Conceptions of Limited Attention and Discourse Focus*

Barbara J. Grosz†
Harvard University

Peter C. Gordon‡
University of North Carolina

Walker (1996) presents a cache model of the operation of attention in the processing of discourse as an alternative to the focus space stack that was proposed previously by Grosz and Sidner (Grosz 1977a; Grosz and Sidner 1986). In this squib, we present a critical analysis of the cache model and of Walker's supporting evidence from anaphora in discourses with interruptions and from informationally redundant utterances. We argue that the cache model is underdetermined in several ways that are crucial to a comparison of the two models and conclude that Walker has not established the superiority of the cache model. We also argue that psycholinguistic evidence does not support the cache model over the focus stack model.

1. Introduction

Attention constrains the structure and processing of discourse. This fact has been important to computational research on discourse since the work of Grosz (1977b). A recent article by Walker (1996) argues that the attentional mechanism has limited capacity, that this limited capacity determines the accessibility of information in discourse processing, and that certain linguistic behavior can only be explained in terms of this limited capacity. Walker presents as an alternative to the focus space stack previously proposed to model global attentional state (Grosz 1977a; Grosz and Sidner 1986) a cache model in which linear recency and a highly constrained cache capacity play primary roles. As critical evidence, Walker presents an analysis of anaphora in discourses with interruptions and of informationally redundant utterances (IRUs). In addition, she cites psychological evidence on the limited capacity of human information processing. In this response, we discuss the relationship between the focus-space stack model and the cache model, examine Walker's evidence with respect to the two models, and review psychological evidence concerning the contributions of limited capacity and recency to the understanding of discourse. We identify problems with Walker's analysis and deficiencies in the cache model.

2. Attentional State, Focus, and Limited Capacity

Walker argues that "[t]he notion of a cache in combination with main memory, *as is standard in computational architectures*, is a good basis for a computational model of human attentional capacity in discourse processing (emphasis added)" (page 258). Walker further claims that this model has advantages in explaining discourse phenomena over what she refers to as the "stack model."

* Preparation of this paper was supported by NSF grants IRI- 94-04756 and IIS-9811129.
† Division of Engineering and Applied Sciences, Cambridge, MA 02138
‡ Department of Psychology, Chapel Hill, NC 27599-3270

Within computer systems, the stack and cache address different concerns and levels of processing. The stack model draws on ideas in programming languages that work on evaluation of variables. The cache model is based on work on memory management. To examine Walker's claims, we first identify those features of each computational construct that play a role in the respective discourse models.

The use of a focus space stack as a model of global attentional state for discourse processing (Grosz 1977b) drew on the use of stacks by programming language interpreters and compilers to determine variable values. In interpreters and compilers, the stack determines the value a variable has when a processor is in a certain state. It provides implicit indexing and an ordering of access to values that is dependent on context. These two properties and the stack's ability to capture hierarchical relationships are the primary attributes adopted in the stack-based model of global discourse structure and processing.

Because Walker describes the cache model only generally and does not identify specific properties of caches that are relevant to her model, we first establish the basic roles of caches in computer systems and then identify those properties that seem most relevant to her claims. In computer systems architectures, the cache and main memory are elements of the memory hierarchy, each level of which has a different speed, size, and cost. Faster memory is "closer" to the processor allowing for quicker processing of information stored there. The term "cache" primarily refers to the smallest, fastest memory that is accessed first. Because the whole point of the cache is to speed up processing, cache management algorithms must be low cost. They are syntactic in the extreme, depending on physical properties of the system and making no reference to content. For discourse-processing algorithms, and for Walker's model, similar physical properties of human memory do not seem the issue and content (semantic and pragmatic) does matter.

Only two of the central memory-hierarchy questions for computer architectures (Patterson and Hennessy 1996) are relevant to the discourse issues Walker raises: replacement strategy and how information is found in the cache. Walker does not specify a replacement strategy, but suggests consideration of a modification of the "least recently used" strategy that accommodates what she calls "preferential retention." Preferential retention is not clearly defined in the squib; we are given no details on how items to be (preferentially) retained are identified nor on differences in treatment of retained items and those in the cache on some other basis. Preferential retention is said to depend on the intentions of the discourse participants, but exactly how is not specified. There is no discussion of how to identify the information relevant to a new intention that should be retained. Because intention-based preferential retention seems to be the main way Walker's cache model breaks out of a strict linear-recency approach, the lack of detail makes it difficult to ascertain exactly how it works, and more generally to establish how the model differs from the stack model.

