. Arnold, 1998; Sanford et al., 1988), but as processing resources are diverted to activate other referents, their activation lessens. This activation decay is partially dependent on the amount of resources required to activate the more recent referents. There will be less decay following a referent that is easily accessible than after a referent requiring more processing. When a referent is re-accessed, its activation level increases to some threshold. According to such a theory, processing costs would be high when a referent with low activation needed to be re-accessed.

(same as above)

8a. RC
The woman who {you/ the boy} had accidentally pushed off the sidewalk got upset and decided to report the incident to the policeman standing nearby.

8b. CC
The woman knew that {you/ the boy} had accidentally pushed the girl but gave him/you a long lecture anyway.

This purely discourse-based resource theory makes the same predictions as the DLT at the embedded verb "pushed" in RC and CC sentences like (8a) and (8b). At "pushed," the referent of "boy" or "you" must be reactivated in order to assign it the discourse role of pusher. Since only *pushed*[6] has been activated since *boy* or *you*, they should be relatively easy to re-access. This is all the processing required at this verb in the CC conditions. There is an additional cost in the RC conditions, because the referent of "the woman" must also be re-accessed to be assigned the discourse role of pushee. In the RC pronoun condition, *you* and *pushed* have both been activated since *the woman*, but *you* is very accessible, so *the woman* should only be slightly harder to access in this condition than in the CC conditions. In the RC full NP condition, *the boy* and *pushed* have both been activated since *the woman*, and they boy is not easily accessible, making it relatively harder to re-access *the woman*. Thus, like the DLT, the purely discourse-based resource theory predicts that the CC conditions will be fastest, the RC pronoun condition slower and the RC full NP condition slowest on the embedded verb.

The predictions of the purely discourse-based theory and the DLT diverge at the main verb region “got upset.” The pure discourse theory predicts no effect of NP-type at the main verb region in the RC conditions. At the main verb, *the woman* must be re-accessed in order to fill the experiencer role introduced by “got upset”. *The woman* was most recently accessed and activated at the verb “pushed.” *The woman* should therefore have the same activation in both conditions at the point of processing the main verb, because the same words - the embedded PP “off the sidewalk”- have been processed in both conditions since *the woman* was last accessed at “pushed”. Hence, unlike the DLT, the purely discourse-based theory predicts no difference in complexity between the RC conditions at the main verb region “got upset.”

To sum up, the DLT and the purely discourse-based resource theory presented above both predict that the RC conditions should be slower than the CC conditions at the embedded verb "pushed." They also predict an interaction between structure and NP-type in this region, such that the RC full NP condition should be slowest. At the main verb in the RC conditions, the DLT predicts that the full NP condition will be slower than the pronoun condition at the main verb, while the purely discourse-based resource theory predicts no difference at this location.

Finally, the DLT also makes predictions with respect to reading times corresponding to differences in syntactic storage across the RC and CC structures. As observed by Grodner, Gibson and Tunstall (2001), reading times are longer over portions of sentences that require the storage of more syntactic predictions. In this experiment, the RC conditions require storing the predictions of a verb and a gap site for the relative pronoun over the embedded clause. There is no similar storage necessary in the CC conditions. Thus the DLT predicts that reading times during the RC will be slower than reading times during the CC.

1.5.2 Method

Participants

80 members of the MIT community participated in this study. Participants took approximately 20-25 minutes to complete the experiment. All of the participants also participated in an unrelated self-paced reading study during the same testing session, and spent approximately one hour in the lab, with a short break between the two sessions of self-paced reading. Participants were paid $8.

Materials

[6] I indicate lexical items with quotation marks, and referents by italicization.

Four versions of 20 items were constructed in a 2 x 2 design crossing structure (RC, CC) and NP type (1^{st} /2^{nd} person pronoun, definite description) as exemplified in (8) above. The CC conditions were formed from the RC conditions by replacing the relative pronoun with a clause-taking verb and the complementizer "that" and adding a direct object after the embedded verb. After the embedded verb, the RC and CC conditions diverged. The 20 target sentences were combined with 40 sentences from two unrelated experiments and 30 filler sentences to form four lists. The four target conditions were counterbalanced across lists and the lists were randomized.

In all conditions, the embedded verb was preceded by the auxiliary "had" and an adverb. This auxiliary region served as a separator, in order to eliminate the possibility that reading times in the embedded verb region were affected by spillover from lexical differences at the embedded subject NP. In eight of the items, the embedded verb was a psychological state verb, while in the other twelve it described an action. The RCs in the RC conditions were always introduced by "who" and always had animate subjects that were either individuals or sets of humans. The embedded full definite NP subjects all had the form "the *noun,*" where the noun was one word, except in one item in which the NP was "the girl scout." A prepositional phrase (PP) modifying the first verb was inserted between the two verbs in the RC condition in order to separate the two verbal regions, where the DLT predicts the longest reading times. As a result, long reading times in the main verb region, "got upset," could not be due to spillover from the embedded verb "pushed." In 13 of the items, the PP was three words long; in five of the items the PP was four words long; in the remaining two items the PP was five and six words respectively. The main verb region in the RC conditions was always two words. In most of the items this region consisted of an auxiliary and a predicate adjective, but in some items it was an auxiliary and progressive verb form. The complete set of items is given in Appendix D.

Procedure

Participants performed self-paced reading in a word-by-word moving window display (Just, Carpenter & Wooley, 1982) on a Macintosh computer running software developed in our lab. At the start of each trial, a sentence appeared on the screen with all characters replaced by dashes. Participants pressed the space bar to change a string of dashes into a word. Each time the space bar was pressed, the next word appeared and the previous word reverted back into dashes. Time between bar-presses was recorded, as a measure of how long participants spent reading each word. After finishing each sentence, participants were required to answer a yes-no comprehension question about the sentence. They answered by pressing one of two keys, and if they answered incorrectly, the word "INCORRECT" flashed on the screen. No feedback was given for correct responses. Participants were asked to read at as naturally as possible and to take incorrect answers as an indication to read more carefully.

Up to 100 characters could appear on each line of the display. Each item spanned from one to one and one-half lines. The embedded verb regions and the main verb in the RC conditions always appeared on the first line. Participants were given a small set of practice items and questions before the experiment in order to familiarize them with the task.

1.5.3 Results

One participant's data was excluded because he fell asleep during testing. Three additional participants' data were excluded from analysis due to accuracy rates of 70% or lower on the comprehension questions pertaining to the experimental sentences. All other participants had accuracy rates of at least 75%. Table 1.1 presents the percentage of questions answered correctly for each condition.

RC full NP	90.4
RC pronoun	97.5
CC full NP	88.2
CC pronoun	89.1

Table 1.1 Percentage of comprehension questions answered correctly for each condition in Experiment 4

There was a significant main effect of NP type, such that questions about the pronoun conditions were more often answered correctly than questions about the full NP conditions (F1(1,75)= 7.23, MSe= .02, p<.01; F2(1,19)= 5.68, MSe= .01, p<.05). A main effect of structure, such that questions about the RC conditions were answered correctly more often than questions about the CC conditions, was significant in the participants analysis (F1(1,75)= 15.05, MSe= .01, p<.001) but not in the items analysis (F2(1,19)= 2.8, MSe= .02, p=.11). This effect of structure is opposite to what would be expected based on complexity, but may reflect differences between questions, since different conditions of the same item often had different comprehension questions. An interaction between structure and NP type was significant by participants, (F1(1,75)= 4.48, MSe= .02, p<.05) but not by items (F2(1,19) = 2.48, MSe= .01, p=.13). Words were grouped into regions for the purpose of analysis as in Table 1.2.

	1	2	3	4	5	6	7
RC:	The woman who	{you/the boy}	had accidentally	pushed off	the sidewalk	got upset	...
CC:	The woman knew that	{you/the boy}	had accidentally	pushed the	girl, but...		

Table 1.2 Regions of analysis for Experiment 4

Region 1 consisted of the initial subject, as well as the relative pronoun "who" in the RC conditions and the matrix verb and the following complementizer for the CC conditions (e.g., "knew that" in (11)). Region 2 contained the embedded subject NP, which varied according to NP type; it was a 1st or 2nd person pronoun or a definite description. Region 3 contained an auxiliary verb and an adverb. Region 4 contained the embedded verb followed by a preposition in the RC conditions. This region contained the embedded verb followed by the determiner "the" in the CC conditions. The word following the verb was grouped with the verb for the purposes of analysis because reaction times from points of high complexity – the verb in the RC conditions as predicted by the DLT - often spill over into following words in self-paced reading. This additional word was a high frequency function word in both the RC and CC conditions. The DLT makes no predictions beyond this region for the CC conditions, so region 5 contained the rest of the sentence for the CC conditions. For the RC conditions, region 5 contained an NP: the object of the preceding preposition. Region 6 was a two-word main verb region in the RC conditions, and region 7 contained the rest of the RC sentence.

Reading times were analyzed for sentences for which participants correctly answered the comprehension question. In addition, reading times were trimmed at 5 standard deviations from the mean for each word position in each condition. This trimming eliminated less than one percent of the remaining data. Unadjusted and adjusted reading time means for each region are presented in Table 1.3.

	The woman {who/knew that}	{you / the boy}	had accidentally	pushed {off/the}	the sidewalk/ *rest of CC*	got upset	...
RC full NP:	380	407	419	507	367	394	361
	(-26)	(10)	(-3)	(112)	(-9)	(-5)	(-45)
RC pronoun:	379	396	371	419	358	367	355
	(-19)	(31)	(-48)	(7)	(-46)	(-39)	(-54)
CC full NP:	383	380	384	405	382		
	(-27)	(-31)	(-39)	(6)	(-17)		
CC pronoun:	380	372	370	388	370		
	(-31)	(6)	(-57)	(-24)	(-33)		

Table 1.3 Mean reading times expressed in milliseconds per word (length adjusted times are in parentheses) for each region in Experiment 4

Repeated measures ANOVAs were performed on regions for which the DLT made predictions. Statistical tests were performed on both raw reading times and reading times that were adjusted for word length following the procedure proposed by Ferreira & Clifton (1986) and Trueswell, Tanenhaus & Garnsey (1994). We report only the statistical tests for raw times; the results were similar for the statistical tests that were performed on length-adjusted times. Mean unadjusted reading times for each region in Experiment 4 are graphed in Figure 1.4.

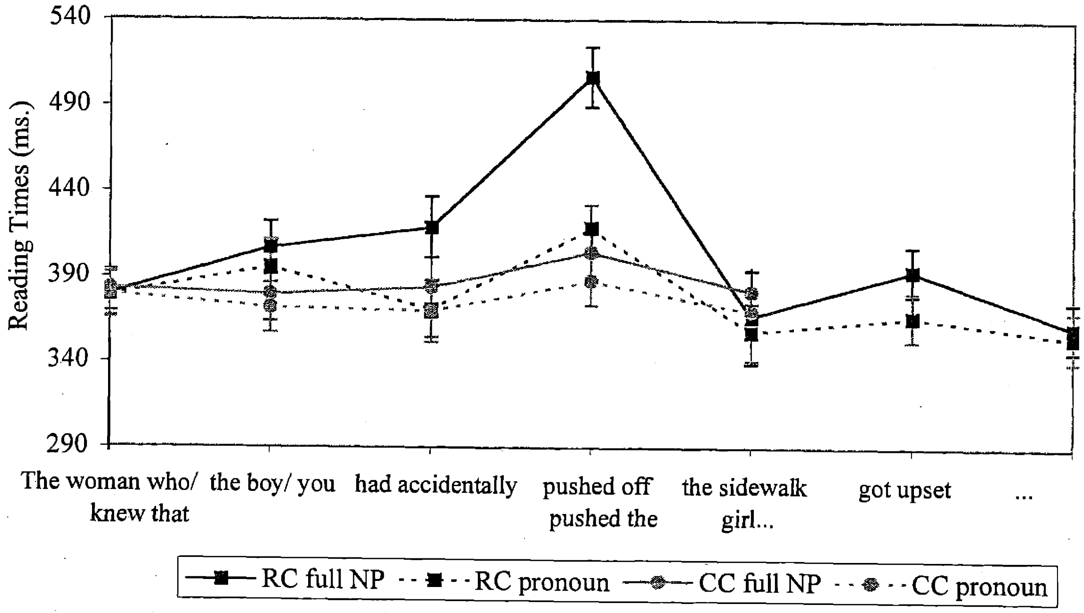

Figure 1.4. Reading times for Experiment 4

In region 1, the RC and CC conditions have different numbers of words and different structures, so no analyses were performed in these regions. In region 2, the embedded subject NP (pronoun or full NP), there was a main effect of structure, such that the RC conditions were read more slowly than the CC conditions (F1(1,75) = 7.66, MSe= 6555, p<.01; F2(1,19)= 6.9, MSe= 2151, p<.05), as predicted by the syntactic storage component of the DLT. NP-type did not affect reading times in this region and there was no interaction (Fs < 1).

In region 3, the auxiliary and adverb, there was a main effect of NP-type (F1(1,75)= 26.15, MSe= 2791, p<.001; F2(1,19)= 11.13, MSe= 2683, p<.005) such that the pronoun conditions were read faster than the full NP conditions. This may have reflected spillover from the previous region, where the NPs were introduced. There was also a main effect of structure similar to that in region 2, such that the CCs were read faster than the RCs (F1(1,75)= 6.84, MSe= 3580, p<.05; F2(1,19)= 3.05, MSe= 1265, p=.10). There was also an interaction between structure and NP-type (F1(1,75)=5.44, MSe= 4192, p<.05; F2(1,19)= 6.49, MSe= 963, p<.05), which was due to the RC full NP condition being slower than the other three conditions.

In region 4, the region containing the embedded verb and the following word, there was a significant main effect of NP type, such that the pronoun conditions were faster than the full NP conditions (F1(1,75)= 23.43, MSe= 8894, p<.001; F2(1,19)=9.95, MSe= 5805, p=.005). There was also a main effect of structure (F1(1,75)= 29.69, MSe= 11418, p<.001; F2(1,19)= 13.43, MSe= 5285, p<.005) with the CC conditions faster than the RC conditions. Finally, there was an interaction between the factors of structure and NP type (F1(1,75)= 11.7, MSe= 8385, p=.001; F2(1,19)=13.7, MSe= 2715, p <.005). These results are as predicted by the DLT. The two CC conditions are fastest and do not differ from each other (F1(1,75)= 1.63, MSe = 6309, p=.21; F2(1,19)= .62, MSe= 1824, p=.44). The RC conditions are both slower and they do differ, with the full NP condition significantly slower than the pronoun condition (F1(1,75)=27.00, MSe= 10970, p<.001; F2(1,19)= 14.00, MSe= 6695, p=.001). The one DLT prediction not fully supported by the data is that the RC pronoun condition should have slower reading times at this region than the CC conditions. Numerically the means are in the right order, but the difference between the RC pronoun and the CC pronoun conditions are only significant by participants and not by items (F1(1,75)= 7.89, MSe= 4587, p=.006; F2(1,19)= .85, MSe= 3197, p=.37).

Because the RC and CC structures are not comparable in region 5 or after, all further analyses were conducted over only the RC conditions. There were no differences between the RC conditions in region 5, the region consisting of the NP object of the preposition following the embedded verb, as predicted by the DLT (Fs < 1.6, ps > .20). Finally, in region 6, the main verbal region, the full NP condition was read more slowly than the pronoun condition (F1(1,75)= 6.58, MSe= 4141, p<.05; F2(1,19)= 4.44, MSe= 2733, p<.05), as predicted by the DLT, but not by the purely discourse-based theory.

1.5.4 Discussion

The results of Experiment 4 provide support for the DLT's discourse-based integration cost metric. Reading time differences in this experiment were evident at the embedded verb and at the main verb in the RC conditions- locations where DLT integration costs are most differentiated among conditions. According to the DLT's integration cost metric, the RC full NP should have been the most complex condition at the embedded verb, followed by the RC pronoun condition, and the CC conditions should have been the least complex. No difference was predicted between the CC conditions at the embedded verb. This pattern was mostly reflected in the reading time data. As predicted there was no difference at the embedded verb between the CC conditions, but there was a difference between the RC conditions, with the RC full NP condition read more slowly than the RC pronoun condition. In region 5, the NP object of the preposition, there were no differences between the conditions, as predicted by the DLT. At the main verb in the RC conditions, region 6, reading times were also as predicted by the DLT, with the full NP condition slower than the pronoun condition.

The results of Experiment 4 support a resource-usage theory including syntactic representations over one that only includes discourse representations. There was a difference between the RC conditions at the main verb, as predicted by the DLT but not by a purely discourse-based resource theory. These results suggest that the appropriate distance metric for computing processing load is based on the amount of discourse processing intervening between syntactic integrations and not just between instances of accessing particular referents.

Reading times in Experiment 4 were also consistent with predictions made by the DLT's storage cost component. In regions 2, 3 and 4, the RC conditions were read more slowly than the CC conditions. This is predicted because in the RC conditions syntactic predictions of a verb and a gap position must be stored until the processing of the embedded verb. Storing these predictions uses processing resources, and consequently increases reading times over the embedded clause. In contrast, the CC conditions do not require the storage of syntactic predictions over the embedded clause, and so their reading times are predicted to be shorter.

1.6 Experiment 5

The evidence from Experiment 4 does not distinguish between a binary distance metric like that proposed in Gibson (1998) and a continuous metric like the one supported by Experiment 2. Experiment 5 was designed to distinguish between the binary and continuous metrics using self-paced reading. Recall that in Experiment 2, sentences with a more accessible referent interrupting multiple long-distance dependencies were judged easier to understand than sentences with a less accessible referent in the same position. Experiment 5 tests whether these graded effects of referent accessibility are apparent in reading times at the conclusions of long-distance dependencies, as predicted by the DLT.

1.6.1 Method

Participants

18 Northeastern undergraduates and 30 members of the MIT community were participants in this study, which took approximately 20 minutes to complete. The participants from Northeastern participated for course credit, while the MIT participants were paid $8 for their participation. As the materials were based on materials from Experiment 2, participation in Experiment 5 was limited to people who had not participated in Experiment 2. All of the participants also participated in another experiment during the same testing session and spent

approximately an hour in the lab, with a short break between experiments. For some participants the other experiment was an unrelated self-paced reading experiment, while for others it was an auditory language processing experiment.

Materials

Experiment 5 tested 24 items, with four conditions per item. The four conditions all had the same structure: a sentence with a subject-modifying object-extracted RC. The subject of the RC was varied between a 1st or 2nd person pronoun, a famous name, a definite description and an indefinite description, as in (9).

9. The writer who {I, NBC, the reporter, a reporter} talked to wrote radical articles about the government.

In order to lessen the likelihood that any differences in reading times could be due to plausibility differences between the conditions, 36 potential items were pre-tested in a plausibility questionnaire. In the questionnaire, sentences were presented in a shortened, non-nested form, created by moving the matrix subject to the object position in the embedded clause. For example, (12) was presented as "{I, NBC, the reporter} talked to the writer." Only the pronoun, famous name and definite description conditions were tested in the plausibility probe, because the indefinite conditions were very similar to the definite conditions. Subjects were asked to rate the naturalness of the events described in the sentences on a scale of 1 to 7, 7 being very unnatural or unlikely. 24 items were selected from the initial set of 36. These 24 items were matched for their plausibility ratings across the three conditions. The mean plausibility of the conditions was 2.71 for the pronoun condition, 2.89 for the famous name condition and 2.65 for the definite condition.

