ELSEVIER

Cognition 85 (2002) 79–112

COGNITION

www.elsevier.com/locate/cognit

The influence of referential processing on sentence complexity

Tessa Warren[a,*], Edward Gibson[b]

[a]*Psychology Department, University of Massachusetts at Amherst, Amherst, MA 01003, USA*
[b]*Massachusetts Institute of Technology, Cambridge, MA, USA*

Received 3 March 2001; received in revised form 18 January 2002; accepted 24 April 2002

Abstract

This paper reports the results of five experiments designed to investigate the effects of referential processing on sentence complexity. Gibson (*Cognition*, *68* (1998) 1) suggested that sentence complexity is related to the locality of integrations between dependent syntactic heads, and that an appropriate measure of locality is the number of new discourse referents intervening between the endpoints of those integrations. The experiments in this paper test, modify and extend Gibson's (1998) claims. Each experiment manipulated noun phrases (NPs) in the subject positions of object-extracted relative clauses in order to determine how different types of NPs affected sentence complexity. Experiments 1, 2 and 3 used questionnaires to gauge sentence complexity, whereas Experiments 4 and 5 used self-paced reading. The results from Experiments 1, 2, 4 and 5 suggest that the complexity of the experimental items was more closely related to the Givenness status of the embedded subject in the Givenness Hierarchy than to whether the embedded subject was old or new to the discourse. Experiment 3 compared materials in which a quantifier was rotated through subject positions of a nested relative clause structure. The results of this experiment support a discourse-processing-based distance metric for computing locality and provide evidence against a pure similarity-based account of structural complexity such as proposed by Bever (Bever, T. G. (1970). The cognitive basis of linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley).

Keywords: Language comprehension; Linguistic complexity; Referential processing

1. Introduction

It has long been known that *nested* (or *center-embedded*) syntactic structures are more difficult to process than non-nested structures. Increasing the number of nestings makes a

* Corresponding author.
E-mail address: tessa@psych.umass.edu (T. Warren).

0010-0277/02/$ - see front matter
PII: S0010-0277(02)00087-2

sentence unprocessable (Chomsky, 1957, 1965; Chomsky & Miller, 1963; Miller & Chomsky, 1963; Miller & Isard, 1964; Yngve, 1960; see Gibson, 1998, for a review of relevant literature). For example, the sentences in (1) are increasingly complex:

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

(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", resulting in a more complex sentence. The sentence in (1c) is doubly nested, consisting of the nested structure in (1b) with a second nesting: the RC "which the neighbors recommended" 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 recommended 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.

Although structures like (1c) are usually very difficult to process, some examples of processable doubly nested RC structures have been noted in the literature:

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

The relevant property of these doubly nested RC structures seems to be that the most embedded NP is a first- or second-person pronoun: an *indexical* pronoun (Gibson, 1991; Kac, 1981). 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, this account does not 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.

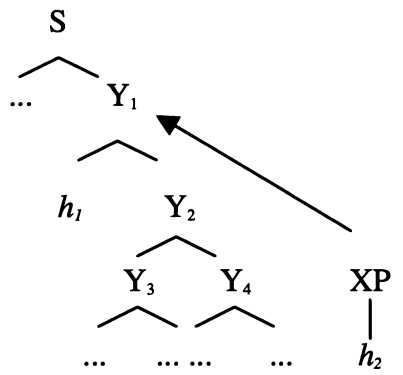

Fig. 1. Structural integration of a maximal projection XP of a newly input head h_2 to an attachment site headed by a head h_1 in the structure for the input so far.

A recent account of the contrast between (1c) and the sentences in (3) is the Gibson (1998, 2000) distance-based dependency locality theory (DLT). 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 relevant aspect of this theory with respect to the contrasts of interest here is that 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 Fig. 1) involves retrieving aspects of h_1 from memory. 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.

Crucial to the Gibson (1998) account of the contrast between (1c) and the sentences in (3) is the assumption that building new discourse structure (e.g. a discourse referent: a representation of an individual or event; Heim, 1982; Kamp, 1981) requires more resources than accessing previously constructed discourse structure. Evidence for this hypothesis from the processing of unambiguous sentence structures is provided by Haviland and Clark (1974), Haliday and Hassan (1976), Garrod and Sanford (1977, 1982, 1994) and Murphy (1984) among others. Evidence for this hypothesis from the processing

[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, Desmet, Grodner, Watson, & Ko, 2001). Consequently, we 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 we will follow this convention here.

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. 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. As an initial simplifying hypothesis, it was 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:

(4) 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, one unit of cost will be counted for each new discourse referent in the intervening region.

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, such as lexical frequency, contextual plausibility, and reanalysis difficulty (see Gibson & Pearlmutter, 1998; Tanenhaus & Trueswell, 1995 for summaries of relevant results). Reading time data in support of the distance-based integration hypothesis with respect to unambiguous sentence materials is provided by 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, and Hickok (1996), Gibson, Pearlmutter, and Torrens (1999), and Pearlmutter and Gibson (2001).

The overall intuitive complexity of a sentence (as measured by questionnaire experiments, for example) depends to a large degree on the maximum intuitive complexity incurred at any state during its processing. For sentences such as (1c) and (3a–c), 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 (1c): integrating the verb "recommended" to its subject "the neighbors", and integrating the object position of the verb "recommended" to the filler "which", 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.

(5) The nanny [who the agency [which you recommended] sent] was adored by all the children.

Each of the integration steps crossing the most embedded subject is less costly when processing sentence (5), in which "the neighbors" is replaced with the indexical pronoun "you". First- and second-person pronouns require little processing, because their referents are assumed to be a default part of the domain of discourse even in a null context, since every discourse has a speaker/writer and a listener/reader (Chafe, 1987; Enç, 1983). Thus, integrating the object position of the verb "recommended" to the filler "who" crosses only one new discourse referent in (5), rather than two for the corresponding integration in (1c). Furthermore, the integration steps at the verb "sent" in (5) cross only two and three new discourse referents, rather than three and four new discourse referents in (1c). And the subject-verb integration at "was adored" in (5) crosses four new discourse referents rather than five, as in (1c).

This paper reports five experiments that investigate 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 the binary distance metric proposed in Gibson (1998) against a continuous metric we propose based on the Gundel, Hedberg, and Zacharski (1993) Givenness Hierarchy. Experiment 3 tested the ordering of subject NPs in doubly nested sentences, a factor predicted to affect complexity by the DLT but not by Bever's similarity theory. Experiments 4 and 5 test the DLT's predictions for singly nested sentences using the self-paced reading method.

2. Experiment 1

Experiment 1 tested the claim that doubly nested sentences with first- and second-person pronouns in the most embedded subject position are easier to process than corresponding sentences with other types of NPs in that position. Experiment 1 tested four types of NP as the most embedded subject: first-/second-person pronouns, short names, third-person pronouns and definite descriptions. Sentences were presented in null context so that first-/second-person pronouns referred to default referents, while names, third-person pronouns and definite descriptions introduced new referents. The third-person pronouns were infelicitous because pronouns do not usually introduce new referents. If the discourse-referent-based distance metric is correct, the conditions with the first-/second-person pronouns should be easier to process than the other conditions.