Walker does not discuss methods for searching the cache although issues of locating information in the cache are important in computer systems design. Instead, the cache model assumes the cache is sufficiently limited in size that everything in it is "almost instantaneously accessible" (Walker 1996, 258). Walker also does not define "strategically retrieve" nor specify how main memory is searched when items are not found in the cache. Because main memory is part of the cache model (see the quotation that starts this section), without such information it is difficult to evaluate the model.

Hence, the main claims of the cache model seem to be that (1) the cache contains a small number of items; (2) as new entities are discussed in the discourse, entities not mentioned recently are removed from the cache to make room, and (3) remention and some (unspecified) connection with the current discourse intention cause an entity to

be retained longer in the cache than it would otherwise be. An important related claim is that (4) only items in the cache are available for various discourse processes like the inference of discourse relations (Walker 1996, 258).

Linear recency thus plays a much more substantial role in the cache model than in the stack model. In this context, claim (1) raises both computational and linguistic issues. From a computational point of view, the cache must be small enough to allow for instant retrieval of all items (or Walker must specify the retrieval part of the processing); from the linguistic point of view, the cache size must accommodate the facts of informationally redundant utterances—being small enough to explain when an IRU is required and large enough to explain when no IRU is needed.

The stack model makes specific claims about how intentions at the discourse level affect attentional state, providing the basis for the hierarchical structure. With respect to (3), Walker says that such intentions influence retention and retrieval in the cache model, but no details are given. Intentional connections might be identified and function identically to the intentional structure in the stack model (which would certainly greatly lessen the differences between the two models) or differently (in which case comparison would shed light on the appropriateness of the different models). An additional difference between the models is the cache model claim (4). The stack model claims only differences in speed and complexity of accessibility, not possibility of accessibility, among those items in the stack and those not.

3. Linguistic Evidence

In her squib, Walker discusses two types of linguistic evidence: pronominal reference following interruptions and informationally redundant utterances. We first examine this evidence from the stack model perspective showing how the data might be accounted for within that model; we then raise some questions about Walker's cache model explanations.

In earlier work (Grosz 1977a; Grosz and Sidner 1986) we have argued that it is important to distinguish between two levels of discourse structure and processing: global and local. A focus-space stack was proposed as a model for the global level. The main claims about its use in processing have been for handling definite descriptions (Grosz 1977a; Grosz 1981) and reasoning about intentional structure (Lochbaum 1998). The local level "refers to the influence of a listener's memory for the linguistic form of an utterance (the actual words and the syntactic structure) on his interpretation of a subsequent utterance" (Grosz 1977b).[1]

According to this theory of discourse structure, pronominal reference depends on the local level of attentional state, not the global level. Initial work on hierarchical discourse structure was motivated by examples of pronouns that were used to refer to entities in stacked focus spaces. This work could be read as suggesting that some memory for local attentional state was attached to each focus space. However, such an account would contradict the local nature of local attentional state, and we have more recently denied it (Grosz and Sidner 1997). A more satisfactory explanation of such pronoun uses has two components: specification of the information that indicates a shift in focus back to the attentional state of some previous discourse segment (typically more than an unstressed pronoun alone) and a determination of the

1 Sidner (1979) first provided algorithms that tied the local level with pronominal reference. In subsequent work we have defined a centering model for attentional state at this level (Grosz, Joshi, and Weinstein 1995) and have explored the ways in which pronominal reference and centering interact (Gordon, Grosz, and Gilliom 1993, interalia).

possible connections between discourse segment purposes and the entities for which such pronoun uses are felicitous (which, we conjecture, should be close connections). Although such pronoun uses provide compelling evidence of the hierarchical nature of discourse structure, the focus space stack in itself is not sufficient to explain their interpretation.

The claim of centering theory is not that centering alone suffices for resolving all pronominal reference (see Kehler [1997]), but that when attentional state plays a role, it is local, not global attentional state. This distinction is important as we reexamine Walker's data and claims.