During the course of Experiment 5, participants read 24 target sentences, 20 sentences from an unrelated experiment and 36 filler sentences. The four conditions were counterbalanced across lists and the lists were randomized. The complete set of items is presented in Appendix E.

Procedure

The procedure was the same as in Experiment 4.

1.6.2 Results

All of the participants answered more than 75% of the comprehension questions correctly across the experiment, so none were eliminated from the analysis. Table 1.4 presents the percentages of comprehension questions that were answered correctly in each condition.

Pronoun	94.1
Famous Name	94.4
Definite	91.3
Indefinite	88.9

Table 1.4 Percentage of comprehension questions answered correctly for each condition in Experiment 5

The pronoun and famous name conditions were comprehended correctly most often, while the definite condition was comprehended correctly less often and the indefinite condition least often. Qualitatively, these percentages mirror the complexity results in experiment 2, except that the pronoun and famous name conditions are virtually identical. A repeated measures ANOVA indicated that there was a difference among the question answering rates for the four conditions that was significant by participants and marginal by items (F1(3,141)= 2.76, MSe= .01, p<.05; F2(3,69)= 2.22, MSe= .01, p=.09).

Reading times were analyzed for sentences for which participants correctly answered the comprehension question. In addition, reading times were trimmed at 5 standard deviations from the mean of each word position in each condition, eliminating less than one percent of the remaining data. Words were grouped into the regions presented in Table 1.5 for the purpose of analysis.

1	2	3	4	5	6	7
The writer	who	{I, NBC, the/a reporter}	talked to	wrote	radical articles	about...

Table 1.5 Regions of analysis for Experiment 5

Mean unadjusted and length adjusted reading times for each region are shown in Table 1.6.

	The writer	who	{I, NBC, the/a reporter}	talked to	wrote	radical articles	about...
Indexical pronoun	334 (-16)	344 (9)	319 (-3)	362 (-7)	373 (5)	359 (2)	361 (11)
Famous name	350 (-3)	365 (27)	359 (2)	352 (-23)	400 (32)	364 (6)	344 (-6)
Definite description	338 (-13)	339 (3)	341 (-14)	381 (11)	416 (47)	369 (12)	353 (3)
Indefinite description	331 (-21)	348 (12)	334 (-14)	415 (47)	428 (61)	386 (31)	355 (7)

Table 1.6 Mean reading times expressed in milliseconds per word (length adjusted times are in parentheses) for each region in Experiment 5.

Only the statistical tests for unadjusted times are reported; the results were similar for the statistical tests that were performed on length-adjusted times. There were no reliable differences between participant groups (Northeastern, MIT), so analysis will be presented over both groups together. Figure 1.5 graphs the mean unadjusted reading times for each region.

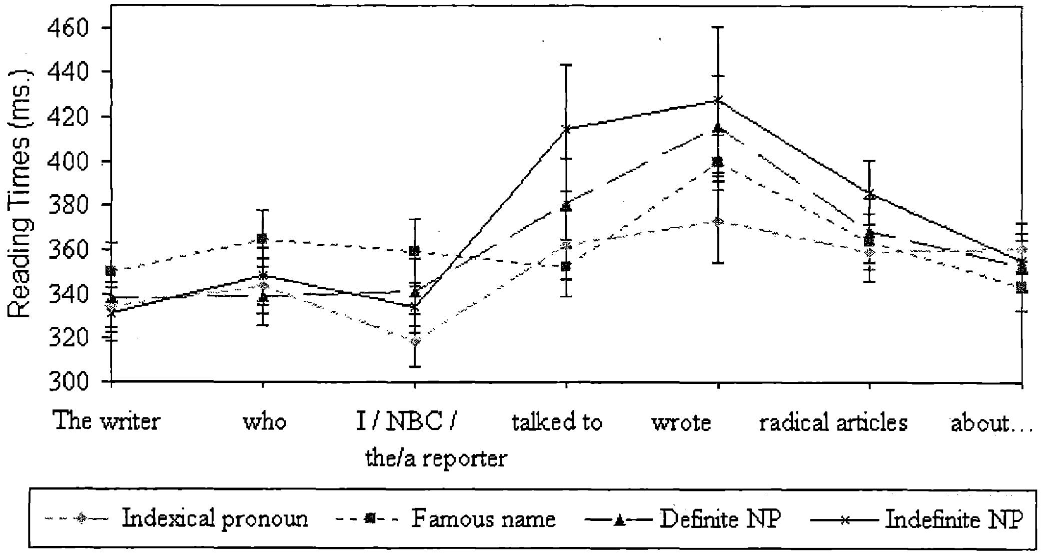

Figure 1.5. Reading times for Experiment 5

All of the conditions contained the same words in regions 1 and 2, so no tests were done in those regions. In region 3, the embedded subject NP, a repeated measures ANOVA showed a difference among the conditions that was significant by participants and marginal by items (F1(3,138)= 5.06, MSe= 2521, p<.005 ; F2(3,69)= 2.71, MSe= 5782, p=.05). The pattern of means seems to reflect the length and number of words in the region. Pronouns, which were one to three letter words, had the fastest reading times. Definite and indefinite descriptions had intermediate average reading times. Famous names, which were often two full words (e.g. "Michael Jordan"), had the slowest reading times. In region 4, the embedded verb "talked to", there was a difference in reading times among the conditions that was significant in the participants analysis but not in the items analysis (F1(3,138)= 3.65, MSe= 10188, p<.05; F2(3,69)= 1.92, MSe=13164 , p =.13). Numerically, the pattern at this region was close to that predicted by the DLT. In particular, the indefinites were slowest, the definites were slightly faster, and the pronoun and famous name conditions were numerically fastest. In region 5, the main verb "wrote," the means were in the order predicted by the DLT, but the overall ANOVA did not show a significant difference, Fs < 2.5, ps> .09. In region 6, there was a difference among conditions that was significant by participants but not by items (F1(3,138)= 2.75, MSe= 2425, p<.05; F2(3,69)=1.3, MSe= 4193, p=.28). Region 7, which contained the remainder of the sentence, showed no reliable differences among the conditions, Fs < 2.

In order to capture possible spillover effects, we performed an additional analysis of region 5 (the main verb) together with the following word, the first word from region 6. On this combined region, a repeated measures ANOVA showed differences in the predicted direction (F1(3,138)= 3.65, MSe= 5384, p<.05; F2(3,69)= 2.65, MSe= 5684, p=.06). Planned comparisons showed that the definite and 1st/2nd pronoun conditions differed marginally (F1(1,47)= 3.61, MSe= 3444, p=.06; F2(1,23)= 3.07, MSe= 2816, p=.09), while the 1st/2nd pronoun and indefinite conditions differed significantly (F1(1,47)= 9.87, MSe= 5774, p<.005; F2(1,23)= 8.72, MSe= 2958, p<.01). The other comparisons revealed no significant differences, although they were numerically in the direction predicted by the DLT.

A repeated measures contrast test at the matrix verb region showed that the increase in reading times from the pronoun to the famous name to the definite to the indefinite condition reflected a significant an increasing linear trend, (t(47)=2.23, p<.05, $r_{contrast}$ =.31; t(23)=2.14, p<.05, $r_{contrast}$ = .41). This trend is the same as the trend evident in the data from Experiment 2. In Experiment 5, nested sentences with embedded subject NPs that required more referential processing had longer reading times at the main verb than sentences with NPs requiring less referential processing. The same pattern was evident in reading times for the pronoun, definite and indefinite conditions at the embedded verb, but the famous name condition reading times were slightly faster than those of the pronoun condition instead of falling between the pronoun and definite conditions, as predicted.

1.6.3 Discussion

The results of Experiment 5 are consistent with a referent-based distance metric that has its foundation in a continuum of referential accessibility. The differences between complexity judgments in Experiment 2 appeared as differences in reading times on verbs in Experiment 5 and also in the rates of correctly answered comprehension questions across the conditions. Though not every prediction of the accessibility-based distance metric was confirmed in every region, the overall reading time pattern was consistent with the DLT's predictions.

1.7 General Discussion

The experiments presented in this chapter: 1) support and extend the DLT, a theory of computational resource usage in sentence processing, 2) provide new evidence about the resource demands of discourse processing during sentence processing and 3) suggest that the resource demands of referential processing have an effect on the resources available for further syntactic processing. In the following, each of these claims will be discussed in turn.

First, these experiments provided evidence supporting many of the basic assumptions of Gibson's (1998) DLT and led to the theory's modification and expansion. Rather than an integration cost distance metric based on binary discourse costs, such as the one proposed in Gibson (1998), the evidence presented here indicates that a metric based on continuous discourse costs is a better fit to the data. Data from Experiments 2 and 4 suggest that the metric should take into account the ease of establishing a referent for an NP, including the difficulty of accessing a previously existing representation or instantiating a new representation if necessary. In light of these findings, one way to improve Gibson's (1998) formulation of integration cost is to assign each referent a graded cost based on its accessibility and then to use those graded costs in calculating integration costs. Easily accessible referents will have lower costs than newly introduced referents, leading to lower costs for the integrations which cross them. Subsequent experiments in this thesis will further test this new version of the DLT.

Second, the results of these four experiments suggest that there is a cost associated with identifying a referent for an NP. The existence of such a cost is not unexpected, considering past discourse research (e.g. Murphy, 1984; Garrod & Sanford, 1982), but the findings showing that the location and amount of the cost are dependent on sentence structure are new. Results in this chapter also extend the findings of corpus research showing correlations between referring forms and discourse status (Ariel, 1988; Arnold, 1998; Gundel et al., 1993) by showing a differential processing cost based on referring form. The fact that changing a referring form can spur differential processing costs, and that those costs pattern in the way predicted by an independently established accessibility hierarchy demonstrates that referring form is an important indicator of discourse status to the referential processing system.

Third, evidence presented in this chapter also bears on the architecture of the human sentence processing mechanism (HSPM). The results presented in this chapter suggest that the resource demands of referential processing can decrease the resources available for subsequent structural processing. The finding in Experiment 4 that syntactic integration difficulty at a verb is affected by the discourse status of intervening NPs demonstrates that the resources used in syntactic processing and discourse processing are not independent. This is important, in that it makes it unlikely that the human sentence processing mechanism is entirely modular and has separate resource stores for each module. At some level in the system there must be processing resources that can be either directly drawn upon by different linguistic processes, or be used by a more general process that encompasses multiple linguistic processes. Note that this does not mean that the process of constructing a syntactic representation and the process of constructing a discourse representation are completely overlapping; there may be resources devoted to each process independently as well.

All of these conclusions are based on the assumption that the results in this chapter were due to differences in the amount of structural and referential processing required in the experimental items, but the way that referential processing requirements were manipulated also changed the frequencies of the reference forms. This made it possible that the results might actually be due to frequency differences between experimental conditions rather than structural and referential factors. The experiments in chapter 2 will address this possibility.

Chapter 2

2.1 Introduction

The five experiments in chapter 1 tested predictions of the linguistic complexity theory introduced in Gibson (1998). The four additional experiments presented in chapter 2 were designed to build on the findings from chapter 1.

2.1.1 Updating the DLT

The experiments presented in chapter 1 provided a range of evidence refining and supporting the basic claims of the linguistic complexity theory introduced in Gibson (1998). This original version of the DLT included a linguistic integration cost metric that incremented cost based on the number of new discourse referents that intervened between the endpoints of the integration. The formulation of this integration cost metric was as follows:

(1) DLT linguistic integration cost (Gibson, 2000): The structural integration cost associated with connecting the syntactic structure for a newly input head h_2 to the projection of a head h_1 that is part of the current structure for the input is dependent on the complexity of the computations that took place between h_1 and h_2. For simplicity, 1 unit of cost will be counted for each new discourse referent in the intervening region.

The experiments in chapter 1 verified the importance of including a measure of referential processing in the integration cost distance metric. Changes in the discourse status of referents intervening between the endpoints of dependencies were shown to affect reading times at the close of the dependencies. Evidence from intuitive complexity questionnaires and self-paced reading time experiments showed that integrations crossing new referents are more difficult to complete than integrations crossing old referents.

But though the results reported in chapter 1 suggested that the DLT's integration cost measure is correct to incorporate referential processing costs in its calculation of complexity, they indicated that the distinction between new and old referents drawn in the formulation of integration cost is overly specific. New referents are a specific category of referents that impose a high referential processing load. Old referents are a specific category of referents that impose a lower referential processing cost. The results presented in Experiment 1 suggested that integration cost should be dependent on the amount of referential processing intervening between the endpoints of dependencies rather than the number of new referents intervening between dependencies. A formulation of integration cost incorporating these changes follows:

(2) Updated DLT linguistic integration cost: The structural integration cost associated with connecting the syntactic structure for a newly input head h_2 to the projection of a head h_1 that is part of the current structure for the input is dependent on the complexity of the syntactic and referential computations that took place between h_1 and h_2. When more referential or syntactic processing takes place between h_1 and h_2, integration cost is greater at h_2.

The experiments in chapter 2 test this updated version of DLT integration cost that takes referential processing load into account when calculating complexity. When the name DLT is used in this chapter, it should be taken to refer to this updated version of the DLT and not the original version that included a binary cost dependent on whether a referent was new or old.

2.1.2 Contextual Manipulations from the Literature

The experiments presented in this chapter used two different kinds of contextual manipulation to vary referential processing load. Experiments 6, 8 and 9 tested sentences preceded by contexts in which a particular referent was or was not introduced. Experiment 7 compared the processing of referents that were either introduced explicitly in context, available through an easy inference, available through a more difficult inference or completely new. In Experiment 7, the amount of processing required to infer the relation between the target referent and the discourse model established in the context sentence was manipulated based on methods used in Haviland & Clark (1974) and Garrod & Sanford (1977).

As discussed in the introduction, Haviland and Clark compared reading times over pairs of sentences where a referent referred to by a definite in the second sentence was either explicitly introduced in the first sentence (3a,c) or could be inferred into the discourse model established by the first sentence (3b,c):

3a. Mary unpacked some beer.
3c. The beer was warm.

3b. Mary unpacked some picnic supplies.
3c. The beer was warm.

In Haviland and Clark's items, the context sentence in the inference condition introduced a schema. For example, in the item above the action of unpacking picnic supplies suggested a picnic schema. The second sentence in the discourse is interpreted within the context of the first, so in this example, the reader must make the inference that one of the unpacked picnic supplies was beer. This extra inference was hypothesized to be the reason that reading times were longer over (3c) when it followed (3b) than when it followed (3a).

Garrod and Sanford (1977) investigated a similar phenomenon. Whereas Haviland and Clark (1974) tested the ease of bridging a new entity into a particular schema, Garrod and Sanford tested the complexity difference engendered by using more and less specific descriptions for an entity. For example, Garrod and Sanford recorded reading times on the second sentence from pairs such as:

4a. A robin would sometimes wander into the house.
The bird was attracted by the larder.

4b. A bird would sometimes wander into the house.
The robin was attracted by the larder.

They found that reading times were shorter in the conditions where the referent was introduced by an exemplar and then subsequently referred to with a category label, as in (4a). Reading times were longer for the second sentence in pairs where the referent was first introduced with a category label and subsequently referred to using an exemplar, as in (4b). Garrod and Sanford argued that this pattern of results occurred because in (4b), the referent was introduced into discourse with a general category description. When the referent was subsequently referred to with an exemplar label, its representation had to be elaborated to include the new information incumbent on its being a particular type of exemplar. This elaboration required processing resources, resulting in slower reading times. In cases where the referent was originally introduced with an exemplar label, the representation for the referent included the fact that it was a member of the more general category. As a result, the subsequent reference using a category label required no change to the representation and required fewer processing resources.

Experiment 7 used manipulations similar to those described above to vary the amount of processing required to interpret a referent. The manipulation similar to the one in Haviland and Clark (1974) involved changing the schema introduced in a context sentence in order to vary the closeness of the relation between the referent and the schema. In the easier inference conditions, the referent was a very natural part of the schema, while in the harder inference conditions, the referent was a less natural part of the schema. The other manipulation was more similar to the one used in Garrod and Sanford (1977). In the items with this manipulation, a set was introduced in context and the target referent (e.g. dog) was an exemplar of that set. In the easier inference conditions, the set was labeled with a category of which the target referent was a typical exemplar (e.g. pet). In the more difficult inference conditions, the set was labeled with a superordinate level category label (e.g. animal). Logically, the amount of referential processing that these manipulations incurred should have been intermediate between the amount of processing required to access an easily accessible referent and the amount of processing required to introduce a completely new, unrelated referent. These manipulations allowed Experiment 7 to test the DLT's predictions about the complexity effects of additional referential processing for four distinct amounts of referential processing.

2.1.3 Profile of Chapter 2

The experiments in chapter 2 used context to manipulate the accessibility of a referent in a position interrupting multiple long distance dependencies. This was preferable to the null context approach taken in the experiments in chapter 1 for a few reasons. First, the state of the discourse model could be more carefully controlled and manipulated. Second, accessibility was no longer a side effect of referential form, as it had been in chapter 1, but now was directly manipulated. Third, the costs or lack of costs for accommodation could be directly identified rather than hypothesized. Fourth, the use of context allowed the accessibility of a referent in a sentence to be varied without changing any of the words in that sentence. This allowed for tests verifying that the effects reported in chapter 1 resulted from changing the amount of discourse processing intervening between the endpoints of a dependency and not from an unrelated side-effect of changing a referential form.

The experiments in chapter 2 were as follows. Experiment 6 was a questionnaire comparing the complexity of sentences that introduced a new referent and sentences that referred to a previously established referent. Experiments 7, 8 and 9 were self-paced reading experiments testing sentences with long distance dependencies crossing a referent whose referential status was manipulated by means of a previous context. Experiment 7 tested sentences with referents at four different levels of accessibility, while Experiments 8 and 9 tested sentences with referents at two different levels of accessibility.

2.2 Experiment 6

Experiment 6 was a complexity questionnaire testing the prediction that increasing the amount of referential processing between the endpoints of long distance dependencies results in increased complexity at their right endpoints. The referential manipulation in Experiment 6 consisted of comparing new referents that were anchored in context with referents that had already appeared in context. Care was taken to make the introduction of the new referent as natural as possible, so that increased processing in the new condition would be due to the building of additional structure and not to extra inferences required to relate the new referent to discourse. Accessing a referent from an immediately preceding context sentence was predicted to require less resource usage than building a representation for a new referent.

Like Experiments 1 and 2 in chapter 1, this experiment compared doubly nested sentences where the subject of the most embedded clause was either new or old to discourse. Unlike those experiments, context sentences in Experiment 6 served to establish the old referent in discourse and to provide a felicitous context for the introduction of the new referent.

2.2.1 Method

Participants

Twenty-four native English speakers were recruited from the MIT community to fill out a questionnaire that took approximately fifteen minutes to complete. They were paid $4 for their participation.