2.1. Method

2.1.1. 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 min to complete.

2.1.2. Materials

Twenty doubly nested experimental items were tested, each with four conditions. Conditions had a first-/second-person pronoun, a third-person pronoun, a short proper name or a definite description as the subject of the most deeply embedded clause. For example:

(6) a. First/second pronoun
The student who the professor who I collaborated with had advised copied the article.
b. Non-referring third pronoun
The student who the professor who they collaborated with had advised copied the article.
c. Short proper name
The student who the professor who Jen collaborated with had advised copied the article.
d. Definite description
The student who the professor who the scientist collaborated with had advised copied the article.

The third-person pronoun conditions were constructed so that the pronoun could not be interpreted as referring to any individual introduced in the sentence. The short proper name condition provided a comparison that kept length similar to the first/second pronoun condition. Fourteen of the proper names were two or three letters long, five names were four letters long and one was five letters long.

Each questionnaire was made up of 20 experimental items and 80 fillers. The fillers were similar in length and complexity to the experimental items. 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.

2.1.3. Procedure

Participants were asked to rate the complexity of sentences on a scale of 1 to 5, 1 being "easy to understand" and 5 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.

2.2. Results

The mean ratings for each condition are presented in graphical form in Fig. 2. When tested with repeated measures F-tests and adjusted for multiple tests using the Bonferroni correction, the first-/second-person pronoun condition was significantly less complex than each other condition as follows: the third-person pronoun condition ($F1(1,39) = 19.99$, $\mathrm{MSe} = 0.27$, $P < 0.005$; $F2(1,19) = 17.38$, $\mathrm{MSe} = 0.16$, $P < 0.01$), the short name condition ($F1(1,39) = 18.01$, $\mathrm{MSe} = 0.14$, $P < 0.005$; $F2(1,19) = 13.65$, $\mathrm{MSe} = 0.09$, $P < 0.01$) and the definite description condition ($F1(1,39) = 32.29$, $\mathrm{MSe} = 0.15$, $P < 0.005$; $F2(1,19) = 33.22$, $\mathrm{MSe} = 0.08$, $P < 0.005$). There were no differences among the third pronoun, name and definite description conditions (all $F < 4$, all $P > 0.2$ with Bonferroni corrections). Furthermore, the results were unchanged in an analysis over the first/second 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 first/second pronoun condition and the other conditions.

2.3. Discussion

The results of Experiment 1 show that doubly nested sentences with first-/second-person pronouns in the innermost subject position are less complex than the same sentences with other NPs in that position. The DLT predicts this result, because (1) the most embedded subject position interrupts multiple long distance dependencies and (2) the first-/second-

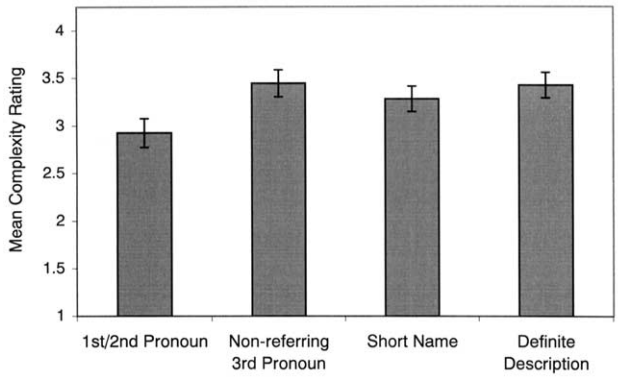

Fig. 2. Mean complexity ratings for Experiment 1.

person pronoun condition does not require the building of a new discourse referent at the most embedded subject position, while the other conditions do.

Alternatively, the observed complexity differences may have been due to differences in the difficulty of *accommodating* or *bridging* a referent for the innermost subject NP (Haviland & Clark, 1974; Heim, 1982). Names, definite descriptions and pronouns carry presuppositions of uniqueness and familiarity. When they are used to introduce new referents, these presuppositions must be accommodated in the discourse. The additional processing cost of accommodation could explain why the name, third-person pronoun and definite description conditions were more complex than the first-/second-person pronoun conditions.

Experiment 2 addresses the above issue and also investigates an alternative formulation of the DLT's integration cost metric. The integration cost metric in Gibson (1998) relies on a binary distinction between old and new referents: processing all old referents is assumed to be cost free, whereas processing new referents is assumed to be equally costly for all new referents. But there is evidence suggesting that not all new referents are difficult to process and not all old referents are easy to process (Garrod & Sanford, 1982; Gordon, Grosz, & Gilliom, 1993; Tanenhaus & Carlson, 1990). Experiment 2 compares the binary version of the DLT's integration cost metric to a continuous version based on the accessibility of the intervening discourse referents.

3. Experiment 2

Like Experiment 1, Experiment 2 tested the complexity of doubly nested RC structures with different types of NPs in the most embedded subject position. As an alternative to the binary distance metric proposed in Gibson (1998), we considered the predictions of a continuous metric, based on the accessibility of referents in discourse (Ariel, 1988, 1990; Arnold, 1998; Chafe, 1976, 1987; Du Bois, 1987; Givón, 1979, 1983, 1984; Gundel et al., 1993; Prince, 1981). Gundel et al. (1993) hypothesized that speakers indicate the discourse accessibility (or *cognitive status*) of the referent they are referring to by using a particular linguistic expression (e.g. a pronoun, a name, or definite description). In particular, Gundel et al. proposed the Givenness Hierarchy in (7), a relationship between the type of an NP and the degree to which its antecedent is accessible (cf. Ariel, 1988; Garrod, Freudenthal, & Boyle, 1994; Garrod & Sanford, 1982; Prince, 1981):

(7) The Gundel et al. (1993) Givenness Hierarchy:

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

A referent for an NP which is low on the hierarchy (*peripheral*[2]) is signaled to be newly introduced or found in long-term memory and thus relatively inaccessible, whereas a

[2] We are introducing the terms "central" and "peripheral" to refer to opposite ends of the Givenness Hierarchy.

referent for an NP high on the hierarchy (*central*) must be highly activated in the current discourse model and very accessible.

In order to use the Givenness Hierarchy in a theory of linguistic complexity, we propose a theory of NP processing such that cost is related to the amount of resources required to access and/or build a referent for an NP. If an NP has a referent in context, the accessibility of that referent determines processing cost. If an NP does not have a referent in context, then the referent access process will fail, and the amount of resources used by that failed process will contribute to cost. The cost of this failed process will be greater for an NP that is usually used for less accessible or more peripheral referents than for an NP that can only be used for very accessible or central referents, because in the former case a larger set of potential referents must be evaluated. Thus, building and/or accessing a referent for a more central NP type may be easier than building/accessing one for a more peripheral NP type in the Givenness Hierarchy. This hypothesis is consistent with recent theories of NP processing and referential access (e.g. Garrod et al., 1994; Myers & O'Brien, 1998).