The stack model requires that a speaker indicate to a hearer when attention is shifting from one discourse intention, and thus segment, to another. As a result of such shifts, focus spaces may be "pushed" or "popped." Typically, isolated references alone (i.e., individual pronouns or definite descriptions) do not suffice. Many different cues have been discussed in the literature, including cue phrases (Hirschberg and Litman 1993; Sidner 1981), intonation (Grosz and Hirschberg 1992; Hirschberg and Litman 1993; Nakatani 1997), repetition of previous content (Sidner 1979), and tense (Polanyi 1988). With this in mind, we can examine dialogues A–C in Walker's squib.

In comparing dialogues A and B, Walker argues that the utterance "as well as her uh husband" is easy to interpret after a short interruption but more difficult to interpret after a longer interruption.[2] She claims that this argues against the stack model because the length of the interruption is not a factor in that model. However, Walker overlooks the fact that "as well as her uh husband" is a sentence fragment, and thus depends on local attentional state for its interpretation. It may be this aspect of the utterance rather than the pronoun that makes interpretation difficult. The alternative continuation utterance, "OK, well, her husband also works" seems much easier to interpret for both versions A and B.

Walker uses dialogue C to support her argument that IRUs provide evidence for linear recency. The gist of her argument is that the material is repeated because it is no longer salient and thus no longer in the cache. However, there are alternative explanations of the function of this particular IRU, all of which are compatible with the stack model. One is that the IRU is being used in part to help identify the discourse intention to which attention is returning. That is, rather than identifying information in main memory that needs to be made salient again, the IRU is a repetition that helps to establish to which part of the dialogue (i.e., which focus space on the stack) attention is now returning. In this example, attention is returning not merely to the certificates, but to the advisor's diversification argument concerning them. With respect to this kind of example, the cache and stack models differ in how they find the information (by looking in main memory or in the stack), more than in what they do once they find it. In either case, the IRU functions to identify the information (discourse segment content and purpose) to which attention is returning, i.e., to focus (for the cache model) or refocus (for the stack model) attention on something previously salient. The difference is whether that material is found in main memory or in the stack. To determine whether the stack model is appropriate, one would need to determine whether other items in the focus space to which attention returns become salient.

2 Intonation plays a major role in the ease or difficulty with which these spoken dialogues could be interpreted. That role must be taken into account for processing claims to be supported.

4. Some Deficiencies of the Cache Model

Although the attentional state model leaves certain details unspecified (e.g., as Walker notes, the depth of the stack), it makes specific claims about discourse segment structure, discourse segment purposes, and shifts in attentional state.[3] The cache model, as Walker presents it, leaves too many details unspecified to allow a similar level of analysis or critique or solid comparison with the stack model. For such purposes the memory organization, replacement policy, and retention must be more completely specified. Some key questions are:

> **Memory organization:** What is the structure and organization of main memory? According to Walker's cache model if an entity is removed from the cache it appears only in main memory and must be retrieved from there. Main memory contains a vast amount of information. How is the relevant information found? In particular, with an IRU, is the information found in the same way on second retrieval as it was initially? Is there a claim that long-term memory in some way separates out information in the current discourse? In what ways does this structure compare to the focus space stack?
> **Cache replacement policy:** Does only a single entry get changed for each new entity mentioned in the discourse or are entities related to the old entry removed and others related to the just-mentioned entity added? If such related entities are deleted or added to the cache, which semantic or pragmatic properties determine sufficient relatedness?
> **Retention:** On what specific basis (e.g., which particular discourse-intentional relationships) are entities (preferentially) retained in the cache?

5. Some Problems with Walker's Comparison of the Models

In Section 5 of the squib, Walker compares the stack and cache models along a number of dimensions. Although her description of the effects in the stack model as participants shift between different discourse intentions is mostly accurate, the statement under the second bullet that entities in the focus space are no longer accessible is misleading. These entities are accessible, but access is more complex and less efficient, because they are no longer stack-accessible, i.e., they cannot be retrieved through the stack. However, they are accessible in memory just as they were before and just as the cache model requires for anything removed from the cache.