Materials

Experiment 6 tested 20 doubly nested items in a 2 x 2 design crossing structure (nested vs. right branching) with referent status (old vs. new). Because nested structures have more long distance dependencies than right branching structures, the DLT predicts that manipulations of referent status will have large effects in nested sentences but small or no effects in right branching sentences. Referent status was manipulated using a context sentence. In all conditions, the context sentence used a proper name to introduce a referent. In the old referent condition, that referent was referred to again by name in the target sentence. In the new referent condition, the name from context was used in the genitive form to introduce a new referent in the target sentence. This design made the introduction of the new referent felicitous, because it was introduced in relation to an existing referent and thus was anchored in the discourse model (Prince, 1981). In addition, the context sentences were specifically designed to make the new referent follow naturally. For example, in (5), the context sentence introduces "Susan" and her "artistic family" in order to make the reference to "Susan's brother" as natural as possible. These precautions were taken in order to minimize any potential effects of accommodation. In the nested condition, the new or old referent was the subject of the most embedded clause. In the right branching conditions, the new or old referent was in the matrix subject position.

Context:

5. Susan comes from a very artistic family.

Old referent/ nested

5a. The classmate who a cartoon character which Susan created fascinated was impatient to see the new comic strip.

Old referent/ right branching

5b. Susan created a cartoon character which fascinated the classmate who was impatient to see the new comic strip.

New referent/ nested

5c. The classmate who a cartoon character which Susan's brother created fascinated was impatient to see the new comic strip.

New referent/ right branching

5d. Susan's brother created a cartoon character which fascinated the classmate who was impatient to see the new comic strip.

Because the new referents in these items are easily accommodated into discourse, differences between the new and old conditions should be primarily due to the resource demands of referent building rather than the resource demands of accommodating new referents.

The questionnaire in Experiment 6 was made up of 20 experimental items and 56 fillers, all of which consisted of two sentences. Many of the fillers contained proper names or genitives, like the target items. The four conditions of each target item were counterbalanced across lists. Participants read only one condition per item and five versions of each condition across the questionnaire. The lists were pseudo-randomized so that at least one filler intervened between each pair of experimental items and the order of the questionnaire pages was varied for each participant. The complete set of items is included in Appendix F.

Procedure

Participants were asked to rate sentences on a scale of one to five, one being "easy to understand" and five being "hard to understand". Participants were told that each sentence would be preceded by a context sentence, and their task was to rate the second sentence in the context provided by the first. The questionnaire began with a page of instructions asking participants to make their judgments based on their first impressions without reading sentences more than once. In the instructions, participants were given five practice items with a brief discussion of the sort of ratings each of the practice items might be assigned. All of the practice items followed naturally from their context sentences. The first two example sentences were relatively comprehensible, while the final three were more difficult to understand. One of the more difficult sentences was doubly nested, with extra prepositional phrases modifying the verbs: "The caterer who the woman who the dog bit on the ankle met at the party talked to her yesterday in the late afternoon." Ratings in the 1 or 2 range were suggested for the easier sentences, and ratings in the 3, 4 or 5 range were suggested for the more difficult sentences, but participants were advised that individuals often differ on which sentences they find easier or harder to understand.

2.2.2 Results

The average complexity ratings for the four conditions in Experiment 6 are graphed in Figure 2.1.

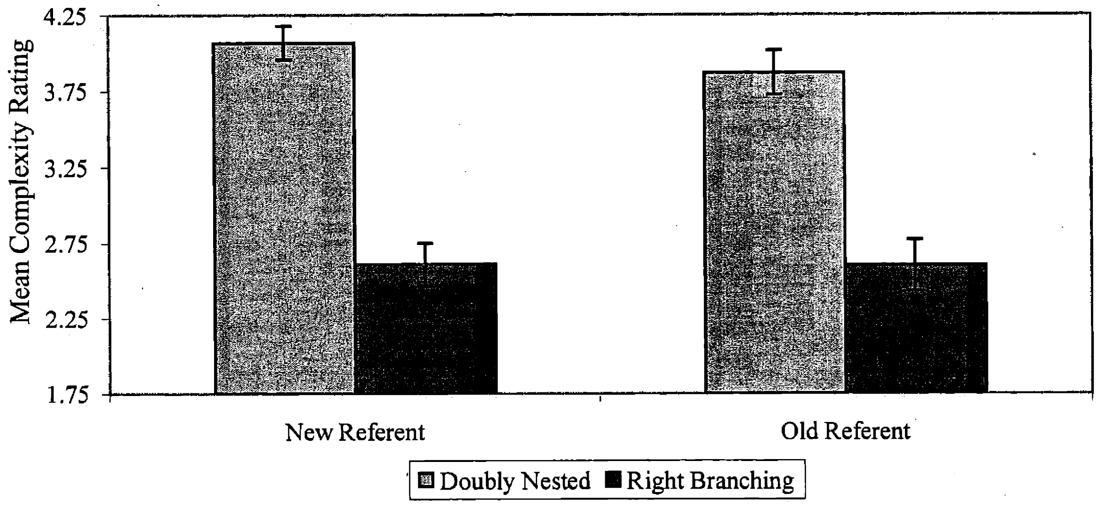

Figure 2.1. Mean complexity ratings in Experiment 3

A repeated-measures ANOVA indicated that there was a reliable main effect of structure, such that nested sentences were more complex than right branching sentences (F1(1,23)=113, p<.001;

F2(1,19)= 171, p<.001). Planned individual comparisons showed a marginal difference between the two nested conditions, such that nested sentences with an old referent were judged less complex than nested sentences with a new referent (F1(1,23)= 3.29, p=.08; F2(1,19)=3.27, p=.09). There were no differences between the right branching conditions (Fs < .1, ps> .9).

2.2.3 Discussion

The results of Experiment 6 are consistent with the prediction that increasing the referential processing load between the endpoints of long distance dependencies increases processing difficulty at the right endpoint of those dependencies. In the right branching conditions, where the referentially manipulated NP interrupted no dependencies, there was no difference between the complexities of sentences with new or old referents. But in the nested conditions, where the referentially manipulated NP interrupted four dependencies, the conditions with new referents were judged more complex than the conditions with old referents. Because accommodation difficulty was minimized for the new referents, this difference must have been primarily due to the extra processing required to build a representation for a new referent.

In Experiment 6, as well as in all of the experiments presented in chapter 1, discourse status manipulations were carried out by varying one of the NPs in a target sentence. Experiment 7 used an alternative method to manipulate a referent's status in discourse more directly.

2.3 Experiment 7

Experiment 7, like Experiment 6, tested the hypothesis that long distance dependencies crossing heavier referential processing loads are more difficult to complete than dependencies crossing lighter referential processing loads. This experiment avoided potential confounds that were present in the first six experiments, because the words and referents in the target sentence were kept constant across conditions. In Experiment 7, the amount of processing required for one of the referents in the target sentence was varied by using different contexts to introduce the target sentence.

2.3.1 Method

Participants

Seventy-three native English speakers were recruited from the MIT and Northeastern communities to participate in a word-by-word self-paced reading study. Participants took approximately 20-25 minutes to complete the experiment. All participants also participated in an unrelated self-paced reading study during the same testing session, and spent approximately one hour in the lab, with a short break between the two sessions of self-paced reading. Participants at MIT were paid $8, while participants at Northeastern participated in partial fulfillment of a requirement of an introductory psychology class.

Materials

Experiment 7 tested 16 experimental items, with four conditions per item. Each item consisted of a context sentence and a target sentence with a subject-modifying object-extracted RC. The difference between conditions was not in the target sentences, but rather in the context sentences. The context sentences served to vary the status of the referent in the embedded subject position of the target sentences. The target sentence was preceded by one of three types of context or presented with no context. The context either explicitly mentioned the entity that served as the subject of the RC (6a), mentioned a set from which the subject of the RC could be easily bridged (6b), or mentioned a set or entity from which the subject of the RC could be bridged with more effort (6c). The final condition, (6d), had no context.

Context:

6a. The company CEO analyzed each employee's performance in January.

6b. The company executives analyzed each employee's performance in January.

6c. The company analyzed each employee's performance in January.

6d. no context

Target:

One of the employees who the CEO promoted completed every project on schedule.

In the no context condition (6d), a sentence was added following the target sentence, so that all items consisted of two sentences.

A set of 24 potential items was run in a norming study in order to confirm that bridging in the easy-bridge conditions was in fact easier than bridging in the difficult-bridge conditions. Participants were presented with one of the three context sentences, followed by the NP that served as the RC subject. They were asked to rate how natural it would be to follow the first sentence with a sentence beginning with the given NP. The norming study also tested the matrix subject from the target sentence in the same way, to determine whether it followed equally naturally from each of the three contexts. The final set of items for the experiment included every item for which: 1) the matrix subject followed equally naturally from the three contexts and 2) the bridged referent followed most naturally from the explicit context, less naturally from the easy bridge context and least naturally from the hard bridge condition. The set of 24 potential items only included a subset of 12 that satisfied these requirements, so four additional items were modified so that they satisfied the criteria according to the experimenters' judgments. The most frequent adjustment made was to increase the difficulty of the hard bridge inference so that it would be more different from the easy condition.

The 16 experimental items in Experiment 7 were combined with 12 sentence pairs from Experiment 8and 32 pairs of filler sentences to form six lists. The four conditions were counterbalanced across lists and the lists were randomized. The complete set of items is presented in Appendix G.

Procedure

Participants performed self-paced reading in a word-by-word moving window display (Just, Carpenter & Wooley, 1982) on a Macintosh computer running software developed in our lab. At the start of each trial, two sentences appeared on the screen with all characters replaced by dashes. Participants pressed the space bar to change a string of dashes into a word. Each time the space bar was pressed, the next word appeared and the previous word reverted back into dashes. Time between bar-presses was recorded, as a measure of how long participants spent reading each word. After finishing each sentence, participants were required to answer a yes-no comprehension question about the sentence. They answered by pressing one of two keys, and if they answered incorrectly, the word "INCORRECT" flashed on the screen. No feedback was given for correct responses. Participants were asked to read at as naturally as possible and to take incorrect answers as an indication to read more carefully.

Up to 100 characters could appear on each line of the display. Each sentence spanned from one to one and one-half lines. The second sentence always began on a new line. The embedded verb region and the main verb region in the target sentence always appeared on the first line. Participants were given a small set of practice items and questions before the experiment in order to familiarize them with the task.

2.3.2 Results

One participant's data was excluded because of below chance performance on the comprehension questions pertaining to the experimental items. All remaining participants answered at least 75 percent of the comprehension questions for this experiment correctly. Table 2.1 presents the percentage of comprehension questions answered correctly for each condition in Experiment 7.

Explicit mention	83
Easy bridge	90
Hard bridge	86
No context	91

Table 2.1 Percentage of comprehension questions answered correctly for each condition in Experiment 7.

Participants were most accurate on questions about the no context condition, less accurate on questions about the easy bridge condition, even less accurate on questions concerning the hard bridge conditions, and least accurate on questions concerning the explicit mention conditions. Comprehension questions were the same across the explicit mention, easy bridge and hard bridge conditions, but seven of the 16 items had a different question for the no context condition. Because of this, no comparison can be made between question answering accuracy in the no context condition and the other conditions. An ANOVA over the means for the three conditions with the same questions showed that there was a marginal difference among them by subjects and a reliable difference by items (F1(2,142)= 2.89, MSe= .03, p=.06; F2(2,30)= 4.59, MSe= .004, p<.02). Questions were answered correctly significantly more often in the easy bridge condition than in the explicit mention condition, (F1(1,71)= 5.30, MSe= .03, p<.05; F(1,15)= 16.30, MSe= .002, p<.001). No other differences among the conditions with the same comprehension questions were reliable.

Words were grouped into regions for the purpose of analysis as in Table 2.2.

1	2	3	4	5	6	7
One of the employees	who	the CEO	promoted	completed	every project	on schedule.

Table 2.2 Regions of analysis for Experiment 7.

Region 1 consisted of the matrix subject. Region 2 contained the relative pronoun "who." Region 3 contained the subject of the embedded RC. Regions 4 and 5 were the embedded verb and main verb respectively. Region 6 contained the NP following the main verb, and region 7 included the remainder of the sentence.

Reading times were analyzed for sentences for which participants correctly answered the comprehension question. In addition, reading times were trimmed at 2.5 standard deviations from the mean for each word position in each condition. This trimming eliminated less than three percent of the remaining data. There were no reliable differences between participant groups (Northeastern, MIT) so we present our analysis over both groups together. Statistical tests were performed on both raw reading times and reading times that were adjusted for word length following the procedure proposed by Ferreira & Clifton (1986) and Trueswell, Tanenhaus & Garnsey (1994). The trends in unadjusted and adjusted reading times were the same. The analyses reported below are over unadjusted times, except where noted. Table 2.3 presents the mean unadjusted and length adjusted reading times for each region in Experiment 7.

	One of the employees	who	the CEO	promoted	completed	every project	on schedule
Explicit	360	329	332	392	390	378	373
mention	(-16)	(-40)	(-60)	(-25)	(-21)	(0)	(-5)
Easy	356	332	355	426	445	384	375
bridge	(-18)	(-37)	(-33)	(-5)	(34)	(8)	(-3)
Hard	348	346	362	411	436	372	372
bridge	(-26)	(-22)	(-30)	(-18)	(25)	(-3)	(-8)
No context	338	352	379	466	516	410	410
	(-38)	(-18)	(-13)	(34)	(108)	(34)	(30)

Table 2.3. Mean reading times (length adjusted times in parentheses) for each region in Experiment 7.

Figure 2.2 graphs the mean unadjusted reading time for each region in Experiment 7.

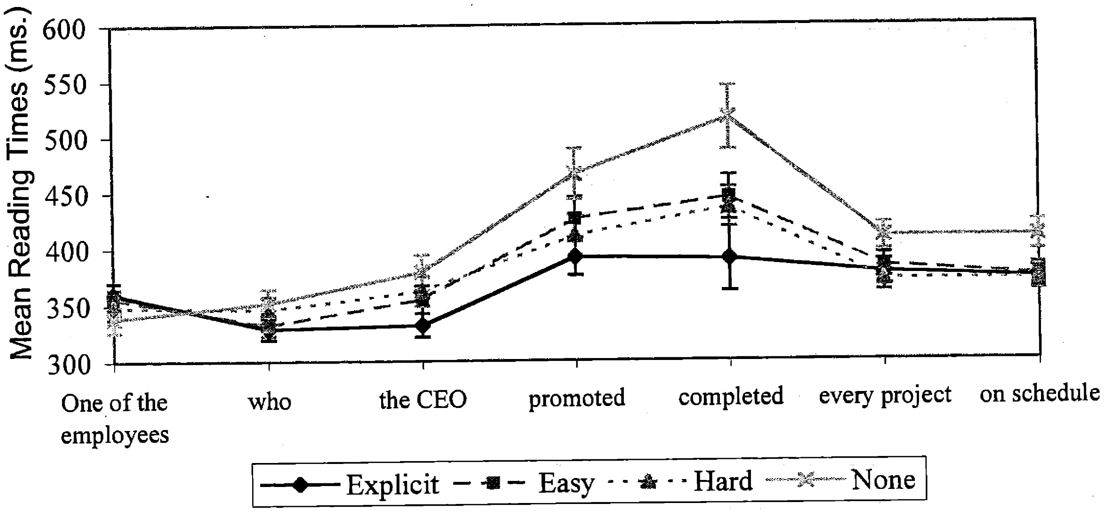

Figure 2.2. Mean reading times for each region in Experment 7

There were no differences among any of the conditions in region 1. In region 2, the relative pronoun, ANOVAs showed differences among the four conditions that were reliable by subjects, but not by items (F1(3,213)=3.99, MSe= 2184, p=.009; F2(3,45)= 1.02, MSe= 1078, p=.395). At region 3, the embedded subject, reading times follow the pattern predicted by previous discourse processing work (Garrod & Sanford, 1994), with the explicit condition being faster than the bridged conditions and the bridged conditions faster than the no context condition. ANOVAs over all four conditions show that there is a reliable difference among reading times in this region (F1(3,213)= 8.90, MSe= 3040, p<.001; F2(3,45)= 3.00, MSe= 1370, p=.04). Reading times on the no context condition were reliably slower than reading times on the explicit mention condition (F1(1,71)= 14.70, MSe= 3329, p<.001; F2(1,15)= 6.87, MSe= 1366, p=.019). No differences between the other conditions approached significance when treating items as a random factor. This same pattern of reading times, with the explicit condition fastest, the bridged conditions slower and the no context condition slowest, also appeared on the embedded verb, region 4, but the differences were greater, as predicted by the DLT (F1(3,213)= 7.09, MSe= 9928, p<.001; F2(3,45)= 3.12, MSe= 4906, p=.035). The explicit mention condition was read faster on the embedded verb than the no context condition (F1(1,71)= 12.30, MSe= 14007, p=.001; F2(1,15)= 6.19, MSe= 6683, p=.025). The hard bridge condition was also read faster than the no context condition in this region (F1(1,71)= 10.07, MSe= 13886, p=.002; F2(1,15)= 4.70, MSe= 7007, p=.047). The explicit mention condition was read more quickly than the easy bridge condition, though this difference reached significance only in the subjects analysis (F1(1,71)= 4.61, MSe= 11154, p=.035; F1(1,15)= 2.14, MSe= 5476, p=.16). None of the remaining pairwise comparisons showed reliable differences. At region 5, the matrix verb, these differences continued and their magnitudes increased (F1(3,213)= 11.05, MSe= 17615, p<.001; F2(3,45)= 5.29, MSe= 6246, p=.003). In region 5, the no context condition was read more slowly than each of the other conditions as follows: the explicit mention condition (F1(1,71)= 15.56, MSe= 21779, p<.001; F2(1,15)= 14.48, MSe= 6106, p=.002), the easy bridge condition (F1(1,71)= 8.45, MSe= 21104, p=.005; F2(1,15)= 6.32, MSe= 7038, p=.024), and the hard bridge condition (F1(1,71)= 11.33, MSe= 20752, p=.001; F2(1,15)= 5.09, MSe= 10707, p=.039). Differences between no other pairs were reliable in the analysis over raw times. In the analysis over length adjusted times, the differences that were reliable in the analysis over raw times remained reliable and two other differences became reliable as well. The difference between the easy bridge and explicit condition was reliable in the analysis over length adjusted reading times (F1(1,71)= 7.66, MSe= 14354, p=.007; F2(1,15)= 8.13, MSe= 2190, p=.012), as well as the difference between the explicit and hard bridge conditions (F1(1,71)= 6.42, MSe= 11800, p=.013; F2(1,15)= 4.56, MSe= 3239, p=.050). The difference between the easy and hard conditions in region 5 was not reliable under any analysis. At region 6, the NP object of the matrix verb, reading times become more similar again, though ANOVAs over the four conditions show there is a reliable difference among the means by subjects but not by items (F1(3,213)= 6.41 MSe= 3138, p<.001; F2(3,45)= 2.02, MSe= 1336, p=.124). In region 7, the remainder of the sentence, there is a difference among the conditions that is reliable by subjects and marginal by items (F1(3,213)= 9.87, MSe= 2511, p<.001; F2(3,45)= 2.56, MSe= 1083, p=.067). This difference is due to the fact that the no context condition is read more slowly than the other conditions.

2.3.3 Discussion

The results of Experiment 7 provided support for a distance metric that takes the costs of referential processing into account. Because the target sentences were the same in every condition, all reading time differences must have been due to differences in the contexts. As predicted by the DLT, reading times at the embedded and main verbs were slowest for the no context condition, faster for the two bridging conditions and fastest for the explicit mention condition. These differences were reliable in an analysis over length adjusted reading times at the main verb position. The same reading time profile was found on the preceding NP "the CEO", though the differences were smaller. This reading time pattern on the NP is predicted on the basis of discourse processing alone (see Garrod et al, 1994; Garrod and Sanford, 1994).