If accessing or introducing referents for NPs from the more peripheral end of the Givenness Hierarchy requires more processing than accessing referents for NPs from more central levels, then, assuming a distance-based hypothesis like that in the DLT, integrations crossing more peripheral NPs should be more complex than integrations crossing more central NPs. Experiment 2 tested this prediction by having participants judge the complexity of doubly nested sentences in which the most embedded subject NP varied across six levels of the Givenness Hierarchy, listed in order from most central to most peripheral: (1) first-/second-person pronouns; (2) third-person pronouns; (3) first names; (4) full names (which are also names of famous people and companies); (5) definite descriptions; and (6) indefinite descriptions. The pronominal elements are the most central of these NP types, because they refer to focused elements of a discourse. The first-/second-person pronouns are more central than the third-person pronouns, because their referents are default in context and unambiguous. In the third pronoun conditions, each sentence included a topicalized name that introduced a highly accessible referent to which the pronoun referred. The names used in the first name and full name conditions fall between the pronouns and the descriptions on the hierarchy. Ariel (1990) and Clancy (1996) have hypothesized that first names are generally used to refer to more active or central referents than full names or descriptions. Hence, the search for a referent for a first name should cause less processing load. The full names used in this experiment referred to famous referents, which were accessible from long-term memory but not highly activated in context. Definite descriptions are the most peripheral form of the definite NPs that were tested. Because they can refer to relatively inaccessible referents, the referent access process they initiate will have a high cost. Unlike the definite NPs, an indefinite description requires no search for an appropriate referent, because indefinites typically introduce referents. But introducing a referent in the subject position of a restrictive RC may cause complexity, because this is not typically where such NPs are introduced (cf. Fox & Thompson, 1990). Indefinites in this position may therefore be more complex than the other NP types. Thus, a continuous distance metric based on the Givenness Hierarchy predicts increasing complexity across the six conditions in the nested structures.

In contrast, the binary metric predicts that there should be only two levels of difficulty among the six conditions. The two pronominal conditions should be in the less complex

group, whereas the short name, definite descriptions and indefinite descriptions should be in the more complex group. The names used in the full name condition all referred to famous people (e.g. "Bill Clinton", "Tiger Woods"), or in a few cases, well-known companies (e.g. "Microsoft", "Barnes and Noble"). Because these names have referents in long-term memory, the binary distance metric predicts that the full name condition should be in the less complex group, patterning with the pronouns.

Finally, the presence of definite and indefinite descriptions allowed us to test the possibility that accommodation failure and its resultant infelicity caused the high complexity in the definite description condition in Experiment 1. Indefinite descriptions carry no presuppositions of uniqueness or familiarity and thus require no accommodation. If accommodation failure caused the complexity of the definite description condition in Experiment 1, then the indefinite condition should be less complex than the definite description condition in Experiment 2.

Experiment 2 also included right branching versions of the target sentences to control for plausibility differences between conditions. Both the binary and continuous distance metrics predict that the reference differences among conditions will be stronger in the nested versions, resulting in an interaction between NP type and nesting.

3.1. Method

3.1.1. Participants

Sixty members of the MIT community participated in this study. Participants took approximately 20–25 min to complete the survey and were paid $4.00 to do so. Most spent an additional 20–25 min participating in an unrelated self-paced reading experiment during the same trip to the lab.

3.1.2. Materials

Experiment 2 tested 36 items in a 2×6 design crossing structure (nested vs. right branching) with subject NP type (indefinite description, definite description, full name, first name, third pronoun or first/second pronoun). In this experiment, the phrase "according to *name*" was added to the beginning of the test sentences in the third pronoun conditions, so that the name could serve as an antecedent for the pronoun. Of the names used in the full name condition, seven were the names of corporations and 23 were full names of famous people. Six were single names such as "Madonna" or "Shakespeare". A sample item from the experiment is included in (8):

(8) a. Nested
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.
b. Right branching
{You, The reporter, A reporter, Bill Clinton, Brad} praised the government assistance program which had saved the old lady who did not have enough money to heat her house.
c. Nested, third pronoun

(*continued*)

According to Brad, the old lady who the government assistance program which he praised had saved did not have enough money to heat her house.

d. Right branching, third pronoun

According to Brad, he praised the government assistance program which had saved the old lady who did not have enough money to heat her house.

The questionnaire in Experiment 2 was made up of 36 target items, 30 items from an unrelated experiment and 54 additional fillers. Conditions 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.

3.1.3. Procedure

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

3.2. Results

The means for all conditions are presented in Fig. 3. A 2×6 repeated measures analysis of variance (ANOVA) comparison showed a main effect of structure type, with the right branching structures rated significantly easier than the nested structures ($F1(1, 59) = 264$, $\mathrm{MSe} = 0.91$, $P < 0.001$; $F2(1, 35) = 333$, $\mathrm{MSe} = 0.43$, $P < 0.001$). There was also a main effect of NP type ($F1(5, 295) = 6.61$, $\mathrm{MSe} = 0.33$, $P < 0.001$; $F2(5, 175) = 5.44$, $\mathrm{MSe} = 0.24$, $P < 0.001$), and an interaction between structure and NP type ($F1(5, 295) = 3.29$, $\mathrm{MSe} = 0.30$, $P < 0.01$; $F2(5, 175) = 2.48$, $\mathrm{MSe} = 0.24$, $P < 0.05$). In order to determine the reason for this interaction, post-hoc contrast tests were carried out between the first/second pronoun condition and each of the rest of the conditions. This was done because in Experiment 1 the first/second pronoun condition was the only condition that differed from the other conditions. None of these individual contrasts revealed significant interactions.

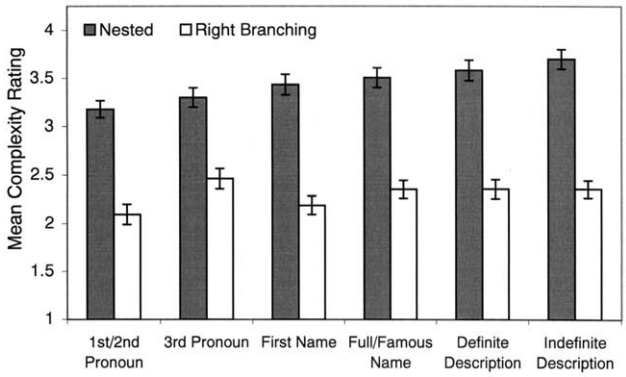

Fig. 3. Mean complexity ratings for Experiment 2.

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 $= 0.35$, $P < 0.005$) and by items ($F2(5, 175) = 6.31$, MSe $= 0.21$, $P < 0.005$). The same ANOVA performed on the right branching conditions showed a significant effect by participants ($F1(5, 295) = 3.68$, MSe $= 0.28$, $P < 0.05$), but the effect did not reach significance in the items analysis ($F2(5, 175) = 2.24$, MSe $= 0.28$, $P > 0.1$) 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.