More importantly, the cache model is incomplete in ways that are essential to deciding between the two models. No definition is given of "related to new intention" (first bullet) or "related to prior intention" (third bullet). Without these specifications it is unclear what material not explicitly mentioned is brought into the cache. A specification of the cache replacement strategy is essential to determining the appropriateness of the cache model when intentions are completed (bullet 2). The statement that the cache "retrieves entities related to the prior intention from the main memory to the cache, unless retained in the cache," leaves unanswered two critical processing

3 There are, in addition, several well-recognized problems with the model. In particular, as used in computer systems, stacks do not differentiate among different kinds of frames, but interruptions seem to operate differently from normal embeddings (Grosz and Sidner 1986) and there are open issues in explaining pronominal reference at discourse segment boundaries.

questions: (1) How is the cache searched for related entities and how is relatedness determined? (2) How is a "prior intention" determined i.e., what memory is there for intentional structure and how is that coordinated with information in main memory so that the relevant information can be found? The stack model coordinates intentional structure and attentional state to address just these issues.

Walker makes two problematic claims in comparing the two models' treatments of interruptions. First, dialogue A differs from dialogue B, in that the interruption in A is three utterances long whereas that in B is five utterances. If this difference affects the cache content as significantly as Walker's model requires, then the cache is very small; it could not accommodate a very large discourse segment. But then one must ask what happens in the cache model with a discourse segment that is longer than three or four utterances. Would such a segment require that the reader or writer explicitly repeat material that is five or more sentences back? Second, in discussing dialogue C Walker says that without (22b), (22c), and (23) the inference required to understand the discourse "requires more effort to process," a claim that requires substantiation; in particular, utterance (22a) followed by (24) without the *and* that is required because of (22c) seems no more difficult to process than the fragment given. In both these cases, more empirical investigation is needed to determine the appropriate model.

With respect to return pops, Walker has misconstrued Sidner's stacked focus constraint. This constraint was postulated before Sidner integrated her work on local focus of attention with the focus-space stack work. Sidner's claim is significantly different from Walker's. In particular, it stipulates which pronouns can*not* be used for entities other than those locally in focus (i.e., those that we now call centers of an utterance). With respect to most of this discussion, Walker does not discuss the potential cost of checking nonfocused entities, but this cost can only be low if very few entities are checked, not all of long-term memory. In the air compressor (not pump!) example to which she refers there was a room full of equipment that could have been made to work and that was visually salient. An alternative to the explanations Walker provides for the IRUs with return pops is that IRUs are a good way to shift attention to a prior discourse intention and segment (hence focus space).

Finally, we note that the cache model does suggest an interesting research issue that is not clearly raised in the stack model. In computer systems, when information is moved from main memory to the cache, it is moved in blocks larger than the individual piece of information that was initially sought. The guiding locality principle here is spatial locality: programs typically need to access information (instructions and data) that have addresses near each other in memory (Patterson and Hennessy 1996). The processing trade-off that computer designers must address is one of processing time versus hit rate: larger blocks take longer to move, but increase the likelihood that subsequently needed information will be in the cache. Implicit focus (Grosz 1977b) and the treatment of functionally related entities (Grosz, Joshi, and Weinstein 1995) respond to a related issue in discourse processing, namely, what other than an entity itself becomes focused when the entity is focused? Although it is widely acknowledged that various related entities become salient (e.g., the door of a house, the time or location of a meeting), the determination of the scope of what becomes salient remains an open question (Grosz 1981). Whereas computer architects can depend on physical proximity in memory, discourse processing requires measures of conceptual closeness.

6. The Psychology of Discourse Processing

Like Walker, we support the integration of research in psycholinguistics and research in computational linguistics, and we support the contention that human information

processing is constrained by capacity limits, a position that is held by most, but not all, cognitive psychologists. However, the psycholinguistic literature does not support Walker's contention that a cache in combination with main memory, as is standard in computational architectures, provides a good basis for a computational model of human attentional capacity in processing discourse (Walker 1996b). The modal model of memory, capturing common aspects of memory models of the 1960s and early 1970s, had a short-term and long-term memory organization resembling cache–main memory organization. However, research since then has caused cognitive psychologists to revise their views of memory in ways that are not consistent with the idea that a cache-like memory contributes to discourse processing.