The results of Experiment 7 did not show the expected difference between the easy and hard bridge conditions. In fact, what little difference there was between these two conditions was in the non-predicted direction, with the easy bridge condition read numerically more slowly than the hard bridge condition. It is possible that this lack of a difference could have resulted from the small magnitude of the difference between the difficulty of the inference in the easy and hard bridge conditions. The difference between these conditions was not great in the norming study, and though some items were subsequently changed to make the difference greater, it is possible that these changes did not have the intended effect, leaving the complexities of the two bridging conditions similar.

The results of Experiment 7 are important, in that reading time differences could not have been due to frequency or number of words or inter-sentential plausibility. However, the target regions in Experiment 7 were all adjacent, allowing for the possibility of reading time spillover. Experiments 8 and 9 test referents at fewer levels of accessibility than were tested in Experiment 7, but they eliminate the possibility of reading time spillover.

2.5 Experiment 8

Like Experiment 7, Experiment 8 tested the prediction that changing the status of a referent in context would affect reading times at the right endpoint of structural dependencies crossing that referent. Experiment 8 compared reading times across nested sentences where the subject of an embedded object-extracted RC either had been introduced in the previous context or had not been introduced in the previous context. The comparison in Experiment 8 was similar to the comparison between the "explicit" and "no context" conditions from Experiment 7, but in Experiment 8 both conditions were introduced by a supportive context.

2.5.1 Method

Participants

Seventy-three native English speakers were recruited from the MIT and Northeastern communities to participate in a word-by-word self-paced reading study. Participants took approximately 20-25 minutes to complete the experiment. All participants also participated in an unrelated self-paced reading study during the same testing session, and spent approximately one hour in the lab, with a short break between the two sessions of self-paced reading. Participants at

MIT were paid $8, while participants at Northeastern participated as partial fulfillment of a requirement in an introductory psychology class.

Materials

Experiment 8 tested 12 experimental items, with three conditions per item. Each item consisted of a context sentence and a target sentence with a subject-modifying object-extracted non-restrictive RC. The conditions reflected two experimental manipulations. The first manipulation was whether the referent that was the subject of the RC in the target sentence had (old) or had not (new) been introduced in the context sentence. In the conditions where the referent was introduced in the context sentence, it was always introduced as the agent of a by-phrase dependent on the passive main verb. The second manipulation was whether the subject of the RC in the target sentence was a definite description or an indefinite description. Only three of the possible four combinations of crossing these two factors were tested: the new/indefinite, old/definite and new/definite combinations. The words that differed among conditions are underlined.

New / Indefinite

7a. A suspect in a bank robbery was caught on Friday.
7a. The suspect, who a detective had sighted on Wednesday, struggled but was eventually subdued.

Old / Definite

7b. A suspect in a bank robbery was caught by a detective on Friday.
7b. The suspect, who the detective had sighted on Wednesday, struggled but was eventually subdued.

New / Definite

7c. A suspect in a bank robbery was caught on Friday.
7c. The suspect, who the detective had sighted on Wednesday, struggled but was eventually subdued.

The new / indefinite condition and the old / definite condition represented the most natural and felicitous ways to refer to new and old referents respectively. New referents are most felicitously introduced with indefinite NPs. Old referents are most felicitously referred to with definite NPs. This means that the new / definite and old / indefinite pairings are both less felicitous. The infelicity in the new/definite condition makes it somewhat awkward to read the definite, but does not interfere with the meaning of the passage because readers can infer the definite's referent into discourse (Haviland & Clark, 1974; Heim, 1982). The old / indefinite condition was not tested, because indefinites introduce new referents and can not be taken to refer to old referents. For example, introducing "a detective" in the context sentence and then referring to "a detective" in the subject of the RC in the target sentence leads the reader to assume that there are two detectives being discussed in the passage.

All target sentences began with a definite description referring to the referent that had been introduced as the subject of the context sentence. This subject was modified by an object-extracted non-restrictive RC, set off by commas. The subject noun of the RC was a single word in eleven of the twelve items and in one item it was two words. This subject was always followed by the auxiliary "had" and a one-word verb. The RC finished with a phrase between one and four words long modifying the verb. This phrase served to separate the embedded and main verbs, in order to eliminate the possibility that reading times on the main verb could be contaminated with spillover from reading times on the embedded verb. After the RC, there was a one-word main verb. After this verb, the target sentences finished with between three and nine additional words.

The 12 experimental items in Experiment 8 were combined with 16 sentence pairs from Experiment 7 and 32 pairs of filler sentences to form six lists. The three conditions were counterbalanced across lists so participants saw one condition from each experimental item and four examples of each condition. The lists were randomized. The complete set of items is presented in Appendix H.

Procedure

The procedure was the same as in Experiment 7.

2.5.2 Results

The data from one participant who correctly answered only 40% of the comprehension questions pertaining to the experimental items was excluded. All of the remaining participants answered at least 66% of such questions correctly. The percentage of comprehension questions answered correctly for each condition in Experiment 8 is shown in Table 2.4. There were no reliable differences between these percentages.

New – Indefinite	89.6
Old – Definite	88.3
New – Definite	92.5

Table 2.4. Percentage of comprehension questions answered correctly for conditions in Experiment 8

Reading times were analyzed on a word-by-word basis, except in two regions. One item had a two-word RC subject (i.e. “wine expert”) rather than a one-word subject (e.g. “detective”). The reading time reported for the subject of the RC includes average reading times for both words from the two-word item. The other region where words were grouped was in the phrase separating the two verbs. The length this phrase varied between one and four words. The first word has been reported individually, but the following zero to three words are presented as one region. This was done for clarity of presentation and also because the small number of data points at most of these word positions meant that there was high variance without grouping. In the tables and figures in this section, the words that represent average reading times over regions rather than over individual words are marked with an asterisk.

Analysis was performed only over sentences for which the comprehension questions had been answered correctly. Reading times that were 2.5 standard deviations from the mean for each position in each condition were trimmed, eliminating 2.7% of the remaining data. There were no differences between participant groups (Northeastern, MIT) that were reliable in both the subjects and items analyses, so we present our analysis over both groups together. Statistical tests were performed on both raw reading times and reading times that were adjusted for word length following the procedure proposed by Ferreira & Clifton (1986) and Trueswell, Tanenhaus & Garnsey (1994). The trends in unadjusted and adjusted reading times were the same. The analyses reported below are over unadjusted times, except where noted. Table 2.5 presents mean unadjusted and length adjusted reading times for each word in Experiment 8.

	The	suspect,	who	the/a	detective *	had	sighted	on	Wednesday, *	struggled	but	was
new-indef	391 (38)	352 (-80)	333 (-30)	321 (-18)	380 (-45)	350 (-14)	387 (-42)	395 (4)	398 (5)	391 (-12)	348 (-32)	356 (-42)
old-def	410 (45)	371 (-54)	346 (-23)	336 (-29)	355 (-57)	348 (-16)	371 (-49)	361 (-35)	381 (-3)	401 (-3)	355 (-17)	372 (-30)
new-def	407 (46)	358 (-74)	345 (-20)	320 (-39)	355 (-63)	349 (-14)	386 (-46)	386 (-8)	410 (21)	391 (-13)	345 (-31)	345 (-53)

Table 2.5. Average reading times (length-adjusted times in parentheses) for each word position in Experiment 8.

Figure 2.3 graphs the mean reading times for each word in Experiment 8.

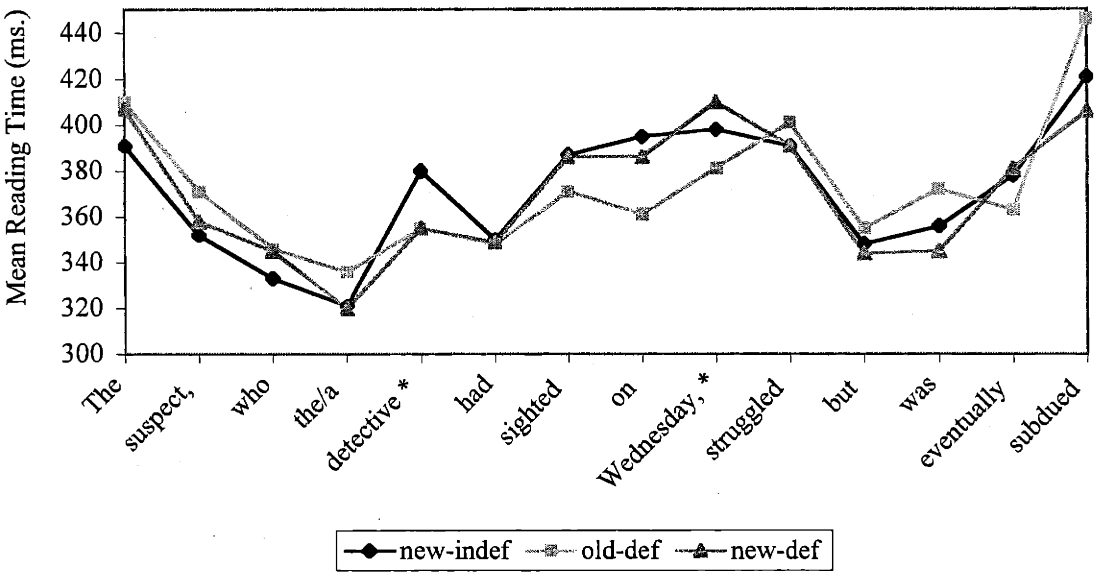

Figure 2.3. Mean reading times for each word position in Experiment 8

The regions of interest in this experiment were the verbs. The DLT predicted that the embedded verb (e.g. "sighted") would be a location of high complexity and that at this word the new conditions would be read more slowly than the old condition. In self-paced word-by-word reading it is not unusual to find reading times spilling over onto the word following a point of high complexity. In this experiment, the difference that the DLT predicted between the new and old conditions did not appear on the embedded verb, but appeared on the immediately following word (e.g. "on"). An ANOVA over all three conditions showed a reliable difference at this word (F1(2,142)= 4.39, MSe= 5197, p=.014; F2(2,22)= 4.40, MSe= 1452, p=.025). Reading times were longer for the new / indefinite condition than they were for the old / definite condition, as predicted by the DLT (F1(1,71)= 9.67, MSe= 4421, p=.003; F2(1,11)= 9.98, MSe= 1234, p=.009). Reading times were also longer for the new / definite condition than the old / definite condition (F1(1,71)=5.19, MSe= 4340, p=.026; F2(1,11)= 6.08, MSe= 899, p=.031).

The DLT also predicted a difference at the matrix verb, (e.g. "struggled"). The new conditions were predicted to be slower than the old condition at this word, but this prediction was not fulfilled. There were no reliable differences between the conditions, and the numerical trend was actually in the opposite direction from the predictions.

2.5.3 Discussion

At the embedded verb, the results of Experiment 8 confirmed the DLT's prediction that integrations crossing new referents would be more difficult to perform than integrations crossing old referents. At the matrix verb, they did not confirm this prediction.

The new / definite condition was included in Experiment 8 to verify that the difference between the new and old conditions in Experiment 8 was not solely due to the different articles that were used in the new and old conditions in order to make them maximally felicitous. The fact that reading times at the embedded verb in the new / definite condition patterned with the new / indefinite condition and were reliably slower than reading times in the old / definite condition showed that the important difference between conditions was referent status and not definiteness. Another reason for including the new / definite condition is that in combination with the new / indefinite condition it replicated the comparison between the definite and indefinite conditions in Experiment 5. The differences between the definite and indefinite conditions were not reliable in Experiment 5, but showed a numerical trend on the verbs with the indefinite condition slower than the definite condition. The same trend appeared on the embedded verb in Experiment 8.

The DLT's prediction of a difference at the embedded verb was confirmed in Experiment 8. The new conditions were read significantly more slowly than the old condition. This difference followed from the DLT's predictions and most likely appeared on the word following the embedded verb as a result of reading time spillover from the embedded verb. The fact that there were no reading time differences at the embedded auxiliary or verb meant that the differences at the word after the embedded verb could not have been the result of spillover from differential processing on the subject of the RC. This finding was important, because the results of Experiment 7 could not rule out that possibility. The DLT's prediction of a difference at the matrix verb was not confirmed in Experiment 8. There was no difference in reading times at the main verb or the following word. The lack of a difference at the main verb in Experiment 8 is consistent with a few alternative hypotheses.

One possible explanation for the lack of a difference at the matrix verb is that the old referents in the items in Experiment 8 were not easy enough to access. Old referents can be easier or harder to access according to their prominence in the discourse model. The old referents in Experiment 8 were introduced in a non- prominent position: the agent of a by-phrase. As a result, the old condition may not have been much different than the new condition in terms of processing load. This explanation would predict that reading times at the verbs would have the pattern predicted by the DLT, but it would also predict weaker effects. Weaker effects would be more likely to be hidden by noise and make it more likely that predicted differences might not appear in the data. If this the case, then putting more focus on the old referent in the old conditions should cause a larger processing load difference between the new and old conditions, making it more likely that reading times at both verbs would show the pattern predicted by the DLT. Experiment 9 tests this hypothesis. Another possible account of the lack of difference found at the main verb in Experiment 8 was that it was a chance occurrence with no particular cause. If so, the lack of difference should not replicate in Experiment 9.

Whereas the finding of differences in reading times at the main verb in Experiment 4 was taken to rule out completely separate resource pools for structural and discourse processing, the lack of a difference in Experiment 8 could be taken as support for the alternative. This possibility will be addressed further in the conclusion.

2.6 Experiment 9

Experiment 9 was very similar to Experiment 8. In Experiment 9, the contexts from Experiment 8 were modified in order to put more prominence on the referent that was old in the old condition. This manipulation tested the hypothesis that the lack of difference in reading times at the main verb in Experiment 8 was due to the fact that accessing the old referent in the old condition may not have been significantly easier than introducing a new referent in the new conditions. Increasing the prominence of the old referent in context should make it more accessible and therefore increase the processing load difference between accessing the old referent and introducing a new referent.

2.6.1 Method

Participants

Forty-one native English speakers were recruited from the MIT community to participate in a word-by-word self-paced reading study. None of the participants had taken part in Experiment 9. Participants took approximately 20 minutes to complete the experiment. All participants also participated in an unrelated self-paced reading study during the same testing session, and spent approximately one hour in the lab, with a short break between the two sessions of self-paced reading. Participants were paid $10.

Materials

Experiment 9 tested 12 experimental items, with two conditions per item. The items were modified versions of the stimuli from Experiment 8. In this experiment, only the new / indefinite and old / definite conditions were tested, because they were the only felicitous conditions and because the new / definite condition in Experiment 8 had confirmed that differences between the conditions were due to changes in referent status rather than definiteness.

Each item in Experiment 9 consisted of two context sentences and a target sentence with a subject-modifying object-extracted non-restrictive RC. In the new condition (8a-c), the subject of both context sentences was a referent that did not appear in the target sentence. In the old condition (8d-f), the subject of both context sentences was the referent that appeared as the subject of the RC in the target sentence. The words that differ between conditions are underlined.

New / Indefinite

8a. A police chief was investigating a bank robbery.
8b. On Friday, he finally caught one of the suspects.
8c. The suspect, who a detective had sighted the previous Wednesday, struggled but was eventually subdued.

Old / Definite

8d. A detective was investigating a bank robbery.
8e. On Friday, he finally caught one of the suspects.
8f. The suspect, who the detective had sighted the previous Wednesday, struggled but was eventually subdued.

Care was taken to make the three-sentence discourses as felicitous and plausible as possible. The only words that differed between the conditions were the subject of the first context sentence and the determiner on the subject of the RC in the target sentence. The referent that served as the subject of the target sentence was always introduced after the verb in the second context sentence.

The target sentences in Experiment 9 were the same as those from Experiment 8, except that the phrase following the embedded verb was standardized to be three words in length.

The 12 experimental items in Experiment 9 were combined with 16 sentence triples from an unrelated experiment and 32 sets of filler sentences to form four lists. The two conditions were counterbalanced across lists and the lists were randomized. The complete set of items is presented in Appendix I.

Procedure

The procedure was the same as in Experiment 7, except that there were three sentences per item in this experiment. As in Experiment 7, every sentence began on a new line and the critical regions of the target sentence always appeared on the first line of the target sentence.

2.6.2 Results

Data from one participant who answered only 66% of the comprehension questions correctly across all fillers and experimental items was excluded. Every other participant answered at least 80% of the comprehension questions correctly across all fillers and experimental items. Table 2.6 presents the percentage of comprehension questions that were answered correctly for each condition in Experiment 9. There was no reliable difference in comprehension accuracy between conditions.

New - Indefinite	89.8
Old - Definite	90.2

Table 2.6. Percentage of questions answered correctly for each condition in Experiment 9

Reading times were analyzed on a word-by-word basis, except in one region. As in Experiment 8, one of the items had a RC subject that was made up of two words rather than one. The mean reading time for the subject of the RC includes average reading times for both of these words. Analysis was performed only over sentences for which the comprehension questions had been answered correctly. Reading times that were farther than 2.5 standard deviations from the mean for each position in each condition were trimmed, eliminating 2.8% of the remaining data.

Statistical tests were performed on both unadjusted reading times and reading times that were adjusted for word length following the procedure proposed by Ferreira and Clifton (1986) and Trueswell, Tanenhaus and Garnsey (1994). The analyses presented below are over unadjusted reading times, but all trends seen in the unadjusted reading times were also apparent in the length-adjusted reading times.

Figure 2.4 graphs the mean reading times for each word in Experiment 9.

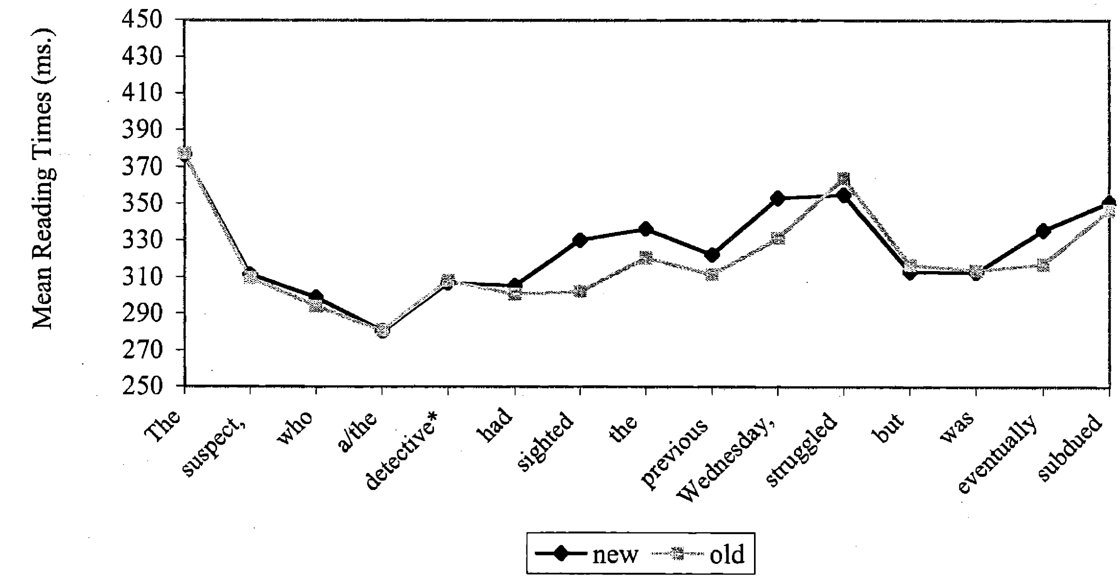

Figure 2.4. Mean reading times for each word position in Experiment 9

The mean unadjusted and length adjusted reading times for each word position in Experiment 9 are presented in Table 2.7.