Finally, repeated measures contrasts were performed to test whether there was a significant relationship between the Givenness Hierarchy and the complexity ratings. A polynomial contrast test indicated that a monotonically increasing trend accounted for a significant amount of the variance in the nested conditions ($t(59) = 5.67$, $P < 0.001$, $r_{\text{contrast}} = 0.59$; $t(35) = 6.77$, $P < 0.001$, $r_{\text{contrast}} = 0.75$ with Bonferroni corrections), but not the right branching conditions ($t(59) = 1.55$, $P > 0.2$, $r_{\text{contrast}} = 0.20$; $t(35) = 1.49$, $P > 0.2$, $r_{\text{contrast}} = 0.24$ with Bonferroni corrections). Thus, as the NP type of the innermost subject became more peripheral, complexity increased in the nested sentences but not in the right branching sentences. A repeated measures contrast test encompassing all conditions was also done, assigning low, even weights to the right branching conditions and higher, increasing weights to the nested conditions. This test directly compared the data to the pattern predicted by the continuous distance metric. The predicted pattern accounted for a highly significant amount of the variance in the data ($t(59) = 16.53$, $P < 0.001$, $r_{\text{contrast}} = 0.91$; $t(35) = 18.38$, $P < 0.001$, $r_{\text{contrast}} = 0.95$ with Bonferroni corrections).

3.3. Discussion

As predicted by both versions of the DLT, effects of NP type were stronger in the nested conditions than in the right branching conditions. This result supports a locality-based account of the difficulty of processing nested structures. In addition, the results of Experiment 2 support a continuous version of the DLT's referent-based integration cost over a binary version. The data suggest that complexity is sensitive to gradations of status between new and old, such as proposed in the Givenness Hierarchy in (8). These data support a referential processing theory based on accessibility and NP type, consistent with referent access models proposed by Garrod et al. (1994) and Myers and O'Brien (1998).

Interestingly, the indefinite condition was numerically, if non-significantly, more difficult than the definite description condition. This is consistent with the hypothesis that indefinites are peculiar in restrictive RCs. It also suggests there was no cost for accommodation. This lack of cost could be due to the fact that each sentence in this experiment defined a separate discourse. Accommodation is necessary in normal discourse, in which factors such as relevance and communicative intentions are important (Sperber & Wilson, 1995), but may not have been necessary in this experiment.

4. Experiment 3

Consistent with the continuous integration metric version of the Gibson (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 could potentially also be accounted for by the Bever (1970, 1974) complexity hypothesis (see Gordon, Hendrick, & Johnson, 2001 for new support of this hypothesis). The relevant prediction of Bever's complexity theory is that nested sentences with similar subject NPs are more difficult to process than ones with dissimilar subject NPs. In particular, doubly nested sentences with a pronoun in one of the subject positions may be less complex than doubly nested sentences with three definite descriptions because the pronoun makes the series of three NPs dissimilar.

Experiment 3 was designed to test whether subject similarity is sufficient to account for nesting complexity effects. The materials for this experiment consisted of doubly nested RC structures whose subject NPs and verbs were the same across conditions, with one quantifier and two definite descriptions rotated through the three subject positions in the sentence, as shown in (9):

(9) a. Doubly nested, Outer
Everyone who the journalist who the photographer met liked was at the party.
b. Doubly nested, Middle
The photographer who everyone who the journalist met liked was at the party.
c. Doubly nested, Inner
The journalist who the photographer who everyone met liked was at the party.

In (9a) the quantifier is in the matrix subject position, in (9b) it is in the middle subject position, and in (9c) it is in the most embedded subject position. Because each condition has the same three subject NPs, the Bever (1970, 1974) similarity-based complexity hypothesis predicts no differences among the three conditions. The DLT, on the other hand, predicts a difference if the quantifiers and the definite descriptions require different amounts of referential processing. We hypothesized that the process of instantiating or identifying a set corresponding to the noun associated with the quantifier contributes to the amount of discourse processing required by the quantified NP. Most discourse contexts – and all of the contexts in the materials for this experiment – involve people. Hence, we hypothesized that processing a quantifier that quantifies over some, all, or none of the people in the context (e.g. "everyone" in (9)) would involve a lower discourse cost than introducing a new referent via a definite description with a more specified noun, such as "the photographer" or "the journalist". All of the quantifiers in this experiment ("everyone", "everybody", "no one" and "many people") quantified over the people in the context. The definite descriptions in this experiment consisted mostly of occupation names.

Given these assumptions, the DLT predicts that the condition with the quantifier as the outermost subject (9a) should be most complex, the condition with the quantifier as the middle subject (9b) should be less complex, and the condition with the quantifier as the innermost subject (9c) should be least complex. This prediction results from the fact that four dependencies cross the innermost subject position (two subject-verb dependencies

and two filler-gap dependencies), two dependencies cross the middle subject position and no dependencies cross the outermost subject position. When NPs with a higher discourse cost (hypothesized to be the definite descriptions in this experiment) interrupt more dependencies, sentences become more complex.

It is worth noting that these same predictions could be made by complexity theories concerned solely with the processing of quantifiers. It could be the case that having more material to divide between the restrictor and the scope makes a structure more complex.[3] It could also be that quantifiers with more restriction (e.g. additional RCs in this case) are more difficult to process. Both of these theories predict, like the DLT, that (9a), in which the restrictor contains the most RCs, should be most complex, followed by (9b), with (9c) least complex. Critically for the design of the experiment, these theories all predict a different pattern of complexity results than Bever's similarity-based hypothesis alone.

4.1. Method

4.1.1. Participants

Sixty members of the MIT community were paid participants in this study. Participants took approximately 20–25 min to complete the survey and were paid $5.00 to do so. About half the participants also participated in a 20–25 min unrelated self-paced reading study during the same trip to the lab. The other half of the participants participated only in this study.

4.1.2. Materials

Experiment 3 tested 18 items in a 2×3 design crossing structure (doubly nested vs. singly nested) with quantifier position (outer, middle or inner subject position). The singly nested conditions provided a plausibility control for the doubly nested conditions. Singly nested versions were used as plausibility controls rather than fully right branching versions in order to keep meaning as constant as possible across the two structures. In the conditions with the quantifier 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 quantifier in the middle subject position, the inner two clauses were extracted and conjoined to the remaining simple sentence. The doubly nested versions of an example item are given in (9) above. The singly nested versions of (9) are presented in (10):

(10) a. Singly nested, Outer
The photographer met the journalist, and everyone who the journalist liked was at the party.
b. Singly nested, Middle
The photographer was at the party, and everyone who the journalist met liked the photographer.
c. Singly nested, Inner

[3] We thank an anonymous reviewer for suggesting this possibility.

(*continued*)

Everyone met the photographer, and the journalist who the photographer liked was at the party.

The questionnaire consisted of 18 target items, 20 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 such 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.

