# Verification of Sentences Containing Anaphoric Metaphors: An ACT-R Computational Model

**Raluca Budiu (raluca@cmu.edu)**
**John R. Anderson (ja@cmu.edu)**
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213

### Abstract

In a sentence-verification experiment, Budiu and Anderson (2001) found that participants took longer to respond to sentences containing anaphoric metaphors than to corresponding sentences containing literals. We present a computational model of this experiment, based on INP, a more general ACT-R (Anderson & Lebiere, 1998) model of sentence processing that has been used to explain various other linguistic and memory phenomena (Budiu & Anderson, 2000; Budiu, 2001; Budiu & Anderson, in preparation). This model shows that metaphors take longer to be processed because their low similarity to their antecedents generates an initial failure of comprehension; that failure may be resolved through an expensive reevaluation process at the end of the sentence, in light of the supplemental information brought in by the other words in the sentence.

## Introduction

Metaphors such as *time is money* or *the war with inflation* are common in every-day language; people often use and comprehend them with great ease, without even being aware of their existence in text. One question that has concerned researchers in this field was how the comprehension of metaphors compares with the comprehension of literals. The results of numerous experiments on this topic have been contradictory: some studies (Ortony, Schallert, Reynolds, & Antos, 1978; Inhoff, Lima, & Carroll, 1984; Shinjo & Myers, 1987; Budiu & Anderson, 2002) found no difference between reading times for metaphoric and literal sentences, whereas others indicated that people take longer to comprehend metaphors (Gibbs, 1990; Onishi & Murphy, 1993; Budiu & Anderson, 2001). Elsewhere (Budiu & Anderson, 2002, in preparation) we have argued that this difference is due to two factors: (1) a speed--accuracy trade-off, with participants in some experiments processing metaphors quickly, but incompletely comprehending them, and (2) supportiveness of sentence-context, with sentences containing different amounts of information that allow participants to relate the metaphor to the more general discourse. One particular class of metaphors that are exposed to comprehension deficits are anaphoric metaphors, which are metaphors typically occurring at the beginning of the sentence and that denote some concept present in previous text (e.g., *The bear beat his opponent*, where *bear* refers to a bulky wrestler introduced in a previous passage). The experiment discussed in this article (Budiu & Anderson, 2001) concerns verification of sentences containing anaphoric metaphors. We present a computational model of this experiment, based on INP (INterpretation-based Processing - Budiu & Anderson, 2000; Budiu, 2001; Budiu & Anderson, in preparation), a more general sentence-processing model developed on top of the ACT-R cognitive architecture (Anderson & Lebiere, 1998). INP is an incremental, real-time model of sentence processing that goes from parsing to semantic processing and was used to successfully simulate data from the metaphor literature, semantic illusions, text priming, and sentence memory (Budiu & Anderson, 2000; Budiu, 2001; Budiu & Anderson, in preparation). In the rest of the paper we briefly overview INP, then we present the experiment in Budiu and Anderson (2001) and the corresponding model. We end with conclusions.

## INP: An Overview

INP is a model of syntactic and semantic sentence processing. Given an input sentence, it produces a syntactic and a propositional representation, and also an **interpretation** for that sentence. The interpretation is the central concept in INP - it is a proposition in the background knowledge that overlaps most with the current input. For instance, if the input sentence were *The man paid the waiter*, a possible interpretation might be *At the restaurant, the customer paid the waiter* (which is part of our general knowledge about restaurants). If the input sentence communicated novel information, such as *Tom Tykwer directed "The princess and the warrior"* may do for those less familiar with non-American movies, a possible interpretation may be *The person directed a play*[1]. Thus, to use Haviland and Clark's (1974) terminology, the interpretation relates the "given" part of the input sentence to the prior knowledge. This is possible because even novel sentences typically have some old or "given" part.

Whereas building the syntactic and semantic

[1] "The princess and the warrior" is actually a movie.

Table 1: Sample materials from Budiu and Anderson (2001)

| *Example 1* | *Example 2* |
|---|---|
| John Quinch, a world champion arm wrestler, was invited to the Jones' house. Mrs. Jones wanted to make a nice nut cake, but the nut cracker was broken. She was quite desperate, but John offered to help her. It took him no more than five minutes to crack all the nuts in his powerful fists. Mrs. Jones was able to bake a delicious cake. | Joe was a massive man and a lumberjack champion. Every time he went with his family at their little chalet in the mountains, he loved to do all the hard work: carrying water from the source and cutting down trees for firewood. His family was very proud of him. |
| *Metaphoric Sentences:* | *Metaphoric Sentences:* |
| The bear cracked the nuts *[target]* | The bear worked hard in his mountain house *[target]* |
| The bear baked the cake *[hard foil]* | The bear worked hard in his city residence *[easy foil]* |
| *Literal Sentences:* | *Literal Sentences:* |
| The athlete cracked the nuts *[target]* | The athlete worked hard in his mountain house *[target]* |
| The athlete baked the cake *[hard foil]* | The athlete worked hard in his city residence *[easy foil]* |

representations is important for comprehension (and INP deals with this part, too), finding an interpretation is the central operation in INP. This process is incremental - roughly speaking, with each new content word read, INP tries to "guess" the interpretation of the sentence. Thus, at each moment, INP searches for a candidate interpretation that matches best the last three content words read. The candidate interpretation is then matched against subsequent words - if a new word does not match it, the interpretation is rejected and another one is searched for.