	The	suspect,	who	the/a	detective *	had	sighted	the	previous	Wednesday,	struggled	but	was
new-	377	311	294	281	307	305	330	336	322	353	355	317	314
indef	(53)	(-47)	(-24)	(-26)	(-44)	(-17)	(-29)	(1)	(-12)	(-7)	(4)	(-11)	(-24)
old-def	377	310	299	281	308	301	302	320	312	331	364	313	312
	(52)	(-55)	(-29)	(-35)	(-40)	(-25)	(-60)	(-19)	(-26)	(-31)	(16)	(-8)	(-28)

Table 2.7. Average reading times (length-adjusted times in parentheses) for each word position in Experiment 9.

As in Experiment 8, the regions of interest were the embedded verb and matrix verb. The DLT predicted that reading times at both verbs would be longer in the new referent condition than in the old referent condition. Reading times at the embedded verb (e.g. "sighted") confirmed the

DLT's prediction, as the new / indefinite condition was read more slowly than the old / definite condition on this word (F1(1,39)= 10.72, MSe= 1457, p=.002; F2(1,11)= 4.00, MSe= 2354, p = .07). This difference does not quite reach the .05 level of significance in the items analysis over the raw reading times, but the difference is reliable by both subjects and items in an analysis over the length adjusted reading times (F1(1,39)= 12.55, MSe= 1469, p = .001; F2(1,11)= 10.05, MSe= 667, p = .009). The same trend of slower reading times on the new / indefinite condition carried over the next three words, which made up a phrase modifying the embedded verb (e.g. "the previous Wednesday"). As in Experiment 8, no differences were found at the matrix verb. In fact, on the matrix verb, "struggled", the old / definite condition was read numerically more slowly than the new / indefinite condition, contrary to the predictions of Gibson (1998) but similar to the finding in Experiment 8.

2.6.3 Discussion

The results of Experiment 9 were very similar to those of Experiment 8. In both experiments, reading times on the embedded verb or the word immediately following it were slower for the new / indefinite conditions than the old / definite conditions, as predicted by the DLT. Also, in both experiments, there were no differences in reading times at the main verb. Experiment 9 showed essentially identical reading times at the subject of the RC for the new and old conditions. This was contrary to findings reported in Garrod and Sanford (1994) and Garrod et al. (1994), where reading time effects of referential status were apparent on NPs.

The findings in Experiment 9 rule out the hypothesis that the lack of reading time differences at the main verb in Experiment 8was a result of the old referent being difficult to access. They leave open the question of why the differences predicted by the DLT appeared on the embedded verb but not the main verb in Experiments 8 and 9. Some possible explanations for this will be addressed in the general discussion.

2.7 General Discussion

On the whole, the results of the four experiments in this chapter confirmed the predictions of the DLT. Experiment 6 showed that building a representation for a new referent caused an increase in overall sentence complexity when the building took place between the endpoints of a long distance structural dependency, but not when the building took place outside of any structural dependencies. Experiment 7 tested the processing of referents which either were old to context, new to context or could be inferred into context by means of a bridging inference. At the right endpoint of structural dependencies crossing the referent whose status was varied, reading times were fastest for the condition where the referent was old, intermediate for the conditions where the referent could be inferred and slowest for the condition where the referent was new. Experiments 8 and 9 compared the reading time ramifications of accessing a referent from context using a definite description with the ramifications of introducing a new referent with an indefinite description. They found the pattern predicted by the DLT only at the embedded verb, not at the main verb.

There are a few possible accounts for the lack of reliable differences at the main verb in Experiments 8 and 9. The fact that Experiments 8 and 9 showed a different pattern of results from all of the other experiments in the thesis could be due to the fact that the RCs in Experiments 8 and 9 were non-restrictive, while the RCs in every other experiment were restrictive. It is possible that the processing of non-restrictive clauses differs from the processing of restrictive clauses in a way that interacted with the complexity effects at the main verb, masking any effect.

The experiments in this chapter manipulated multiple factors in order to change referential processing load. Experiment 6 showed that building new referential structure increases referential processing load even when no inferences are necessary to situate the new referent in context. Experiment 7 showed that making a contextually supported bridging inference results in a processing load that is intermediate between the loads caused by accessing a referent that is already in the discourse model and introducing a new referent without any contextual support. Experiments 8 and 9 showed that the processing load incurred by introducing a new referent in a supportive context with an indefinite description is greater than referring to a currently active referent in the discourse model with a definite description. None of these findings are surprising, but except for the finding in Experiments 8 and 9, which had previously been shown in Murphy (1984), none of these had been previously experimentally demonstrated.

The results in this chapter blurred the picture of resource allocation in the HSPM that had emerged from chapter 1. In the exposition of Experiment 4, it was argued that the prediction of differences on the main verb in the structures used in this thesis relied on a formulation of the DLT that assumed an overlap between the computational resources used for structural and referential processing. When differences were found at the main verb in Experiment 4, they were taken as evidence for overlapping resource pools. But the results of Experiments 8 and 9 conflicted with this evidence and showed no reading time differences at the main verbs. Until the cause of the null result at the main verb in Experiments 8 and 9 can be determined, the evidence concerning resource allocation will remain indeterminate. More investigation will be necessary to resolve this issue.

Whereas the experiments in chapter 1 identified and broadly described the phenomenon that became the version of the DLT's integration cost based on referential processing load, the experiments in chapter 2 delved more deeply into testing the specific predictions of the theory that had been developed in chapter 1. As more precise investigation was undertaken, the methodology of the experiments moved from complexity questionnaires to self-paced reading in order to provide more detail. This meant that every test sentence had multiple locations at which the DLT made specific predictions about reading time patterns. Future experiments will be necessary in order to determine why predicted differences were not found at the main verb in Experiments 8 and 9. Every experiment in chapter 2 showed reading time patterns that were consistent with the DLT's predictions in at least one of the two locations where they were predicted. The overall pattern of results in chapter 2 provides support for the DLT and for a formulation of integration cost that incorporates referential processing load in the distance metric.

Conclusion

3.1 Overview of Results

This thesis presented the results of nine experiments investigating the effects of referential processing on complexity. The experiments tested predictions of Gibson's (1998) DLT and motivated the inclusion of referential processing costs in the DLT's calculation of complexity. The finding in Experiment 1, that having indexical pronouns in the most embedded subject position of doubly nested sentences made them easier to understand, motivated Gibson's inclusion of referential processing in the original formulation of integration cost. The exposition of integration cost in Gibson (1998) included binary referential processing costs: one unit of cost was associated with introducing a new referent and no cost was associated with accessing an old referent. Experiment 2 provided evidence that this binary version of referential processing cost was too limited. In Experiment 2, six different types of referential forms were tested in doubly nested and right branching structures. The pattern of complexities in the doubly nested structures matched a Givenness hierarchy suggested by Gundel et al. (1993), supporting the hypothesis that referential cost was graded rather than binary and corresponded to the difficulty of accessing a referent from memory. Experiment 3 tested the DLT's prediction that increased referential processing at different subject positions in a doubly nested structure would cause different increases in overall complexity. Subject positions more internal to the structure interrupt more syntactic dependencies, so changes in the amount of referential processing in those positions should have a greater effect on overall complexity. Experiment 3 showed that this was the case, and also suggested that certain types of quantified NPs incurred no or little referential processing cost. Experiments 4 and 5 extended the findings of Experiments 1 and 2 to self-paced reading and showed that the complexity effects of changing the referential form of and thus the amount of referential processing required by a referent occurred at the word positions predicted by the DLT.

Because it incorporated findings from chapter 1 into the formulation of integration cost, the version of the DLT tested in chapter 2 was different from the one tested in chapter 1. In chapter 2 referential processing load was no longer apportioned in a binary way, but now was considered a graded cost that increased as referents became harder to access or to build. The experiments in chapter 2 differed from the experiments in chapter 1 because they used context sentences to vary referential status. Experiment 6 was a bridge between the experiments in chapter 1 and chapter 2, in that it included a context sentence, but still accomplished the manipulation of referential processing load by changing the referential form of an NP in the target sentence. Experiment 6 tested the complexity of doubly nested sentences where the most embedded subject was either a name from context or a referent anchored to that name by a possessive. The results of Experiment 6 showed that introducing a new referent, even when it was contextually anchored, caused a greater processing load than accessing an old referent. As predicted by the DLT this effect appeared in doubly nested structures but not in right branching structures. Experiment 7 tested reading time predictions of the DLT using singly nested sentences where the referential processing load of the embedded subject was manipulated in context. Four different levels of processing load were tested; the referent had either been previously explicitly introduced, was easily inferable in the context, was more difficult to infer in context or was completely new. The different processing loads incurred by these contexts were apparent in reading times at the locations predicted by the DLT. Experiments 8 and 9 tested reading times over sentences with a referent that had either been explicitly mentioned in context and was referred to with a definite or had not been introduced in the context and was referred to with an indefinite. The results of these experiments showed the DLT's predicted pattern of reading times in one of the two positions where the DLT predicted it would appear.

The experiments in this thesis investigated the effects of different types of referential manipulations on complexity. Their results supported a version of the DLT where increased referential processing causes increased integration cost. Results from these experiments also provided evidence concerning: 1) possible mechanisms underlying accessibility differences and 2) the organization of resource stores. This conclusion will discuss the contributions of this thesis to these two issues, before summarizing and evaluating two new contributions to the literature that test claims in this thesis. It will conclude with suggestions for future directions of this work.

3.2 The Mechanisms Underlying Accessibility

Evidence presented in this thesis suggested that increased referential processing early in a sentence has ramifications on complexity at specific points later in the sentence. The experiments in this thesis manipulated the level of referential processing in multiple ways. Some experiments compared new and old referents directly. Accessing old referents was hypothesized to require less referential processing than introducing new referents. Further experiments focused on costs associated with accessing old referents at multiple levels of accessibility, while others varied the amount of difficulty associated with introducing new referents.

The complexity effects demonstrated in these experiments raise the question of what properties of the referential or discourse processing system cause some referential processes to require more resources than others. Rather than attempting to explain the root mechanisms of accessibility differences, most linguistic and psycholinguistic research has focused on describing the phenomenon of accessibility through a collection of diagnostics. Experiments and corpus analyses have shown that an entity is more accessible if: 1) it has been mentioned more recently (e.g. Ariel, 1990; Clark & Sengul, 1979), 2) it is in a syntactically focused or syntactically available position (e.g. Almor, 1999; Ariel, 1990; Arnold, 1998; McKoon et al., 1993), 3) it is introduced by a proper name rather than by a description (Sanford, Moar & Garrod, 1988), 4) it is a subject rather than an object or a goal rather than a source (Arnold, 1998), or 5) it is more fully elaborated or has more modifiers (Myers & O'Brien, 1998). This list of diagnostics is only partial, but it demonstrates that accessibility is affected by factors associated with a wide range of causes, from syntax to event structure to communicative intent. The linguistic form of NPs correlate with these diagnostics, and thus with accessibility (see e.g. Givon, 1983; Ariel, 1988, 1990; Gundel et al., 1993). Gundel et al. suggested that the correlation between accessibility and NP type occurs because the different types of NPs are used to indicate that referents have different cognitive statuses. In Experiment 2, an activation-based hypothesis based on Gundel et al. (1993) was suggested as a reason for the correlation between referential form, status in the Givenness hierarchy, and accessibility. This activation-based hypothesis was founded in observations about properties of referential forms from the Givenness hierarchy, which suggested that some forms refer only to very active and accessible referents while other forms have fewer restrictions on the activation and accessibility of their referents. This hypothesis accounted for the data in Experiment 2, but leaves open questions about the mechanisms of the referent access process and the structure of the discourse model.

3.2.1 The Resonance Model: Myers and O'Brien (1998)

Myers and O'Brien (1998) present an overview of the resonance model, a theory of the mechanism underlying referent access. The resonance model was presented as an alternative to a search process involving conscious effort on the part of the reader. It hypothesizes that new linguistic input and the contents of working memory act as constant signals to the rest of the memory system. As the contents of working memory change due to the processing of new linguistic input, the signal being sent to the memory system changes. The signal works through a simple pattern-matching process of activation. Information in memory that shares more conceptual features with or appears in more propositions with the signal resonates more strongly. The contents of the discourse model have a resonance advantage over the contents of general semantic memory, but elements from either can be activated through resonation. The elements that resonate most strongly become most active and enter working memory, whether or not they will be relevant in further processing. A subsequent process checks the contents of working memory to verify that all of the information required to process the text has been activated, for example that each anaphor has been assigned a single antecedent, and that all of the information in working memory is internally consistent. Much of the research supporting the resonance model has been concerned with determining how the accessibility of high level concepts, such as goals, motivations or causes, change with discourse manipulations (e.g. Albrecht & Myers, 1995; Rizzella & O'Brien, 1996). The evidence that Myers and O'Brien cite as supporting the resonance model over a directed search model involves assumptions about the assignment of motivation to characters in a passage; it is not clear that such evidence is relevant to the process of finding a referent for a NP. Still, when considering the process of anaphor resolution under normal circumstances, the resonance model accounts for some of the properties of accessibility in a natural way. Myers and O'Brien implemented some of the basic features of the resonance model in a neural network and found it successfully modeled a subset of the factors that have been shown to affect accessibility.

The relevant question for this thesis is how the functioning of the resonance process affects computational resource usage. Myers and O'Brien (1998) and Greene, McKoon and Ratcliff (1992), who present a similar conception of the referent activation process, assume that the resonance process is automatic, constant and undirected, like a perceptual process. If this is the case, then it may be a self-contained system and not use resources that are available to other language subsystems. This would pose a problem for the referential processing theory proposed to account for the results of Experiments 2 and 5, which relied on the assumption that accessing a less-active referent increased processing load in the referential processing system. But even if resonance itself has a dedicated system, the post-access processes that are necessary for verifying that the elements returned by resonance are correct and consistent must be general referential processing systems. It is possible that the processing load differences attributed to referent access in Experiments 2 and 5 could be due to differences in the amount of processing required to check through the information output by resonation and to select only the correct referent and relevant information. It would require another thesis to develop and test such a theory, but one might predict that the amount of output from the resonance system depended on the type of the NP used as a cue, because of differences in the amount of features on different types of NPs. Such a theory might argue that indexical pronouns have very few semantic features and plausibly activate only one referent, 3rd person pronouns possibly activate a few low-threshold referents, famous names have more features and thus activate a single referent as well as extraneous information, and definite descriptions have many features and activate multiple referents and extra information. Checking through more output could cause the post-access processes to require more resources, driving up processing load. Another possibility is that the resonance process draws on resources that are used by other referential processing systems, and the time the resonance process requires to activate a referent or set of referents in response to a cue correlates with the amount of resources the process consumes. This theory is one possible implementation of the hypothesis suggested in Experiment 2.

The on-line processing demands of accessing and assigning referents to their antecedents were left underdetermined in the presentation of the resonance model in Myers and O'Brien (1998). The results of Experiment 2 suggested that the processing demands of referential access for an NP are related to the lowest possible referent activation level that could be referred to with the NP. These findings provide information about the mechanism and on-line resource demands of referent access, but a full theory of referential processing will also require an account of the structure of the discourse model and an explanation for why certain referents are more active than others.

3.2.2 The Focus Memory Framework: Garrod, Freudenthal and Boyle (1994)

There are few well-elaborated theories of the structure of discourse models. One of the most detailed is the Focus Memory Framework (FMF) introduced in Sanford and Garrod (1981) and extended in Garrod et al. (1994). The FMF is consistent with an underlying system of activation like the resonance model. The FMF hypothesizes that there are two partitions in working memory used for discourse processing. Explicit focus contains a representation of the referents that have been active in the discourse. Implicit focus contains the background information that is activated by the scenario that relates the referents. Explicit and implicit focus are separate partitions, but they are related by mappings between referent tokens in explicit focus and "role slots" in implicit focus. Referent tokens are assumed to include individuating information about different referents, while role slots contain more transient information such as a referent's current role, location and state in the scenario. Referents that are more important in the current discourse are referenced more often and have more links to implicit focus than background referents, making them more likely to resonate in response to a wider set of signals.

Referents with few or weak connections between explicit and implicit focus are more difficult to activate, and as the scenario changes, referents that are no longer relevant to the current scenario and have no links to implicit focus are eliminated from the representation.

Garrod et al. (1994) hypothesized that pronouns directly access referent tokens in explicit focus. Descriptions and proper names, on the other hand, introduce temporary referent tokens in explicit focus. These temporary tokens are necessary because descriptions and names are often used to introduce new referents or to reactivate referents that are no longer represented in the focus system. In the case of definite descriptions and names, a subsequent process attempts to map the new referent token to an extant role in implicit focus. If there is already a referent token that maps to the relevant role in implicit focus, the temporary token and the already extant referent token are unified. For indefinite descriptions, a new referent token is introduced in explicit focus and a new role is introduced in implicit focus.

3.2.3 A More Comprehensive Combination

The FMF and the resonance model are consistent with the results reported in this thesis. Combining the two theories produces a theory with a possible explanation of the referent activation process, a detailed account of the properties of the discourse representation and an account of how activation spreads beyond the discourse representation in order to access entities that are no longer in the discourse representation. According to this combined theory, pronouns should be easily processed because they directly access the referents that resonate most easily because they have the most links in the discourse model. First names usually indicate referents that are important in the discourse model, so they should also have many links between explicit and implicit focus and thus resonate easily. Garrod et al. (1994) suggested that definite descriptions and full names cause more difficulty because they require extra processes to instantiate a temporary token and check for redundancy. But names and definite descriptions must also sometimes spur a search of long term memory, because they can be used to refer to referents that had once been in the discourse model but are no longer relevant to the current scenario as well as to referents stored in long term memory. The FMF says nothing about referent access beyond the discourse model. But the resonance model suggests that information from all of memory can be accessed. The FMF does not predict a difference between full names and definite descriptions, but it is possible the difference between names and definite descriptions could result from differences in their statuses as resonance cues.

This overarching theory is consistent with the results of the experiments in this thesis. The resonance model provides a mechanism for referent access, but leaves unexplained vital processes such as the secondary processes that winnow the information returned through resonation. The FMF provides an account of the structure of the discourse model, but does not extend the account to referents stored in long term memory and does not explain differences between names and descriptions. Together, they provide the basis for an account of referential processing. The findings in this thesis contribute to this account by identifying processing costs that are associated with referential access processes, even in cases where referent access failed.

The properties of accessibility were used in this thesis in order to demonstrate that the resources used for referential processing at particular locations in a sentence had ramifications on the complexity of subsequent processing. The results of the accessibility manipulations also provided new evidence relevant to the mechanisms underlying accessibility.