4.1.3. Procedure

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

4.2. Results

The means for each condition are shown in Fig. 4. A repeated measures 2×3 ANOVA revealed a significant interaction between the factors of structure and quantifier placement ($F1(2, 118) = 11.32$, MSe $= 0.30$, $P < 0.001$; $F2(2, 34) = 10.82$, MSe $= 0.09$, $P < 0.001$). There was also a significant main effect of structure, such that the doubly nested conditions were rated more complex than the singly nested conditions ($F1(1, 59) = 162.06$, MSe $= 1.07$, $P < 0.001$; $F2(1, 17) = 516.78$, MSe $= 0.10$, $P < 0.001$). No effect of quantifier position was apparent at this level (all $F < 1$). A repeated measures ANOVA across the three doubly nested conditions showed a significant difference among the conditions ($F1(2, 118) = 7.69$, MSe $= 0.25$, $P < 0.001$; $F2(2, 34) = 4.88$, MSe $= 0.12$, $P < 0.05$). 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 $= 0.27$, $P < 0.001$; $F2(1, 17) = 19.08$, MSe $= 0.06$, $P < 0.001$) or the middle subject condition, although this effect was marginal in the items analysis ($F1(1, 59) = 6.86$, MSe $= 0.26$, $P < 0.05$;

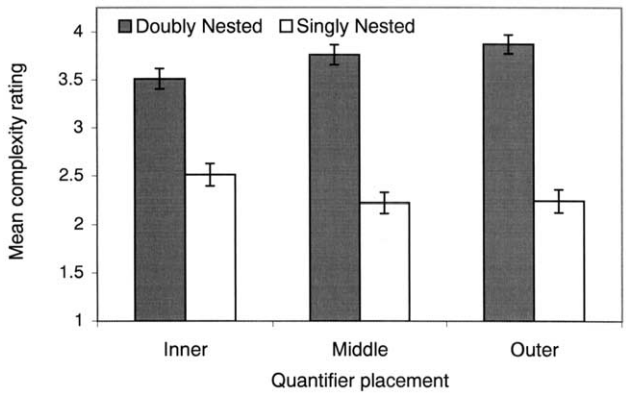

Fig. 4. Mean complexity ratings for Experiment 3.

$F2(1,17) =) = 3.53$, MSe $= 0.15$, $P = 0.08$). Finally, the middle subject condition was numerically easier than the outermost subject condition, but this difference did not reach significance (all $F < 1.5$, all $P > 0.2$). Interestingly, the complexity pattern for the singly nested structures was different than for the doubly nested structures. The singly nested "inner" condition was rated numerically more complex than the other two. Post-hoc comparisons corrected for multiple tests indicated that this difference was marginally significant: "inner" vs. "outer" ($F1(1,59) = 5.37$, MSe $= 0.40$, $P = 0.07$; $F2(1,17) = 6.52$, MSe $= 0.10$, $P = 0.06$) and "inner" vs. "middle" ($F1(1,59) = 7.17$, MSe $= 0.35$, $P < 0.05$; $F2(1,17) = 4.68$, MSe $= 0.16$, $P = 0.14$).

4.3. Discussion

The overall pattern of results in Experiment 3 suggests that subject similarity is not sufficient to account for nesting complexity effects. In particular, this theory predicted no differences among the doubly nested conditions, but significant differences among these conditions were observed. The data pattern was as predicted by the DLT, under the assumption that in a null context, quantified NPs like "everyone" and "everybody" have a lower discourse processing cost than definite NPs like "the journalist". This theory predicted that having the quantified NP in the more embedded positions in the doubly nested structures would lead to less processing difficulty, which was the observed pattern. In addition, this pattern of data was not likely due to differences in plausibility of the thematic relationships between NPs and verbs, because a very different pattern of results was observed in the singly nested structures, which included the same thematic relationships as the doubly nested versions.

The pattern of results in the doubly nested structures is also consistent with theories attributing greater complexity to quantifiers that are restricted by more RCs. Suggestive evidence against these alternative theories comes from an examination of the results for the singly embedded structures. The singly embedded structures in which the quantifier was modified by a RC were rated marginally *less* complex than the condition in which the quantifier was not modified by an RC. This pattern of results is opposite from what is predicted by a theory attributing processing difficulty to quantifiers with more restriction. In fact, the pattern of results in the singly embedded structures was not predicted by any of the theories that we considered here. It is possible that the observed pattern in the singly embedded structures may be due to coherence differences among the conditions.

5. Experiment 4

Experiments 4 and 5 were designed to test the DLT hypotheses from Experiments 1 and 2 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; Grodner et al., 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.

The items in Experiment 4 consisted of RC and complement clause (CC) structures,

where the subject of the embedded clause was either a first/second pronoun or a definite description, as in (11):

(11) a. Relative clause
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.
b. Complement clause
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 (11). For simplicity, following Gibson (1998, 2000) we will count one unit of cost for each new discourse referent which has been processed since the attachment site was introduced. At the embedded verb, "pushed", both the CC and the RC conditions require an integration to the embedded subject, "you" or "the boy". This integration has one unit of cost because "pushed" is a new discourse referent. The RC structure has an additional integration at "pushed": the integration of the object-gap with the relative pronoun. In the RC pronoun condition this integration crosses the discourse referents "pushed" and "you", but since a referent for "you" is provided in the discourse, the integration requires only one unit of cost. In the RC definite description condition the filler-gap integration crosses "pushed" and "the boy", both of which indicate new referents, leading to an integration cost of two units. Hence, the DLT predicts that at the verb "pushed" the CC conditions will have an integration cost of one unit each, the RC pronoun condition will have a cost of two units and the RC definite description condition will have a cost of three units. Thus, the DLT predicts a main effect of clause type (RC vs. CC), a main effect of NP type (pronoun vs. definite description) and an interaction between the two factors.

At the main verb in the RC conditions ("got upset" in (11)), the matrix subject and verb must be integrated. This integration crosses the intervening RC, which contains one more new referent in the definite description condition than in the pronoun condition. Thus, the DLT predicts a main effect of NP type at the main verb, such that the RC definite description condition will be read more slowly than the RC pronoun condition.

Finally, the DLT also predicts reading time differences due to differences in syntactic storage between the RC and CC structures during the processing of the embedded subject "you/the boy" and the auxiliary verb and the adverb "had accidentally". Integrations are the same in the RC and CC structures in these regions, but there are more syntactic predictions in the RC structures than in the CC structures, because the RC requires storing the predictions of a verb and a gap site, while the CC does not. As observed by Grodner, Gibson, and Tunstall (2002), reading times are longer over portions of sentences that require the storage of more syntactic predictions.

5.1. Method

5.1.1. Participants

Eighty members of the MIT community participated in this study in exchange for an $8

payment. Participants took approximately 20–25 min 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 1 h in the lab, with a short break between the two sessions of self-paced reading.

5.1.2. Materials

Four versions of 20 items were constructed in a 2×2 design crossing structure (RC, CC) and NP type (first/second pronoun, definite description) as exemplified in (11) above. 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 made it unlikely that reading times at the embedded verb 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 12 it described an action. The relative pronoun in the RC conditions was always "who" and the subject of the RC was always an animate individual or a set of humans. A prepositional phrase (PP) modifying the first verb intervened between the two verbs in the RC condition to eliminate the possibility of reading time spillover from the first verb to the second verb. In 13 of the items, the PP was three words long, in five the PP was four words long, and in the remaining two the PP was five and six words long. The main verb region in the RC conditions was always two words: usually an auxiliary and a predicate adjective, but sometimes an auxiliary and progressive verb form. A complete list of items is included in Appendix D.

5.1.3. Procedure

Participants performed self-paced reading in a word-by-word moving window display (Just, Carpenter, & Woolley, 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. The time between bar-presses was recorded. After each sentence, participants answered a yes/no comprehension question about its content. 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 natural a pace 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.