The prominent work of Baddeley (1986) retains the notion of sharply limited capacity in working memory, but the component subsystems that have these limitations (the articulatory loop and the visual-spatial sketchpad) represent information at levels that are not directly useful in discourse processing of the sort with which Walker is concerned. Baddeley advances the additional idea of a limited-capacity central executive that controls processing, but for this executive, the appropriate computational analogy is the processor rather than the cache or memory. Kintsch's (1988) position on short-term memory capacity (articulated in Kintsch & van Dijk [1978]) depends on a general model of discourse-processing that incorporates many other processing assumptions. In fact, Kintsch and van Dijk (1978) report that reducing the capacity of the short-term buffer from four propositions to one proposition has no effect on how well the discourse-processing model fits human subjects' performance in recalling and summarizing stories. In more recent work, Ericsson and Kintsch (1995) have argued that the amount of information that is required for working memory to perform the tasks ascribed to it far exceeds the capacity of the kinds of memory stores that are studied using traditional short-term memory tasks. They have proposed that working memory makes use of highly organized long-term memory. One source of evidence for this view is that frequently people can easily resume a task that has been interrupted, a kind of evidence that was also used to motivate the original stack model (Grosz 1977a). Walker also cites experimental research on pronoun interpretation and other types of inference in support of her cache model. Clark and Sengul (1979) is cited in support of the notion that pronoun interpretation is based on a linear backward search of the text, but this research has been criticized for confounding distance between the pronoun and its antecedent with topic shifts (Garnham 1987). Studies of local coreference (within discourse segments) clearly show that recency is not the primary factor in human pronoun interpretation (see Gordon and Scearce [1995] for a review). Studies of coreference beyond the local domain (called "reinstatement" in the psychological literature) do not provide evidence of a powerful effect of recency in determining ease of comprehension (O'Brien et al. 1995). Other research cited by Walker does support the idea that inference in human language comprehension is constrained, but it does not provide a basis for distinguishing capacity-limited and focus-based approaches to this constraint.

7. Summary

In sum, we agree with Walker that: (1) a model of attentional state is important for explaining the uses of certain linguistic expressions (e.g., cue phrases and pronouns); (2) human mental processes have limitations of both memory and processing. We disagree with Walker's claims that (1) limited memory capacity is the key architectural feature for attentional state; (2) IRUs should be explained on the basis of this limited capacity; (3) the felicity of pronominal processing after an interruption provides ev-

idence of this limited memory capacity; (4) the psychological literature supports the cache model.

References

Baddeley, Alan D. 1986. *Working Memory*. Oxford University Press, New York.

Clark, Herber H. and C. J. Sengul. 1979. In search of referents for noun phrases and pronouns. *Memory and Cognition*, 7:35–41.

Ericsson, K. Anders and Walter Kintsch. 1995. Long-term working memory. *Psychological Review*, pages 211–245.

Garnham, Alan, 1987. *Understanding Anaphora*, volume 3, pages 253–300. Erlbaum, London.

Gordon, Peter, Barbara Grosz, and Laura Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 3(17):311–347.

Gordon, Peter C. and Kimberly A. Scearce. 1995. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, 23(3):313–323.

Grosz, Barbara and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In John Ohala et al., editors, *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP-92)*, pages 429–432, Edmonton, Canada. Personal Publishing Ltd.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, Barbara and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara J. 1977a. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Cambridge, MA. Reprinted in Grosz et al. 1986, pages 339–352.

Grosz, Barbara J. 1977b. The representation and use of focus in dialogue understanding. Technical Report 151, Artificial Intelligence Center, SRI International, Menlo Park, CA.

Grosz, Barbara J. 1981. Focusing and description in natural language dialogues. In A. Joshi, I. Sag, and B. Webber, editors, *Elements of Discourse*. Cambridge University Press, Cambridge, England, pages 84–105.

Grosz, Barbara J. and Candace L. Sidner. 1997. Lost intuitions and forgotten intentions. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse*. Oxford University Press. To appear.

Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*.

Kehler, Andrew. 1997. Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475.

Kintsch, Walter. 1988. The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, pages 393–394.

Kintsch, Walter and Tuen van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85:393–394.

Lochbaum, Karen E. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

Nakatani, Christine. 1997. *The Computational Processing of Intonational Prominence: A Functional Prosody Perspective*. Ph.D. thesis, Harvard University.

O'Brien, Edward J., Jason E. Albrecht, Christopher M. Hakala, and Michelle L. Rizzella. 1995. Activation and suppression of antecedents during reinstatement. *Journal of Experimental Psychology*, pages 626–634.

Patterson, David A. and John H. Hennessy. 1996. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12.

Sidner, Candace L. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology. Technical Report 537.

Sidner, Candace L. 1981. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231.

Walker, Marilyn A. 1996. Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264.