3.3 Evaluating The Evidence Concerning Resource Stores

Another issue that the experiments in this thesis provided evidence about was the resource usage of different subprocesses in the HSPM. In his (1998) presentation of the DLT, Gibson formulated integration cost in a way consistent with a single resource pool or with overlapping pools for syntactic and discourse processes. The resource demands of referential processing were predicted to have an effect on the amount of resources available for syntactic structure building. But Gibson could have formulated integration cost so that it treated the resource use of syntactic and discourse processing separately. For example, such a version of integration cost could have been expressed as:

Separate syntactic and discourse DLT linguistic integration costs:

Syntactic integration cost: The structural integration cost associated with connecting the syntactic structure for a newly input head h_2 to the projection of a head h_1 that is part of the current structure for the input is dependent on the complexity of the syntactic computations that took place between h_1 and h_2.

Discourse integration cost: The discourse integration cost associated with accessing previous discourse structure r_1 is dependent on the complexity of the discourse computations that took place between the last access of r_1 and the current access of r_1.

Such a formulation of integration cost would have been consistent with all of the data presented in Gibson (1998). But reading time evidence in Experiment 4 was relevant to deciding the issue. As discussed previously in the exposition of Experiment 4, the single resource pool formulation of integration cost predicts that effects of changing the discourse status of a referent in a nested clause will be apparent at the endpoints of dependencies in the matrix clause that cross the nested clause. For example, in (1), introducing a previous context that makes the congressman accessible is predicted to cause faster reading times at both the embedded verb "visited" and the matrix verb "found," as opposed to the case when the congressman is less accessible.

1) The judge who the congressman had recently visited in the courthouse found the corporation innocent.

This prediction results because any change in the amount of discourse processing between the endpoints in a dependency, whether the extra processing be in the same clause or a different clause, is predicted to change the amount of resources available for completing the syntactic dependency. Reading times in Experiment 4 showed exactly this pattern. Reading times at the matrix verb were slower when a referent in the nested clause was difficult to process than when a referent in the nested clause was easily processed. In the discussion of chapter 1, this finding was described as not being consistent with a version of integration cost that allowed no overlap between the resources accessed by syntactic processes and discourse processes. According to a formulation of integration cost that separates discourse processing resources from syntactic processing resources, like the one presented above, the amount of referential processing necessary for an argument in a nested clause should not affect reading times on the matrix verb. This follows because all of the arguments of the embedded verb are accessed so that they can be related into a proposition at the embedded verb. For example, in sentence (1), the referents for the judge and congressman will be accessed at "visited" in order that the proposition "the congressman visited the judge" can be created. The argument that is shared by both clauses ("judge" in (1)) must also be accessed at the matrix verb, but its most recent access was at the embedded verb. According to the definition above, only a difference in the amount of discourse processing between the embedded verb and matrix verb will appear in reading times on the matrix verb. If the ease of accessing "in the courthouse" above does not differ between conditions, then no differences should be apparent at the matrix verb.

While Experiment 4 showed a difference at the main verb, consistent with a formulation of integration cost where discourse and syntactic processes tap the same resources, Experiments 8 and 9 showed no difference at the main verb. The question, then, becomes how to account for the difference between Experiment 4 and Experiments 8 and 9. One possible account is that the lack of a difference at the main verbs in Experiments 8 and 9 was simply an uninterpretable null result, while the differences evident in Experiments 4 were real. This account is possible but somewhat unlikely, because predicted differences were found at the embedded verbs in Experiments 8 and 9, showing that the experimental manipulation was successful and that experiment participants were sensitive to it. Also, the null result of Experiment 8 was replicated in Experiment 9. Another possible account is based on the difference between the embedded clauses in Experiment 4 and Experiments 8 and 9. The embedded RCs in Experiment 4 were restrictive, while the embedded RCs in Experiments 8 and 9 were non-restrictive. It is possible that a discourse constraint associated with non-restrictive RCs may have led to longer reading times at the end of RCs containing new referents than at the end of RCs containing old referents, and this effect may have spilled over and canceled out the DLT's predicted effect at the main verb. This account is also unlikely, for almost any discourse constraint that could be hypothesized to slow reading times for old referents as opposed to new referents, for example increased relevance, would most likely slow reading times over the entire clause, not just its last word. Experiments 8 and 9 show no evidence of slower reading times over the embedded clause in the old referent condition than in the new referent condition.

The evidence in this thesis relating to resource allotment is mixed. One experiment shows a pattern that would be expected if syntactic and discourse processes accessed the same resources, while two experiments show a pattern that would be expected if they accessed different resource pools. In order to settle the question, more directed experiments will be needed.

3.4 Responses from the Literature

The work in this thesis has already provoked responses in the literature. The results of Experiment 1 were described in Gibson (1998) and spurred a response by Peter Gordon and his research group. Gordon, Hendrick & Johnson (2001) argued against the version of the DLT from Gibson (1998), which has been modified and expanded by the work in this thesis. Some of the results that they describe are compatible with the version of the DLT that emerged from the work in this thesis, but some provide a challenge for the DLT. Julie Van Dyke presented a poster at the 2001 CUNY conference with an experiment that she claimed showed longer reading times for dependencies crossing old referents than new referents. These potential challenges to the DLT and the claims in this thesis will be discussed below.

3.4.1 Gordon, Hendrick & Johnson (2001)

In a paper currently in press, Gordon, Hendrick & Johnson disputed the finding in Experiment 1 that nested sentences with short names as the most embedded subject were judged similarly complex to the same sentences with definite descriptions in the most embedded subject position. In two experiments directly testing this finding, Gordon et al. compared the processing of singly nested object and subject extracted RCs, varying the lexical NP in the RC between either the indexical pronoun "you" and a definite description, as exemplified in (2), or a three-letter name and a definite description.

2a) The barber that the lawyer/ you admired climbed the mountain.
2b) The barber that admired the lawyer/ you climbed the mountain.

These experiments were self-paced reading experiments in which each word in a sentence appeared in the center of a computer screen until a button press replaced it with the subsequent word. In both experiments, Gordon et al. found that the condition with a description as the subject of an object-extracted RC had significantly longer reading times at the verbs than the condition with a pronoun or name as the subject of the object-extracted RC. Reading times in the subject-extracted RC conditions did not show as large an effect of referential type variation. Gordon et al. concluded that these data were inconsistent with a theory like the DLT, which explains differences in reading times using a referent-based distance metric. The reasoning behind this conclusion was that in these experiments, names and definite descriptions both introduced new discourse referents but caused different reading time patterns.

But though Gordon et al.'s data from these two experiments were not consistent with the formulation of integration cost from Gibson (1998), they are consistent with the formulation of integration cost suggested in this thesis. Though Gordon et al. did not compare names and pronouns within a single experiment, their stimuli were matched across the two experiments in such a way that directly comparing the name and pronoun conditions was possible. They found that average reading times on the critical words in the name condition were 190 ms slower than average reading times on the same words in the indexical pronoun condition. These data are consistent, though the magnitudes of the differences are bigger, with the findings from

Experiments 2 and 5 in this thesis, where names were intermediate between descriptions and pronouns in the amount of complexity they contributed to the processing of dependencies that crossed them. Though Gordon et al. framed the results of these experiments as disproving referent-based complexity metrics such as integration cost in Gibson's (1998) DLT, they only provided evidence against the simplified binary metric based on incrementing cost for new referents but not old referents. The more elaborated metric based on referent accessibility that is argued for in this thesis not only withstands, but finds support in, their results.

In another experiment, Gordon et al. tested the effects of referring to referents with either names or definite descriptions in subject- and object-extracted clefts. Clefts are similar to RCs in that they have a dependency between a relative pronoun and a gap position, but they are different than RCs in that the cleft head can be a wider range of referential type. This property of clefts allowed Gordon et al. to further test their hypothesis that complex structures with multiple NPs of the same type are more difficult to process than complex structures with more varied NP types. Gordon et al. gathered word-by-word reading times over subject- and object-extracted clefts, and varied the cleft head and the lexical NP in the embedded clause independently between names and definite descriptions. These variations defined eight conditions. An example of one of their stimuli is shown in (3a) in the object-extracted condition and in (3b) in the subject-extracted condition:

3a. It was the editor/Pam that the author/Jen recommended after a new merger was announced.
3b. It was the editor/Pam that recommended the author/Jen after a new merger was announced.

The DLT predicts that reading times in these structures will vary depending on the amount of referential processing that intervenes between the relative pronoun "that" and its gap site. Names are predicted to require less referential processing than descriptions. Because the cleft head is processed before the beginning point of any long distance dependencies, the DLT predicts that varying the cleft head will have no effect on reading times at "recommended." In the subject-extracted condition, the lexical NP in the embedded clause is in object position and does not intervene between any long distance dependencies, so the DLT predicts that changing its referential type will not affect reading times at "recommended" either. The DLT subsequently predicts that there will be no differences among any of the four subject-extracted cleft conditions. In the object-extracted clefts, the DLT predicts a difference in reading times at the verb "recommended" depending on whether the subject of the embedded clause is a name or a description. The DLT predicts that reading times will be longer at "recommended" when the subject of the embedded clause is a description and shorter when it is a name, in the object-extracted cleft conditions.

Gordon et al. report reading times only for the verb "recommended." In the subject-extracted cleft conditions, there were no differences in reading times among any of the conditions, as predicted by the DLT. In the object-extracted cleft conditions, Gordon et al. found a pattern not predicted by the DLT, namely that reading times were slower when both NPs were names or descriptions and faster when one NP was a name and the other was a description. Gordon et al. interpret these results as providing evidence against the DLT and for a theory attributing more complexity to sentences where a particular category of NP is repeated and less complexity to sentences with NPs from different categories.

If the results in Gordon et al. are correct, then this finding poses a considerable problem for the DLT. But there are indications that reading times in Gordon et al.'s studies may have been confounded by memory strategies unrelated to normal reading. First of all, Gordon et al.'s reading times were unusually long. They found average per-word reading times ranging between 500 ms and 1000 ms in every experiment, including a preliminary study holding referential form constant but comparing the processing of subject and object-extracted RCs. The materials in this preliminary study were similar to those used by King and Just (1991), who reported reading times for low-memory-span participants performing an additional memory task during reading that were similar to those found by Gordon et al.. In an independent experiment testing the same structures, Gibson and Ko (1998) found average per-word reading times ranging between 300ms and 450 ms, well within the range of reading times found in the experiments in this thesis.

Secondly, in every experiment Gordon et al. had a low ratio of fillers to experimental items and asked predictable comprehension questions. Gordon et al.'s participants were presented with blocks of sentences in which there were ten filler sentences and eight target sentences, four of which contained subject-extracted RCs or clefts and four of which contained object-extracted RCs or clefts. This means that in 44% of trials in the RC experiments participants read simple sentences containing RCs. In these trials, participants were always questioned about the first two NPs in the sentence and one of the two verbs. In the cleft experiment, the questions for the 44% of trials which contained target items were even more predictable, since there were only two NPs and one verb to be queried. With such a high percentage of target items and such predictable questions, it is possible that participants recognized the target structures and developed strategies for remembering the relationships between the NPs in the RCs and clefts.

If participants in the Gordon et al. experiments were using a memory strategy, reading times should have reflected the strategy. A simple strategy participants may have used was to pause and rehearse the actors and events from the sentence's first proposition before reading the remainder of the sentence. In the experimental sentences the last word in the RC or cleft was the last word necessary to build the proposition that was usually tested in the comprehension questions. If readers were pausing to consolidate propositions in memory in anticipation of the comprehension questions, then reading times should have been high on the last word in the RC. This pattern was found in the preliminary experiment where Gordon et al. compared simple subject and object-extracted RCs. Gordon et al. found an increase in reading times on the last word in the subject-extracted RC conditions, while Gibson and Ko, who had a higher ratio of fillers to target items and less predictable comprehension questions, found no increase in reading times on that word. Both experiments found an increase in reading times on the last word in the object-extracted RCs, but this increase is predicted independently by Gibson's (1998) DLT.

Two of the experiments in Gordon et al.'s paper provide converging evidence for the version of the DLT argued for in chapter 1. Their cleft experiment provides an interesting challenge to the DLT. But the methods and experimental procedure used in all of the experiments may have led participants to develop particular reading strategies that could have affected their reading times. For this reason, it will be important to replicate the experiments from Gordon et al. (2001) with more fillers and less predictable comprehension questions before accepting their results as the reflections of normal reading processes.

3.4.2 Van Dyke (2001)

In a poster presented at the fourteenth annual CUNY conference, Van Dyke (2001) reported an experiment testing reading times over a sentence containing a long distance syntactic dependency crossing either an old referent or a new referent. The structures she tested were like the following; they differed in whether an indexical pronoun, i.e "you", a 3rd person pronoun, i.e. "him", or a new definite description, i.e. "the dock" was located inside the embedded clause:

4. The older boy realized that the girl who was swimming next to {him/ you/ the dock} was paranoid about dying.

The two pronouns refer to entities that are old to the discourse, while the definite description refers to a new entity. Van Dyke found that reading times were significantly longer over the most embedded clause in the pronoun conditions than in the definite description condition. She argued that this finding was not consistent with the DLT's predictions.

There are a few reasons why this conclusion is premature. For one, the pronoun conditions require an additional predication in the embedded clause as compared to the definite description conditions. The RC “who was swimming next to {him/ you}” presupposes the information that *you* or *the boy* was swimming. The RC “who was swimming next to the dock” introduces no such additional predication. Since the DLT predicts that integrations crossing more discourse processing will be more difficult, and an additional predication will cause more discourse processing, the DLT predicts Van Dyke’s finding. Still, Van Dyke’s results should be regarded as preliminary, because she did not control for possible plausibility differences between conditions.

3.5 Future Directions

The experiments in this thesis have been integral to the development of the DLT. But in no way have they completed the work of developing and testing the DLT. The following paragraphs offer some suggestions as to future directions for this work.

3.5.1 Extending the Methodology

All of the experiments in this thesis were conducted either as complexity questionnaires or as self-paced reading experiments. Complexity questionnaires were effective in testing broad predictions of the DLT, but all of the detailed testing relied on self-paced reading data. The reading time supported the DLT over a more general theory predicting that complexity ramifications of referential processing will be localized at the few words following the extra referential processing. Reading times in Experiments 5 and 7 could not distinguish between these two theories because of possible spillover effects, but Experiments 4, 8 and 9 provided evidence that complexity differences found at the verbs were not simply spillover from the point of extra referential processing. Experiments 8 and 9 showed differences on the embedded verb, even though there were no differences over the auxiliary and adjective intervening between the point of additional referential processing and the verb. Experiment 4 showed differences at the main verb, even though there were no differences over the words intervening between the two verbs. This evidence suggests that the pattern of complexity that has been studied in this thesis is not a general amorphous increase in complexity, but it is discrete complexity increases at predicted locations.

Still, not all of the differences that the DLT predicted were found in every experiment. Self-paced reading is an artificial task and reading times can be affected by a wide range of factors unrelated to normal reading, such as fatigue, motivation, question-answering strategies, and issues related to button pressing. Other methods, such as eye-tracking, are more natural and arguably introduce less noise into reading data. It would be worthwhile to replicate some of the experiments from this thesis using an eye-tracking methodology to see whether the predicted effects are found more consistently than in the self-paced reading data.

3.5.2 Extending the Structures Tested

All of the experiments in this thesis directly tested the predictions of the DLT. But they all tested processing in structures with subject-modifying object-extracted RCs. This restriction of experimental structures was intentional, because the aim of these experiments was to identify and fully describe the effects of changing referential processing demands on complexity. The experiments in this thesis varied referential factors but held syntactic factors essentially constant. Restricting study to one or two syntactic structures allowed a quicker, deeper investigation into the particulars of the referential phenomena than would have been possible if multiple structures had been investigated. But restricting study to a single structure also had the disadvantage that some of the effects attributed to a general cause could be the result of a peculiarity of the structure rather than a general property. For these reasons, while it was useful to restrict the structures investigated in this thesis in order to learn as much about the effects of different referential processing manipulations, the next step in the development of this theory will be to test it on as many different structures as possible.

3.5.3 Further Investigating the Processing of Quantified NPs

Many of the experiments in this thesis made use of the fact that different referential forms are used to refer to referents with different accessibilities. Experiment 3 used quantified pronouns as low discourse cost NPs to test the DLT's predictions about processing high and low discourse cost NPs in positions interrupting different numbers of syntactic dependencies. But the fact that the experiment was successful raised some interesting issues about the processing of quantified pronouns. Quantified pronouns could have low discourse costs for two reasons: 1) because they are pronouns or 2) because they are quantified. This is an interesting question to pursue, because it bears on important issues in formal linguistics as well as on accessibility.

3.6 Summing Up

This thesis has played an integral part in the development and testing of a new theory of linguistic complexity. The experiments in this thesis provided new evidence about the complexity ramifications of referential processing. As a result of this evidence, DLT has been expanded to predict complexity effects based on both syntactic and referential processing costs. This work has also provided evidence concerning the mechanisms underlying referent accessibility and the allocation of resources to different subprocesses of the HSPM. This thesis has provided a broad base on which future work investigating the complexity ramifications of discourse and syntactic processing can build.

Appendix A Experimental items for Experiment 1

The items used in Experiment 1 are listed below. Individual conditions used one of the four NPs separated by slashes.

1. The student who the professor who I/ they/ Jen/ the scientist collaborated with had advised copied the article.
2. The hockey player who the fans who I/ she/ Ed/ the sports writer ridiculed cheered for scored a goal to win the game.
3. The Republicans who the senator who I/ she/ Ann/ the citizens voted for chastised were trying to cut all benefits for veterans.
4. The warden who the prisoners who I/ she/ Sarah/ the social worker visited hated didn't give the inmates enough food.
5. The librarian who the child who I/ he/ Dave/ the teenager babysat for respected gave many presentations to elementary school classes.
6. The singer who the pianist who I/ they/ Rob/ the manager admired toured with had a beautiful voice.
7. The inventor who the lawyer who I/ they/ Beth/ the patent office disliked swindled lost the $2000 he invested in his latest invention.
8. The administrator who the nurse who you/ they/ Ed/ the doctor supervised had fired lost the medical reports.
9. The umpire who the baseball player who you/ she/ Ted/ the little boy liked had threatened sent a letter of protest to the commissioner.
10. The politician who the actor who you, they/ Sam/ the critic admired had supported gave a moving speech.
11. The nanny who the agency who you/ he/ Al/ the neighbors recommended sent was adored by all the children.
12. The engineers who the technicians who you / he/ Joe/ the students lived with worked for were making important discoveries.
13. The boy who the teacher who you/ she/ Pat/ the principal respected had disciplined was arrested again last week.
14. The inmate who the judge who we/ she/ Ben/ the Democrats voted for convicted escaped from prison last week.
15. The prophet who the religious leader who we/ they/ Mary/ the journalist distrusted worshiped proclaimed the end of the world was near.
16. The defendant who the attorney who we/ she/ Barb/ the consultant hired questioned incriminated himself several times during the course of the trial.
17. The sergeant who the lieutenant who we/ they/ Ron/ the governor met reported to was hard on all the new recruits.
18. The actress who the director who we/ he/ Mike/ the studio admired seduced won an Emmy award for her portrayal of a homeless woman.
19. The immigrant who the con-man who we/ she/ Liz/ the children saw on the news had tricked went to the police immediately.
20. The producers who the playwrights who we/ he/ Jim/ the community supported ignored wanted to make some last minute changes to the script.

Appendix B Experimental items for Experiment 2

The items used in Experiment 2 are listed below. Only the nested versions of the items are given, except for the first item. The right branching conditions can be computed by extracting the innermost clause, following it by the middle clause and finishing with the outermost clause. Individual conditions contained one of the five NPs separated by slashes. The 3rd person pronoun conditions were created by adding "According to *name*," to the beginning of the item, and replacing the varied NP with the gender-appropriate pronoun.