5.2. Results

One participant's data were excluded from analysis because he fell asleep during testing. Three additional participants' data were excluded due to accuracy rates of 70% or

Table 1
Percentage of comprehension questions answered correctly for each condition in Experiment 4

Condition	Percentage correct
RC full NP	90.4
RC pronoun	97.5
CC full NP	88.2
CC pronoun	89.1

lower on the comprehension questions pertaining to the experimental sentences. All other participants had accuracy rates of at least 75%. Table 1 presents the percentage of questions answered correctly for each condition.

Questions about the pronoun conditions were answered more accurately than questions about the full NP conditions ($F1(1,75) = 7.23$, MSe $= 0.02$, $P < 0.01$; $F2(1,19) = 5.68$, MSe $= 0.01$, $P < 0.05$). Questions about the RC conditions were answered more accurately than questions about the CC conditions, but this difference was significant only in the participants analysis ($F1(1,75) = 15.05$, MSe $= 0.01$, $P < 0.001$; $F2(1,19) = 2.8$, MSe $= 0.02$, $P = 0.11$). This difference probably indicates that the questions for the RC conditions were easier than those for the CC conditions, as questions differed for different conditions of the same item. Finally, there was an interaction between structure and NP type in the participants analysis ($F1(1,75) = 4.48$, MSe $= 0.02$, $P < 0.05$) but not in the items analysis ($F2(1,19) = 2.48$, MSe $= 0.01$, $P = 0.13$).

Fig. 5 presents mean reading times per word or per word in a phrase 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, eliminating less than 1% of the remaining data. Repeated measures ANOVAs were performed on regions for which the DLT made predictions. Statistical tests were performed on both unadjusted reading times and reading times adjusted for word length following the procedure proposed by Ferreira and Clifton (1986) and Trueswell, Tanenhaus, and Garnsey (1994). We report only the statistical tests for unadjusted times; the results were similar for the statistical tests that were performed on length-adjusted times.

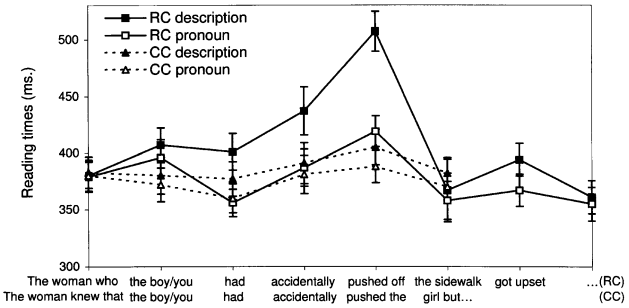

Fig. 5. Mean reading times for Experiment 4.

The words preceding the subject of the embedded clause were not analyzed because they comprised different syntactic structures and the DLT made no interesting predictions in this region. At the subject of the embedded clause, the RC conditions were read more slowly than the CC conditions ($F1(1,75) = 7.66$, MSe $= 6555$, $P < 0.01$; $F2(1,19) = 6.9$, MSe $= 2151$, $P < 0.05$). This effect of structure is predicted by the storage component of the DLT. No other effects reached significance in this region. At the next word, the auxiliary "had", there was an effect of NP type, such that the description conditions were read more slowly than the pronoun conditions ($F1(1,75) = 18.10$, MSe $= 4150$, $P < 0.001$; $F2(1,19) = 15.60$, MSe $= 1931$, $P = 0.001$). This effect probably reflects spillover from the previous word ("boy" or "you" in (11)). No other effects reached significance on this word.

The next two regions consisted of (1) the adverb, and (2) the embedded verb and the following word, a high frequency function word (e.g. "off" or "the" in (11)). The embedded verb and following word were analyzed together because reaction times from points of high complexity often spill over into following words in self-paced reading. The RC conditions were read more slowly than the CC conditions in both regions (adverb: $F1(1,75) = 9.73$, MSe $= 5323$, $P < 0.005$; $F2(1,19) = 3.35$, MSe $= 1829$, $P = 0.08$; embedded verb region: $F1(1,75) = 29.69$, MSe $= 11418$, $P < 0.001$; $F2(1,19) = 13.43$, MSe $= 5285$, $P < 0.005$). In addition, the description conditions were read more slowly than the pronoun conditions in both regions (adverb: $F1(1,75) = 13.29$, MSe $= 5244$, $P = 0.001$; $F2(1,19) = 6.79$, MSe $= 4382$, $P < 0.05$; embedded verb region: $F1(1,75) = 23.43$, MSe $= 8894$, $P < 0.001$; $F2(1,19) = 9.95$, MSe $= 5805$, $P = 0.005$). Furthermore, there was an interaction between NP type and structure in both regions, due to the RC description condition being slower than the other conditions (adverb: $F1(1,75) = 5.11$, MSe $= 6132$, $P < 0.05$; $F2(1,19) = 9.17$, MSe $= 1078$, $P < 0.01$; embedded verb region: $F1(1,75) = 11.7$, MSe $= 8385$, $P = 0.001$; $F2(1,19) = 13.7$, MSe $= 2715$, $P < 0.005$). A repeated measures contrast test on the embedded verb region using weights corresponding to DLT integration cost predictions for the embedded verb shows that the DLT predictions capture a significant amount of the variance in the data ($t(75) = 5.79$, $P < 0.001$, $r_{\text{contrast}} = 0.56$; $t(19) = 4.57$, $P < 0.001$, $r_{\text{contrast}} = 0.72$).

The DLT makes no predictions about the remainder of the CC conditions, but predicts a main effect of NP type at the main verb in the RC conditions, such that the description condition should be read slower than the pronoun condition. Such an effect is apparent on the main verb ($F1(1,75) = 6.58$, MSe $= 4141$, $P < 0.05$; $F2(1,19) = 4.44$, MSe $= 2733$, $P < 0.05$), and is separate from the effect at the embedded verb, as the effect is not apparent on the NP intervening between the verbs (all $F < 1.6$, all $P > 0.2$).

5.3. Discussion

The results of Experiment 4 provide support for the DLT's discourse-based integration cost metric. The DLT's integration cost metric predicts that at the embedded verb the RC definite description condition should be slowest, the RC pronoun condition should be faster, and the CC conditions should be fastest and not differ. This pattern was largely reflected in the reading time data. The DLT also predicted that the RC definite description condition should be slower than the RC pronoun condition at the main verb. This pattern

was observed, and could not have been due to spillover from processing at the embedded verb, because there were no reading time differences on the region separating the verbs.

Interestingly, the interaction that the DLT predicted at the embedded verb began to appear at the adverb immediately preceding the verb. This may indicate that readers sometimes assigned the object position thematic role associated with the filler "who" before they processed the embedded verb, where the actual thematic role becomes known. This effect may be related to an effect observed by Boland, Tanenhaus, Garnsey, and Carlson (1995) and Altmann and Kamide (1999) that in certain predictive situations, people can assign thematic roles to fillers before the role-assigning verb is processed. Having read the auxiliary "had" and the adverb, the reader is in a situation highly predictive of a verb with at least two arguments. Following the discussion by Boland et al. (1995) of provisional interpretation, at this point in the sentence the reader may project a prototypical verb with a generalized theme theta role and begin interpretation before the verb is actually read.