1a. The writer who the professor who I/ the reporter/ a reporter/ CBS/ Ann talked to disliked had written many radical articles.
1b. I/ the reporter/ a reporter/ CBS/ Ann talked to the professor who disliked the writer who had written many radical articles.
2. The old professor who the students who I/ the scientist/ a scientist/ Stephen Hawking/ Jane lectured to liked told interesting stories.
3. The schoolboy who the cartoon character which I/ the cartoonist/ a cartoonist/ Walt Disney/ Susan created fascinated was impatient to see the new episode.
4. The programmers who the company which I/ the accountant/ an accountant/ Steve Jobs/ Alan worked for had competed with invented a new way to design web pages.
5. The assistant who the lawyer who I/ the partner/ a partner/ Marcia Clark/ Alice talked to had hired repeatedly mistook the addresses on the letters.
6. The salesman who the woman who I/ the company/ a company/ Microsoft/ Bob hired dealt with was very polite.
7. The man who the drag queen who I/ the dancer/ a dancer/ Dennis Rodman/ Mike recognized had accompanied was wearing lizard skin boots.
8. The congressman who the passers-by who I/ the bicyclist/ a bicyclist/ Newt Gingrich/ Polly avoided had recognized was catching a taxicab.
9. The hairdresser who the image consultant who I/ the rock star/ a rock star/ Madonna/ Kathy visited had recommended has cut hair in Hollywood for thirty years.
10. The actress who the producer who I/ the manager/ a manager/ Howard Stern/ Tom had talked with liked wasn't very good.
11. The artist who the interior decorator who you/ the heiress/ an heiress/ Martha Stewart/ Dan had asked about worked with painted beautiful portraits.
12. The child who the little girl who you/ the neighbor/ a neighbor/ Santa Claus/ Carl had surprised played with had a cast on her arm.
13. The radio station which the man who you/ the politician/ a politician/ NPR/ Mary sued had worked for went out of business.
14. The model who the photographer who you/ the designer/ a designer/ Versace/ Sam had hired took pictures of was very famous.
15. The old lady who the government assistance program which you/ the reporter/ a reporter/ Bill Clinton/ Brad praised had saved did not have enough money to heat her house.
16. The surgeon who the old man who you/ the insurance company/ an insurance company/ Medicaid/ Jane subsidized requested performed the operation.
17. The movie star who the trainer who you/ the weight lifter/ a weight lifter/ Arnold Schwartzenegger/ Alex worked with worshiped had made a new movie.
18. The food critic who the waiter who you/ the food critic/ a food critic/ Julia Child/ Kate ordered from served was a reviewer for the Michelin guide to restaurants.
19. The nurse who the doctor who you/ the basketball player/ a basketball player/ Michael Jordan/ Fred consulted had called arrived promptly.

20. The housewife who the secretary who you/ the phone company/ a phone company/ Bell Atlantic/ George employed had contacted was late paying her bill.
21. The consultant who the business analyst who we/ the phone company/ a phone company/ Bell Atlantic/ George argued with called did not fix the problem.
22. The pianist who the conductor who we/ the famous violinist/ a famous violinist/ Andrew Lloyd Webber/ Frank respected chose had won several international competitions.
23. The TV show which the teenagers who we/ the comedian/ a comedian/ Jerry Seinfeld/ Jenn entertained loved was going to be cancelled next season.
24. The librarian who the salesperson who we/ the book store/ a book store/ Barnes and Noble/ Jim had employed consulted recommended some great new books.
25. The agent who the celebrity who we/ the official/ an official/ Ted Kennedy/ Jane had invited hired contacted the travel agency.
26. The circus performer who the audience who we/ the magician/ a magician/ Houdini/ Ellen entertained watched did some new tricks during the break.
27. The family who the benefit which we/ the celebrity/ a celebrity/ Dolly Parton/ Peter sang at helped had a baby with bone cancer.
28. The boy who the chess player who I/ the master/ a master/ Kasparov/ Judy almost lost to taught became very good at chess.
29. The woman who the poet who we/ the playwright/ a playwright/ Shakespeare/ Jeff inspired was seeing liked artistic men.
30. The bohemian artist who the woman who we/ the musician/ a musician/ Bob Dylan/ Jennifer sang with dated could not pay his bills.
31. The boxing fan who the coach who I/ the boxer/ a boxer/ Mike Tyson/ John trained with shouted at was sitting too close to the ring.
32. The girl who the psychologist who I/ the neurotic man/ a neurotic man/ Woody Allen/ Ken met with counseled hated her movie star mother.
33. The schoolteacher who the character who I/ the actor/ an actor/ Robin Williams/ Matthew played laughed at was not very open minded.
34. The illustrator who the author who we/ the talk show host/ a talk show host/ Oprah/ Kelly had interviewed praised had done a fantastic job.
35. The golfer who the amateur who I/ the golf pro/ a golf pro/ Tiger Woods/ Josh beat had learned from had a cart with a racing stripe.
36. The judge who the lawyer who I/ the criminal/ a criminal/ OJ Simpson/ Judith hired liked was usually sympathetic to the defense.

Appendix C Experimental items for Experiment 3

The items used in Experiment 3 are listed below. All six conditions are given for the first item, with only the doubly nested - quantified NP in the innermost subject position condition given for the rest. The other conditions can be computed as described in the materials section of Experiment 3.

1a. Everyone who the journalist who the photographer met liked was at the party.
1b. The photographer met the journalist, and everyone who the journalist liked was at the party.
1c. The photographer who everyone who the journalist met liked was at the party.
1d. The photographer was at the party, and everyone who the journalist met liked the photographer.
1e. The journalist who the photographer who everyone met liked was at the party.
1f. Everyone met the photographer, and the journalist who the photographer liked was at the party.
2. The taxi driver who the tourist who everyone honked at cut off hated driving in downtown Boston.
3. The stranger who the beautiful girl who everyone warned stayed away from had ties to the Russian mafia.
4. The hairdresser who the manicurist who everyone gossips with talks about is either pregnant or gaining lots of weight.
5. The activist who the guest of honor who everyone greeted admired was having a wonderful time at the fund raiser.
6. The mechanic who the dealer who everybody consulted recommended had twenty years of experience with Toyotas.
7. The cheerleader who the football star who everybody had a crush on was friends with was very popular around school.
8. The engineer who the technician who everybody ran into praised had gone to school at MIT.
9. The manager who the associate who everybody disliked kissed up to had a lot of power in the company.
10. The eskimo who the anthropologist who no one visited spoke with had enough seal meat to eat during the winter.
11. The administrator who the teacher who no one talked to commended was fired at the end of the school year.
12. The columnist who the politician who no one trusted communicated with thought that the Republicans had a chance of winning the election.
13. The sergeant who the corporal who no one supervised argued with was considered a rising star in the army.
14. The sailor who the bartender who no one liked talked to wanted to tell a dirty joke.
15. The judge who the lawyer who many people praised agreed with thought the first ammendment should be interpreted as widely as possible.
16. The secretary who the temp worker who many people helped made friends with was very competent and a hard worker.
17. The nun who the priest who many people admired spoke with believed religion was becoming obsolete in modern society.
18. The park ranger who the ecologist who many people had heard of worked with devoted his life to saving old growth forests.

Appendix D Experimental items for Experiment 4

The items used in Experiment 4 are listed below. Both the RC (a) and CC (b) versions of each item are given, and individual conditions used one of the two NPs separated by slashes.

1a. The professor who the student/ I had recently met at the party was famous, but no one could figure out why.
1b. The professor said that the student/ I had recently met the philosopher, but he/ I might not have met the mathematician.
2a. The chairman who the consultant/ we had previously interviewed about the company was knowledgeable, but very resistant to changes in the structure of his company.
2b. The chairman thought that the consultant/ we had previously interviewed the employee and was reluctant to allow them to meet again.
3a. The student who the family/ we had willingly hosted during the summer was friendly and her English really improved during her stay.
3b. The student noticed that the family/ we had willingly hosted the German girl and wondered if they/we would be part of the exchange program again next year.
4a. The teacher who the child/ I had really admired after the lesson was talented, because she could explain very technical ideas in a simple way.
4b. The teacher saw that the child/ I had really admired the scientist, because she/ I started reading all about biology research.
5a. The policeman who the bicyclist/ we had not obeyed on the street was friendly and only issued a warning instead of a fine.
5b. The policeman realized that the bicyclist/ we had not obeyed the law because he/ we was/ were going the wrong way on a one way street.
6a. The advisor who the students/ you have always appreciated for her clear thinking is excited because she recently won a teaching award.
6b. The advisor noticed that the students/ you have always appreciated the help they/ you have gotten and decided to hire an assistant for them/ you.
7a. The counselor who the teenager/ I had previously called on the phone was helpful since she really cared about his/ my problems.
7b. The counselor concluded that the teenager/ I had previously called the psychologist because he/ I knew the medical terminology for his/ my illness.
8a. The doctor who the patient/ we had faithfully trusted with his/ our health was skillful, but it was a dangerous procedure so everyone was worried.
8b. The doctor thought that the patient/ we had faithfully trusted the surgeon, but this time he went out of his way to reassure her/ us since it was such a difficult procedure.
9a. The singer who the fan has/ you have always adored with all of her heart/ your heart is coming to town for a concert to promote her new record.
9b. The singer assumed that the fan had/ you have always adored her music because she / you had shown up at every concert on the East coast for the past ten years.
10a. The salesperson who the shopper/ I had immediately disliked from the beginning was unhelpful and refused to look for a bigger size.
10b. The salesperson feared that the shopper/ I had immediately disliked the pants and tried to convince her/ me of their worth.
11a. The woman who the boy/ you had accidentally pushed off the sidewalk got upset and decided to report the incident to the policeman standing nearby.
11b. The woman knew that the boy/ you had accidentally pushed the girl but gave him/ you a long lecture anyway.

12a. The judge who the lawyer/ we had greatly respected by the end of the trial was brilliant, but he had difficulty keeping the court in order.
12b. The judge realized that the lawyer/ we had greatly respected the decision, though some of the lawyer's/ our enemies were trying to make it look like he/we hadn't.
13a. The author who the editor/ I had really liked for his creativity was young but very talented.
13b. The author believed that the editor/ I had really liked the book despite evidence to the contrary.
14a. The candidate who the Democrat/ I had wholeheartedly supported during the campaign was liberal and wanted to increase welfare.
14bThe candidate figured that the Democrat/ I had wholeheartedly supported the plan and called to ask for another donation.
15a. The plumber who the landlord/ we had already hired for the job was incompetent but there was nothing to do because the contract had already been signed.
15b. The plumber feared that the landlord/ we had already hired the electrician who had recently lost his license.
16a. The freshman who the volunteer/ I had willingly tutored on a daily basis was bright, but he had difficulty concentrating.
16b. The freshman heard that the volunteer/ I had willingly tutored the sophomores and asked if he/ I would help him with his own work.
17a. The comedian who the teenager/ you had genuinely enjoyed during the talent show is staying and will do another show at the club tonight.
17b. The comedian sensed that the teenager/ you had genuinely enjoyed the routine and was happy about her performance.
18a. The landlord who the tenant/ you had previously met at a friend's house is pleased to have someone who she knows is responsible in the apartment.
18b. The landlord knew that the tenant/ you had previously met the neighbor so he was not surprised to find the neighbor visiting the apartment.
19a. The visitor who the host/ we had belatedly invited to the party was shy but ended up having a fantastic time.
19b. The visitor figured that the host/ we had belatedly invited the guest who came to the party in sweats, without a gift.
20a. The neighbor who the girl scout/ I had faithfully visited at the nursing home was old and sick and needed help fixing her dinner.
20b. The neighbor believed that the girl scout/ I had faithfully visited the nursing home and was so impressed with her/ my charity that she sent a letter to the newspaper.

Appendix E Experimental Items for Experiment 5

The items used in Experiment 5 are listed below. Individual conditions used one of the four NPs separated by slashes.

1. The writer who you/ the reporter/ a reporter/ CBS talked to wrote radical articles about the government.
2. The company which I/ the accountant/ an accountant/ Steve Jobs founded invented new software for web design.
3. The assistant who I/ the partner/ a partner/ Marcia Clark hired answered the phone with a growl.
4. The salesman who I/ the company/ a company/ Microsoft liked remembered the name of every customer.
5. The drag queen who I/ the dancer/ a dancer/ Dennis Rodman recognized wore red boots with leather fringe.
6. The hairdresser who I/ the rock star/ a rock star/ Madonna visited weaves long extensions into natural hair.
7. The actress who I/ the manager/ a manager/ Howard Stern talked with wore tight pants and a sweater.
8. The artist who you/ the heiress/ an heiress/ Martha Stewart admired painted beautiful portraits of little children.
9. The little girl who you/ the neighbor/ a neighbor/ Princess Diana comforted had a cast on her arm.
10. The girl scout who you/ a reporter/ the reporter/ Bill Clinton praised founded recycling programs at local schools.
11. The surgeon who you/ the insurance company/ an insurance company/ Medicaid reimbursed performed the operation with great success.
12. The doctor who you/ the basketball player/ a basketball player/ Michael Jordan consulted repaired torn ligaments in athletes' ankles.
13. The consultant who we/ the chairman/ a chairman/ Donald Trump called advised wealthy companies about tax laws.
14. The pianist who we/ the violinist/ a violinist/ Andrew Lloyd Webber respected won several competitions for young musicians.
15. The TV show which we/ the comedian/ a comedian/ Jerry Seinfeld loved portrayed everyday life in New York.
16. The editor who we/ the book store/ a book store/ Barnes and Noble consulted recommended some novels for summer reading.
17. The celebrity who we/ the official/ an official/ Ted Kennedy invited donated some money to the Democrats.
18. The acrobat who we/ the magician/ a magician/ Houdini watched did some tricks during the break.
19. The benefit which we/ the celebrity/ a celebrity/ Dolly Parton organized raised money for breast cancer research.
20. The bohemian artist who we/ the musician/ a musician/ Bob Dylan liked made garden ornaments from rusty nails.
21. The psychologist who I/ the neurotic man/ a neurotic man/ Woody Allen saw counseled many people from New York.
22. The author who we/ the talk show host/ a talk show host/ Oprah interviewed wrote a novel at age thirteen.

23. The lawyer who we/ the criminal/ a criminal/ OJ Simpson hired won every case he worked on.
24. The secretary who you/ the businessman/ a businessman/ Bill Gates greeted controlled all access to the office.

Appendix F Experimental Items for Experiment 6

The items used in Experiment 6 are listed below. Only the nested versions of the items are given, except for item 1. The right branching conditions can be computed from the nested conditions by extracting the innermost clause, following it by the middle clause and finishing with the outermost clause. Individual conditions contained either the name or the name plus the bracketed material.

1a. My friend Ann and I had a conversation yesterday about the academic squabbling that takes place in journalism school. A writer who a professor who Ann{'s boyfriend} talked to disliked had written many radical articles.
1b. My friend Ann and I had a conversation yesterday about the academic squabbling that takes place in journalism school. Ann{'s boyfriend} talked to a professor who disliked a writer who had written many radical articles.
2. I have a friend, Jane, who is working as a teaching assistant at a summer program for academically talented teenagers. An old professor who the students who Jane{'s roommate} lectured to liked told interesting stories.
3. I just met a woman named Susan who is helping to organize an anti-drug symposium next week. A secretary who a celebrity who Susan{'s co-worker} invited employed wrote an apologetic note declining the invitation.
4. There is a guy that I know, Alan, who claims to be the only fifty year old on the planet who really knows his way around the internet. Some programmers who the company which Alan{'s son} worked for had competed with invented a new way to design web pages.
5. Alice and I were complaining about the incompetence of corporate America yesterday. An assistant who a lawyer who Alice{'s friend} talked to had hired mistook the addresses on some letters.
6. When Bob went to work in Belgium, he was surprised at how friendly everyone there was. A salesman who a woman who Bob{'s supervisor} hired dealt with was very polite.
7. My friend Mike lives in the theater district, near a wild all-night club. A man who a drag queen who Mike{'s neighbor} recognized had accompanied was wearing lizard skin boots.
8. Polly's family lives right in the heart of Washington DC. A congressman who some passers-by who Polly{'s sister} avoided had recognized was catching a taxicab.
9. When I told Kathy I was going to Los Angeles, she suggested that I get my hair done. A hairdresser who an image consultant who Kathy{'s sister} visited had recommended has cut hair in Hollywood for thirty years.
10. Tom is an aspiring playwright who dreams of having his plays performed on Broadway, but he is currently stuck very far off Broadway. An actress who a producer who Tom{'s agent} had talked with liked wasn't very good.
11. Josh and the other neighborhood kids spend a lot of time hanging around the golf course. A golfer who an amateur who Josh{'s friend} beat had learned from had a cart with a racing stripe.
12. I saw Carl, my little nephew, playing hide and seek at the park yesterday. A child who a little girl who Carl{'s babysitter} had surprised was searching for had a cast on her arm.
13. My friend Mary was telling me the other day about the huge settlements that courts have been awarding to the victims of slander. A radio station which a disc jockey who Mary{'s boss} sued had worked for went out of business.
14. I have a cousin Sam who works for a fashion magazine. A model who a photographer who Sam{'s editor} had hired took pictures of was very famous.
15. My friend Brad belongs to a church which is active in many socially liberal causes. An old lady who a government assistance program which Brad{'s pastor} praised had saved did not have enough money to heat her house.

16. Jenn and a lot of other girls her age make a little extra pocket money over the summer by babysitting at night. A TV show which some kids who Jenn{'s classmate} babysat loved was going to be cancelled next season.
17. I met a guy named Alex at a gym in Bel Air. A movie star who a trainer who Alex{'s girlfriend} worked with worshiped has just made a new movie.
18. A prodigy on the violin, Frank participated in a competition where the prize was to play at Lincoln Center, but he lost. A soloist who the judge who Frank{'s teacher} respected chose had won several international competitions.
19. My neighbor Fred went to the hospital the other day because he had a severe pain in his left foot. A nurse who a specialist who Fred{'s doctor} consulted had called arrived promptly.
20. I know a man, George, who works in the customer service department of the local telephone company. A housewife who a secretary who George{'s manager} supervised had contacted was late paying her bill.

Appendix G Experimental Items for Experiment 7

The items used in Experiment 7 are listed below. The sentences marked with an (a) are the explicit mention contexts, the sentences with a (b) are the easy bridge contexts and the sentences with a (c) are the hard bridge contexts. The sentences with only a number are the target sentences that followed the contexts.