Reading times in Experiment 4 were also mostly consistent with the DLT's storage cost prediction that reading times over the RC should be slower than reading times over the CC, because the RC requires storing the predictions of a verb and gap site while the CC does not.

6. Experiment 5

Experiment 5 was designed to distinguish between the binary and continuous DLT integration cost metrics using self-paced reading. Experiment 5 tests whether graded effects of referent accessibility are apparent in reading times at the conclusions of long-distance dependencies, as predicted by the DLT.

6.1. Method

6.1.1. Participants

Eighteen Northeastern undergraduates and 30 members of the MIT community were participants in this study, which took approximately 20 min 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, no participants in Experiment 5 had previously participated in Experiment 2. All of the participants also participated in another experiment during the same testing session and spent approximately 1 h 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.

6.1.2. 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 first/second pronoun, a famous name, a definite description and an indefinite description, as in (12). In 17 of the 24 items the famous name was a full name, in four items it was the name of a company and in three items it was a first name such as "Madonna".

(12) The consultant who {we, the chairman, a chairman, Donald Trump} called advised wealthy companies about tax laws.

In two items the embedded verb was followed by a particle, which was grouped with the verb in analysis.

In order to control for plausibility differences among the conditions, 36 potential items were pre-tested in a 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 "{We, the chairman, a chairman, Donald Trump} called the consultant". Only the pronoun, famous name and definite description conditions were tested, because the indefinite NP conditions were very similar to the definite description conditions. Participants 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. Twenty-four items that were matched for their plausibility ratings across the three conditions were chosen from the original set of 36. The mean plausibility of the conditions was 2.71 for the pronoun condition, 2.65 for the definite condition and 2.89 for the famous name 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. A complete list of items is included in Appendix E

6.1.3. Procedure

The procedure was the same as in Experiment 4.

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 2 presents the percentages of comprehension questions answered correctly in each condition. In Experiment 5, unlike in Experiment 4, questions were the same for every condition of an item. The pronoun and famous name conditions were comprehended most often, while the definite description condition was comprehended less often and the indefinite condition the least often. Qualitatively, these percentages mirror the complexity results in Experiment 2, except that there is no numerical difference between the pronoun and famous name conditions. 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 ($F1(3, 141) = 2.76$, $\mathrm{MSe} = 0.01$, $P < 0.05$; $F2(3, 69) = 2.22$, $\mathrm{MSe} = 0.01$, $P = 0.09$).

Fig. 6 presents the mean reading times per word 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 1% of the remaining data. We report only the statistical tests for unadjusted times; the results were similar for the statistical tests that were performed on length-

Table 2
Percentage of comprehension questions answered correctly for each condition in Experiment 5

Condition	Percentage correct
Pronoun	94.1
Famous name	94.4
Definite	91.3
Indefinite	88.9

adjusted times. There were no reliable differences between participant groups (Northeastern, MIT), so we present our analysis over both groups together.

All of the conditions contained the same words in the first three words of the sentence (e.g. "the writer who"), so no analyses were performed in this region. At 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 $= 2521$, $P < 0.005$; $F2(3,69) = 2.71$, MSe $= 5782$, $P = 0.05$). The pattern of means likely reflects the frequency and length of words in the region. Pronouns, which were one to three letters long, were read fastest, while definite and indefinite descriptions had intermediate average reading times. Famous names (e.g. "Donald Trump") were read most slowly. At the embedded verb (e.g. "called"), reading times differed significantly by participants but not by items ($F1(3,138) = 3.65$, MSe $= 10188$, $P < 0.05$; $F2(3,69) = 1.92$, MSe $= 13164$, $P = 0.13$). Numerically, the pattern at this region was close to that predicted by the DLT. The indefinites were slowest, the definite descriptions were slightly faster, and the pronoun and famous name conditions were numerically fastest. At the main verb of the sentence ("advised") the means were in the order predicted by the DLT, but the overall ANOVA did not show a significant difference (all $F < 2.5$, all $P > 0.09$). In order to capture possible spillover effects, we also analyzed the main verb together with the following word, the first word from the following NP. On this combined

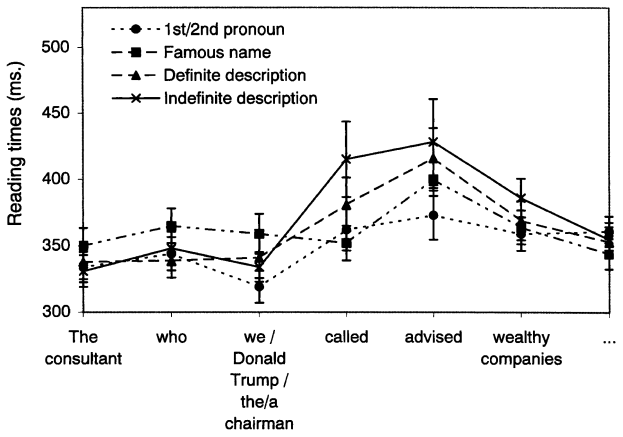

Fig. 6. Mean reading times for Experiment 5.

region, a repeated measures ANOVA showed differences in the predicted direction ($F1(3,138) = 3.65$, MSe $= 5384$, $P < 0.05$; $F2(3,69) = 2.65$, MSe $= 5684$, $P = 0.06$). Planned comparisons showed that the definite description and first/second pronoun conditions differed marginally ($F1(1,47) = 3.61$, MSe $= 3444$, $P = 0.06$; $F2(1,23) = 3.07$, MSe $= 2816$, $P = 0.09$), while the first/second pronoun and indefinite conditions differed significantly ($F1(1,47) = 9.87$, MSe $= 5774$, $P < 0.005$; $F2(1,23) = 8.72$, MSe $= 2958$, $P < 0.01$). The other comparisons revealed no significant differences, although they were numerically in the direction predicted by the DLT.

The final region of the sentence showed no reliable differences among the conditions (all $F < 2$).

A repeated measures contrast test at the region made up of the matrix verb and following word showed that the increase in reading times composed a significant monotonically increasing trend ($t(47) = 3.18$, $P < 0.005$, $r_{\text{contrast}} = 0.42$; $t(23) = 3.14$, $P < 0.005$, $r_{\text{contrast}} = 0.55$). This trend mirrors the trend in Experiment 2, such that reading times at the main verb were longer for sentences whose embedded subject NPs were from the more peripheral end of the Givenness Hierarchy. A similar pattern was evident at the embedded verb in the pronoun, definite and indefinite conditions, but the famous name condition was read slightly faster than the pronoun condition instead of falling between the pronoun and definite description conditions.

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. 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.

One issue that has not been discussed much thus far is the role that lexical frequencies might play in accounting for some of our results. In Experiment 5, as in most of the other experiments in this paper, varying the type of an NP not only changed the amount of referential processing necessary at the NP, but also the frequency of the NP. This makes it possible that the results we report might actually be due to frequency differences between experimental conditions rather than structural and referential factors. Under perhaps the simplest frequency-based hypothesis, the frequency of all the words back to an integration site could affect the difficulty of the integration. This account could account for the observed differences at the verbs for comparisons involving pronouns in the above experiments, because pronouns are generally more frequent than names or common nouns. However, it is unlikely that this account could explain the effects involving comparisons that do not involve pronouns. For example, it was shown that sentences with proper names (e.g. "Donald Trump") in their embedded subject positions are easier to process than sentences with full definite NPs (e.g. "the chairman"). Although it is hard to know what the right corpus is to search, the common nouns that were used are likely more frequent than the proper names that were used. Indeed, in the plausibility survey that we conducted prior to the on-line study, the proper name condition was rated as numerically (non-

significantly) *less* plausible than the full definite NP condition (proper name mean 2.89, full definite NP mean 2.65 on a 7-point scale where 1 was natural and 7 was unnatural). If the plausibility ratings are sensitive to word frequency, then, if anything, the definite descriptions are more frequent than the proper names in the items that were tested.

A second reason that it seems unlikely that lexical frequencies alone are driving the results is that similar integration cost effects have been shown in experiments which manipulate the context before a target sentence, leaving the words between two endpoints of an integration constant (Warren, 2001). For example, in one experiment the accessibility of a referent was varied by introducing the referent explicitly in a previous sentence, by introducing a related set in a previous sentence or by having no prior context for the referent. Reading times at the verbs varied according to the difficulty of establishing a referent for the embedded subject NP: the more difficult it was to establish a referent for this NP, the slower people read the embedded and main verbs in the sentences. These results were as predicted by the DLT, but cannot be accounted for by a complexity theory based on lexical or structural frequencies.

7. General discussion

The experiments presented in this paper (1) provide new evidence about the resource demands of discourse processing during sentence processing, (2) support and extend the DLT, a theory of computational resource usage in 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, we discuss each of these claims in turn.

First, the results of the experiments suggest that there is a cost associated with identifying a referent for an NP, consistent with earlier research (e.g. Garrod & Sanford, 1982; Murphy, 1984). Furthermore, the findings demonstrate that the location and amount of processing cost are dependent on sentence structure. In addition, these data extend the findings from 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 as predicted by an independently established accessibility hierarchy demonstrates that referring form is an important indicator of discourse status to the referential processing system.

Second, these experiments provided evidence supporting many of the basic assumptions of the Gibson (1998) DLT and led to the theory's extension. The evidence presented here indicates that a distance metric based on continuous discourse costs is a better fit to the data than a metric based on binary discourse costs, such as the one proposed in Gibson (1998). Data from Experiments 2 and 5 suggest that the metric should take into account the ease of establishing a referent for an NP, including both the search for a representation and the instantiation of a new representation if necessary. In light of these findings, one way to improve the Gibson (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.

Third, evidence presented in this paper also bears on the architecture of the human sentence processing mechanism (HSPM). In particular, the results suggest that the resource demands of referential processing can decrease the resources available for subsequent structural processing. The strongest evidence for this is the finding in Experiments 4 and 5 that syntactic integration difficulty at the matrix verb is affected by the discourse status of intervening NPs. While some integrations at the embedded verb involved the NP whose status was changed, integrations at the matrix verb only involved elements whose discourse status was constant across conditions. Thus, the differences found at the matrix verb suggest that the resources used in syntactic processing and discourse processing are not independent. This is important because it makes it unlikely that the HSPM 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.

Acknowledgements

We would like to thank the following people for helpful discussions on presentations of portions of this work: the audience of the 12th CUNY Human Sentence Processing Conference, the audience of the 6th Architectures and Mechanisms of Language Processing Conference, Lyn Frazier, Simon Garrod, Daniel Grodner, Barbara Grosz, Edson Miyamoto, Neal Pearlmutter, Steve Pinker, Molly Potter, San Tunstall, Duane Watson, Ken Wexler, the MIT/NEU sentence processing lab group and two anonymous reviewers. Funding for the project was provided by a grant awarded by the NSF to the second author, award number SBR-9729037. An earlier version of this work was a part of the first authors PhD Thesis, copyright MIT. This version appears here with MIT's permission. While at MIT, the first author was supported by the Research Training Grant "Language: Acquisition and Computation" awarded by the NSF to the Massachusetts Institute of Technology (DIR 9113607), by a NSF graduate research fellowship and by a Burnett Fellowship.

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 third-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 (Section 4.1.2).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 amendment 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.14b. The 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.

References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73* (3), 247–264.

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*. PhD dissertation, Stanford University, Palo Alto, CA.

Bever, T. G. (1970). The cognitive basis of linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

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* (pp. 173–200). Washington, DC: Hemisphere.

Boland, J., Tanenhaus, M., Garnsey, S., & Carlson, G. (1995). Verb argument structure in parsing and interpretation: evidence from wh-questions. *Journal of Memory and Language*, *34*, 774–806.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. N. Li (Ed.), *Subject and topic* (pp. 25–55). New York: Academic Press.

Chafe, W. (1987). Cognitive constraints on information flow. In R. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 21–51). Philadelphia, PA: John Benjamins.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. 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* (pp. 269–321), Vol. 2. New York: Wiley.

Clancy, P. M. (1996). Proper names as a referential option in English conversation. In B. Fox (Ed.), *Studies in anaphora* (pp. 95–143). Philadelphia, PA: John Benjamins.

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* (pp. 320–358). Cambridge: 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.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, *63*, 805–855.

Enç, M. (1983). Anchored expressions. In: M. Barlow, D. Flickinger & M. Westcoat (Eds.), *Proceedings of the west coast conference of formal linguistics* (pp. 79–88), Vol. 2. Standford, CA: Stanford Linguistics Association.

Ferreira, F., & Clifton Jr., C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language*, *66* (2), 297–316.

Frank, R. (1992). *Syntactic locality and tree-adjoining grammar: grammatical, acquisition and processing perspectives*. PhD thesis, University of Pennsylvania, Philadelphia.

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

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* (pp. 675–698). San Diego, CA: Academic Press.

Gibson, E. (1991). *A computational theory of human linguistic processing: memory limitations and processing breakdown*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

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* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2001). Reading relative clauses in English. MIT manuscript.

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.

Givón, T. (1979). *On understanding grammar*. New York: Academic Press.

Givón, T. (1983). *Topic continuity in discourse*. Amsterdam: John Benjamins.

Givón, T. (1984). *Syntax*, I. 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*, *27*, 1411–1423.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language*, 26(2).

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

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

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

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.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

Kac, M. B. (1981). Center-embedding revisited. *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, T. Janssen & M. Stokhof (Eds.), *Formal methods in the study of language* (pp. 277–322). Amsterdam: Mathematisch Centrum.

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* (pp. 419–491), Vol. 2. New York: Wiley.

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

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*, 574–590.

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

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, Handbook in perception and cognition* (pp. 217–262), Vol. 11. San Diego, CA: Academic Press.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language*, *33*, 285–318.

Warren, T. (2001). *Understanding the role of referential processing in sentence complexity*. PhD thesis, MIT, Cambridge, MA.

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