1a. The company CEO analyzed each employee's performance in January.
1b. The company executives analyzed each employee's performance in January.
1c. The company analyzed each employee's performance in January.
1. One of the employees who the CEO promoted completed every project on schedule.
2a. A physicist attended the cross-disciplinary conference at the university.
2b. Some scientists attended the cross-disciplinary conference at the university.
2c. Some professors attended the cross-disciplinary conference at the university.
2. A speaker who the physicist questioned explained the results in more detail.
3a. The sportscaster was disappointed in the Red Sox game last night.
3b. The TV broadcast crew was disappointed in the Red Sox game last night.
3c. The crowd was disappointed in the Red Sox game last night.
3. One of the players who the sportscaster criticized made three errors in the field.
4. The nursery school director prepared kids for kindergarten.
4b. The nursery school prepared kids for kindergarten.
4c. The educational program prepared kids for kindergarten.
4. One of the children who the school director praised shared some toys with another child.
5a. The book reviewer published a new list of interesting books before Christmas.
5b. The magazine's book section published a new list of interesting books before Christmas.
5c. The magazine published a new list of interesting books before Christmas.
5. One of the novels which the reviewer liked recounted the voyages of Darwin.
6a. The pigeon easily learned tricks which built on its natural behavior.
6b. The birds easily learned tricks which built on their natural behaviors.
6c. The animals easily learned tricks which built on their natural behaviors.
6. One of the tricks that the pigeon loved involved pecking at a yellow bar to get a treat.
7a. Some lucky city kids got the opportunity to hike with a wilderness guide.
7b. Some lucky city kids got the opportunity to hike on a wilderness expedition.
7c. Some lucky city kids got the opportunity to hike in the wilderness.
7. One of the kids who the guide encouraged twisted an ankle on a rock.
8a. The high school high-jumper competed at the city championships.
8d. The high school track team competed at the city championships.
8c. The high school students competed at the city championships.
8. A referee who the high-jumper insulted disqualified the team for unsportsmanlike behavior.
9a. The star quarterback was interviewed by a group of reporters after the win.
9b. The star football players were interviewed by a group of reporters after the win.
9c. The star athletes were interviewed by a group of reporters after the win.
9. One of the reporters who the quarterback trusted asked about the state of the contract negotiations.
10a. The dean of Engineering was looking into which departments could be cut.
10b. The university officials were looking into which departments could be cut.
10c. The university was looking into which departments could be cut.
10. One of the departments which the dean reviewed had a strong research program.
11a. The architect was in a meeting about plans for a new MIT building yesterday.

11b. The building design committee was in a meeting about plans for a new MIT building yesterday.
11c. The administrator was in a meeting about plans for a new MIT building yesterday.
11. A decorator who the architect recommended worried that the third floor would be too dark.
12a. The public wanted an apology from the police chief after the scandal broke.
12b. The public wanted an apology from law enforcement personnel after the scandal broke.
12c. The public wanted an apology from city officials after the scandal broke.
12. A citizen's group who the police chief approached asked for an admission of guilt.
13a. The copilot was angry about being asked to fly more hours than his union contract specified.
13b. The cockpit crew was angry about being asked to fly more hours than their union contracts specified.
13c. The flight crew was angry about being asked to fly more hours than their union contracts specified.
13. The union representative who the copilot telephoned contacted the union president to talk about the issue.
14a. The mother put the Snickers bar on the top shelf so that it was out of the children's reach.
14b. The mother put the candy bars on the top shelf so that they were out of the children's reach.
14c. The mother put the groceries on the top shelf so that they were out of the children's reach.
14. The child who the Snickers bar tempted used a chair to get it.
15a. The sorority sisters tried on some earrings while they were getting dressed to go out.
15b. The sorority sisters tried on some jewelry while they were getting dressed to go out.
15c. The sorority sisters tried on some accessories while they were getting dressed to go out.
15. The girl who the earrings flattered wore a matching necklace to go with them.
16a. John bought a computer for his study, but had trouble deciding where to put it.
16b. John bought some electronic equipment for his study, but had trouble deciding where to put it.
16c. John bought some equipment for his study, but had trouble deciding where to put it.
16. The outlet that the computer needed occupied a space that was out of reach from everywhere.

Appendix H Experimental Items for Experiment 8

The items from Experiment 8 are listed below. The new / indefinite condition included the indefinite article and did not include the by-phrase. The old / definite condition included the definite article and the by-phrase. The new / definite condition included the definite article and did not include the by-phrase.

1. A suspect in a bank robbery was caught {by a detective} on Friday. The suspect, who {a/the} detective had sighted on Wednesday, struggled but was eventually subdued.
2. A gymnast at a meet was penalized {by a judge} for her floor routine. The gymnast, who {a/the} judge had warned during warm ups, exceeded the allotted time on the event.
3. A columnist for the Boston Globe was invited {by an activist} to a debate. The columnist, who {an/the} activist had criticized on the radio, opposed rent control very strongly.
4. A restaurant in New York was praised {by a wine expert} on the Food-TV network. The restaurant, which {a/the} wine expert had bought in April, served wonderful French cuisine.
5. A stonecutter at a quarry was sent home {by a foreman} after straining his back. The stonecutter, who {a/the} foreman had excused earlier that week, missed many days of work.
6. A legal expert for NBC was interviewed {by an anchorman} about the breaking news in the murder case. The expert, who {an/the} anchorman had scheduled on the day before, thought that the prosecution's case was strong.
7. A student in Cambridge elementary school was shoved {by a bully} on the playground this afternoon. The student, who {a/the} bully had harassed all week, started to cry loudly.
8. A student in the fourth grade was sent {by a teacher} to the principal's office. The student, who {a/the} teacher had scolded earlier, sat quietly and listened to the principal's advice.
9. A waiter at Legal Seafood was tipped generously {by a customer} last night. The waiter, who {a/the} customer had complimented previously that evening, gave fantastic recommendations about wines.
10. A singer with a beautiful voice was discovered {by a recording agent} in a shopping mall. The singer, who {an/the} agent had overlooked earlier, signed a contract for three albums with a record company.
11. A bouncer on his way home was attacked {by a drunk customer} outside a bar. The bouncer, who {a/the} customer had threatened earlier in the evening, quit his job later in the week.
12. An engineer at Microsoft was advised {by an accountant} to sell his stock options soon. The engineer, who {an/the} accountant had advised before, ignored the advice and bought more stock instead.

Appendix I Experimental Items for Experiment 9

The items from Experiment 9 are listed below. The new / indefinite condition included the first of the two nouns in brackets and the indefinite article. The old / definite condition included the second of the two nouns in brackets and the definite article.

1. A {police chief/ detective} was investigating a bank robbery. On Friday, he finally caught one of the suspects. The suspect, who {a/the} detective had sighted the previous Wednesday, struggled but was eventually subdued.
2. A {TV commentator/ judge} was watching a gymnastics meet. She monitored a gymnast's floor routine. The gymnast, who {a/the} judge had warned during the warm-ups, exceeded the allotted time on the event.
3. A {talk show host/ activist} was driving home. During a morning talk show he had spoken with a newspaper columnist. The columnist, who {an/the} activist had criticized on the radio, opposed rent control very strongly.
4. An {investor/ wine expert} was starting to invest in some local businesses. This month she noticed that one little restaurant generated high profits. The restaurant, which {a/the} wine expert had bought four months ago, served wonderful French cuisine.
5. A {payroll officer/ foreman} was working at a stone quarry. On payday he noticed that one stonecutter was absent. The stonecutter, who {a/the} foreman had excused earlier that week, missed many days of work due to a knee injury.
6. An {anchorman/ reporter} on NBC was covering a high-profile murder. Last night on TV, he interviewed a famous legal expert. The expert, who {a/the} reporter had briefed the day before, thought that the prosecution's case was strong.
7. A {boy/ bully} was walking around his school playground. Suddenly, he shoved a girl off a swing. The girl, who {a/the} bully had harassed previously that morning, started to cry loudly.
8. A {parent volunteer/ teacher} was supervising a fourth grade class. During recess, she sent a misbehaving girl to the principal's office. The girl, who {a/the} teacher had scolded on the playground, sat quietly and listened to the principal's advice.
9. The {owner/ manager} of a restaurant was deciding on Employee of the Month Awards. Suddenly she realized there was one waiter who really deserved the award. The waiter, who {a/the} manager had complimented previously that evening, gave fantastic recommendations about wines.
10. A {model/ agent} was hosting a talent search at a shopping mall. Yesterday morning, she discovered a singer with a beautiful voice. The singer, who {an/the} agent had overlooked in previous sessions, signed a contract for three albums with a record company.
11. A {gang member/ customer} was leaving a bar last night. On his way out, he insulted the bartender. The bartender, who {a/the} customer had threatened earlier that evening, quit his job later in the week.
12. A {financial advisor/ accountant} was working at Microsoft. This morning, she suggested to an engineer that he should sell some of his stock options. The engineer, who {an/the} accountant had advised many times before, ignored the advice and bought more stock instead.

Works Cited

Albrecht, J. & Myers, J. (1995). Role of context in accessing distant information during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21 (6),* 1459-1468.

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review, 106*:(4) 748-765.

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

Altmann, G., VanNice, K., Garnham, A. & Henstra, J. (1998). Late closure in context. *Journal of Memory and Language, 38*, 459-484.

Ariel, M. (1988). Referring and Accessibility. *Journal of Linguistics, 24*, 65-87.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London; Routledge.

Arnold, J. (1998). *Reference Form and Discourse Patterns*. Palo Alto: Stanford University dissertation.

Arnold, J. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes, 31(2),* 137-162.

van Berkum, J., Brown, C. & Hagoort, P. (1999). Early referential context effects in sentence processing: evidence from event-related brain potentials. *Journal of Memory and Language, 41*, 147-182.

Bever, T.G. (1970). The cognitive basis of linguistic structures. In J.R. Hayes (Ed.), *Cognition and the development of language*. John Wiley, New York, NY.

Bever, T.G. (1974). The ascent of the specious, or there's a lot we don't know about mirrors. In D. Cohen (Ed.) *Explaining linguistic phenomena*. Hemisphere Pub. Co., Washington.

Brennan, S. (1995). Centering attention in discourse. *Language and Congitive Processes, 10*(2), 137-167.

Chafe, W. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed.), *Coherence and Grounding in Discourse*. Philadelphia, PA: John Benjamins. 21-51.

Chambers, G. & Smyth, R. (1998). Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language, 39* (4*)*, 593-608.

Chomsky, N. (1957). *Syntactic Structures*, Mouton, The Hague, The Netherlands.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In

R.D. Luce, R. R. Bush & E. Galanter (Eds.) *Handbook of mathematical psychology*, volume 2, 269-321, John Wiley, New York, NY.

Clark, H. & Sengul, C. (1979). In search of referents for nouns and pronouns. *Memory & Cognition, 7*, 35-41.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language processing: Psychological, computational and theoretical perspectives.* Cambridge, UK: Cambridge University Press.

De Roeck, A., Johnson, R., King, M., Rosner, M., Sampson, G. & Varile, N. (1982). A myth about center-embedding. *Lingua*, 58, 327-340.

Enç, M. (1983). Anchored expressions. In *Proceedings of the west coast conference of formal linguistics, 2*, pp. 79-88.

Frank, R. (1992). *Syntactic locality and tree-adjoining grammar: Grammatical, acquisition and processing presepctives*. PhD Thesis, University of Pennsylvania, Philadelphia, PA.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart, ed., *Attention and performance XII*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Frazier, L. (1999). *On Sentence Interpretation*. Boston; Kluwer Academic Publishers.

Garnham, A. (1987). Mental Models as Representations of Discourse and Text. New York, NY: Ellis Horword Limited.

Garnham, A. (1997). Representing information in mental models. In M. Conway (Ed.), *Cognitive Models of Memory*. Cambridge, MA; MIT Press. 149-172.

Garnham, A. & Oakhill, J. (1988). "Anaphoric islands" revisited. *The Quarterly Journal of Experimental Psychology*, 40A (4), 719-735.

Garnham, A. & Oakhill, J. (1989). The everyday use of anaphoric expressions: Implications for the 'Mental Models' theory of text comprehension. In N.E. Sharkey (Ed.), *Models of cognition: A review of cognitive science*. Norwood, N.J.; Ablex Publishing Co. 78-112.

Garnham, A. & Oakhill, J. (1990). Mental models as contexts for interpreting texts: Interpretations from studies of anaphora. *Journal of Semantics, 7*, 379-393.

Garnham, A., Oakhill, J., Ehrlich, M-F., & Carreiras, M. (1995). Representations and processes in the interpretation of pronouns: new evidence from Spanish and French. *Journal of Memory and Language, 34*, 41-62.

Garrod, S. & Sanford, A.J. (1977). Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior, 16*, 77-90.

Garrod, S. & Sanford, A.J. (1982). The mental representation of discourse in a focused memory system: implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 1, 21-41.

Garrod, S & Sanford, A.J. (1994). Resolving sentences in a discourse context: How discourse representation affects language understanding. In M.A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. San Diego, CA:Academic Press.

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, 32*, 1-30.

Gernsbacher, M.A. (1989). Mechanisms that improve referential access. *Cognition, 32*, 99-156.

Gernsbacher, M.A. (1991). Comprehending conceptual anaphors. *Language and Cognitive Processes*, 6 (2) 81-105.

Gibson, E. (1998). Syntactic complexity: Locality of syntactic dependencies. *Cognition, 68* (1), 1-76.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In Y. Miyashita, A.P. Marantz & W. O'Neil (Eds.), *Image, Language, Brain*. Cambridge, MIT press.

Gibson E. & Ko, K. (1998). Processing main and embedded clauses. Manuscript, MIT, Cambridge, MA.

Gibson, E. & Pearlmutter, N.J. (1998). Constraints on sentence comprehension. *Trends in Cognitive Science, 2*, 262-268.

Gibson, E., Pearlmutter, N.J., Canseco-Gonzalez, E. & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, 59, 23-59.

Gibson, E., Pearlmutter, N.J., & Torrens, V. (1999). Recency and lexical preferences in Spanish. *Memory and Cognition, 27*, 603-611.

Givon, T. (1983). *Topic continuity in discourse: A quantitative cross-language study.* Amsterdam: John Benjamins.

Gordon, P., Grosz, B. & Gilliom, L. (1993). Pronouns, names and the centering of attention in discourse. *Cognitive Science, 17*, 311-347.

Gordon, P.C., Hendrick, R. & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition.* In Press.

Graesser, A., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101 (3),* 371-395.

Greene, S., McKoon, G. & Ratcliff, R. (1992). Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18(2),* 266-283.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L.Morgan, (Eds.), *Syntax and Semantics, 3: Speech acts*, 41-58. New York: Academic Press.

Grodner, D., Gibson, E. & Tunstall, S. (2001). Syntactic complexity in ambiguity resolution.

Manuscript accepted for publication, *Journal of Memory and Language*.

Grodner, D., Watson, D. & Gibson, E. (2000). *Locality effects in processing unambiguous sentences*. Talk at CUNY XIII conference on human sentence processing, San Diego, CA.

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

Gundel, J., Hedberg, H. & Zacharski, R. (1993). Referring expressions in discourse. *Language, 69*, pp. 274-307.

Haliday, M.A.K. & Hassan, R. (1976). *Cohesion in English*. Longman, London.

Hankamer, J., Sag, I. (1976). Deep and surface anaphora. *Linguistic Inquiry, 7*, 391-428.

Haviland, S.E. & Clark, H.H. (1974). What's new? Acquiring new information as a process of comprehension. *Journal of Verbal Learning and Verbal Behavior,* 13, 521-521.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. PhD Thesis. University of Massachusetts, Amherst, MA.

Heim, I & Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell Publishers Inc.

Kac, M.B. (1981). Center-embedding revisited. *In Proceedings of the third annual conference of the Cognitive Science Society* (pp. 123-124). Hillsdale, NJ: Lawrence Erlbaum.

Kamp, H. (1981). A theory of truth and semantic representations. In J. Groenendijk (Ed.*) Formal methods in the study of language*. Mathematisch Centrum, Amsterdam, The Netherlands.

Kamp, H. & Reyle, U. (1993). *From discourse to logic*. Dordrecht: Kluwer Academic Publishers.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition, 2*, 15-47

King, J & Just, M. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*, 580-602.

Kintsch, W. (1998). *Comprehension*. New York, NY: Cambridge University Press.

Lucas, M. M., Tanenhaus, M. K. & Carlson, G. N. (1990). Levels of representation in the interpretation of anaphoric reference and instrument inference. *Memory and Cognition,* 18 (6), 611-631.

MacDonald, M. & Christiansen, M. (2001). Reassessing working memory: A comment on Just & Carpenter and Waters & Caplan. *Psychological Review*. In Press.

MacDonald, M., Pearlmutter, N. & Seidenberg, M. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*. 101 (4), 676-703.

Maratsos, M. (1976). *The use of definite and indefinite reference in young children*. New York, NY: Cambridge University Press.

Marslen-Wilson, W., Levy, E. & Tyler, L.K. (1982). Producing interpretable discourse: The establishment and maintenance of reference. In R.J. Jarvella and W. Klein (Eds.), *Speech, Place and Action*. New York, NY: John Wiley.

Mauner, G., Tanenhaus, M. K. & Carlson, G. N. (1995). Implicit arguments in sentence processing. *Journal of Memory and Language*, 34, 357–382.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99 (3), 440-466.

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. *Journal of Memory and Language, 32*, 56-75.

McKoon, G., Ratcliff, R., Ward, G., & Sproat, R. (1993). Syntactic prominence effects on discourse processes. *Journal of Memory and Language, 32*, 593-607.

Miller, G.A. & Chomsky, N. (1963). Finitary models of language users. In R.D. Luce, R.R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (volume 2, pp. 419-491). New York, NY: John Wiley.

Miller, G.A. & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control, 7*, 292-303.

Morrow, D. & Greenspan, S. (1989). Situation models and information accessibility. In N.E. Sharkey (Ed.), *Models of cognition: A review of cognitive science*. Norwood, N.J.; Ablex.

Murphy, G.L. (1984). Establishing and accessing referents in discourse. *Memory and Cognition, 12* (5), 489-497.

Myers, J. & O'Brien, E. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26*(2&3), 137-157.

Pearlmutter, N.J. & Gibson, E. (2001). Recency and verb phrase attachment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27* (2), 574-590.

Pollard, C. & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago, IL: University of Chicago Press.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics*, 223-256. New York, NY: Academic Press.

Reinhart, T. (1983). Coherence and bound anaphora- A restatement of the anaphora questions. *Linguistics and Philosophy, 6* (1), 47-88.

Rizzella, M. & O'Brien, E. (1996). Accessing global causes during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22 (5), 1208-1218.

Sag, I. & Hankamer, J. (1984). Toward a theory of anaphoric processing. *Linguistics and Philosophy, 7*, 325-345.

Sanford, A. (1989). Component processes of reference resolution in discourse. In N.E. Sharkey (Ed.), *Models of Cognition: A Review of Cognitive Science*, 113-140. Norwood, NJ: Ablex Publishing Co.

Sanford, A. & Garrod, S. (1981). *Understanding written language: Explorations in comprehension beyond the sentence*. Chichester: Wiley.

Sanford, A.J., Moar, K., & Garrod, S. (1988). Proper names as controllers of discourse focus. *Language and Speech. 31*, 1, 43-55.

Sperber, D. & Wilson, D. (1995). *Relevance*: *Communication and cognition.* (2nd ed.). Oxford:Blackwell.

Tanenhaus, M. K. & Carlson, G. N. (1990). Comprehension of deep and surface verb-phrase anaphors. *Language and Cognitive Processes, 5*, 257-280.

Tanenhaus, M.K. & Trueswell, J.C. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *Speech, Language and Communication*, Academic Press, San Diego, CA.

Van Dyke, J. (2001). *Syntactic and referential effects in processing complex sentences*. Poster presented at the 14th annual CUNY conference on human sentence processing, Philadelphia, PA.

Webber, B. L. (1979). *A Formal Approach to Discourse Anaphora*. New York, NY:Garland Publishing Inc.

Yngve, V.H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104, 444-466.