This search-and-match process is computationally inexpensive due to ACT-R's mechanism of activation spreading. In ACT-R, the items that are currently in the focus of attention spread activation to other items to which they are associated, and the amount of activation spread is proportional to the strength of association. INP assumes that strengths of associations reflect semantic similarities. Thus, as INP reads the sentence, the last three content words are kept in the focus and spread activation to all propositions that are semantically similar to them. The model picks the most active proposition in memory as the candidate interpretation (provided that its activation is above some threshold), because, in most cases, that proposition has received the most spreading activation and, thus, is the most similar to the input words in the focus. The final interpretation of the sentence is the interpretation that the model has reached after processing all the words in the sentence.

Sometimes no proposition in memory may be active enough (i.e., above the ACT-R retrieval threshold) and, thus, INP cannot find any candidate interpretation. Whenever such an event happens, INP creates a chunk called **bug**. The bug encapsulates information about the state of the model when the failure was encountered. Bugs allow INP to estimate the truth (or non-novelty) of a sentence. For instance, if a sentence reiterates some information already contained in the previous text or in the prior knowledge (e.g., *The man paid the waiter at the restaurant*), the process of finding an interpretation will be smooth, bug-free; if the sentence communicates novel or false information that does not match any fact in the prior knowledge, the model will form one or more bugs during comprehension. However, even if at some point INP has formed a bug, it is possible that subsequent words in the sentence help it reach a final interpretation.

We saw that the activation-spreading process controlling the search for an interpretation in INP is driven by semantic similarities between words and/or propositions. To set semantic similarities between words we use Latent Semantic Analysis (LSA - Landauer & Dumais, 1997). LSA is a mathematical technique that computes similarity as the distance between two words, which are seen as points in a multidimensional space. It has been used successfully to simulate a number of psycholinguistic phenomena (Landauer & Dumais, 1997; Laham, 1997; Kintsch, 2000). Although LSA may not always offer a perfect definition of similarity, it is a convenient way for estimating average similarities and provides a solid, independently defined constraint for our model.

## Overview of the Experiment

Experiment 2 in Budiu and Anderson (2001) illustrates a situation in which metaphors are processed more slowly than literals. In this experiment, participants read a short passage and then had to verify whether a probe sentence was true based on that passage. The probe could be either metaphoric or literal and either true (target) or false (foil). Metaphors were always used anaphorically: the metaphoric word referred to some concept previously introduced in the passage, but absent from the current sentence. Ratings of metaphors by human participants indicated that the metaphors were in a range of goodness and familiarity comparable to other metaphors used in psycholinguistic experiments. Table 1 shows examples of passages and possible probes. The foils could be further classified as easy or hard: the easy foils were designed so that the participants could reject them even without understanding the metaphor; the hard foils could not be answered correctly without first

Table 2: Percentage of correct responses and judgment times for correct responses in the first block of Budiu and Anderson's (2001) experiment: data and model.

| | | % Correct | | RT (s) | |
|---|---|---|---|---|---|
| | | Data | Model | Data | Model |
| Targets | Met | 53 | 54 | 4.50 | 4.67 |
| | Lit | 90 | 88 | 3.59 | 3.24 |
| Easy foils | Met | 100 | 98 | 4.18 | 4.83 |
| | Lit | 85 | 85 | 4.46 | 3.91 |
| Hard foils | Met | 73 | 69 | 4.41 | 4.34 |
| | Lit | 88 | 86 | 3.80 | 3.52 |

resolving the referent of the metaphor. For the Example 2 in Table 1, the foil is easy: even if they did not know the referent of *bear*, participants could reject *The bear worked hard in his city residence* because nobody in the story works in his city residence. For the Example 1 in Table 1, the foil *The bear baked the cake* is hard: it cannot be rejected by looking only at the predicate *baked the cake*, because, if *bear* referred to *Mrs. Jones*, the probe would be true.

Table 2 shows the results from the first block of Budiu and Anderson's (2001) experiment. Note that participants were less accurate and slower on metaphoric targets or metaphoric hard foils than on literal targets or literal hard foils, respectively (the effects were significant); however, they performed comparably on metaphoric and literal easy foils (the latency effects were not significant).

## Simulation of the Experiment

Table 2 also shows the results of the simulation. Our model's central assumption is that what distinguishes metaphors from literals is the similarity to their respective antecedents (i.e., the concepts from the previous passage that they denote): literals are more similar to their antecedents than metaphors. This assumption is the reason for the differences in reaction times and latencies for the two kinds of items and is confirmed by the LSA analysis. Indeed, the results in Table 2 were obtained by setting a similarity of 0.19 between the metaphors and their antecedents in the stories. This value corresponds to the average LSA distance between the metaphor and the story. Note that we used the whole story, not just the antecedent, to compute the LSA, because sometimes the antecedent was diffused across multiple sentences (for instance, one sentence mentioned that a character was a wrestler and another that he was very strong). In the same way, the similarity between the literal and its antecedent was set to 0.36 - the average LSA distance between the literal and the passage.

For each trial, INP starts with a representation of the facts corresponding to the passage. To perform passage-based verification of the probe sentence, the model searches for interpretations of the probe among the propositions coming from the passage. If it succeeds in comprehending the probe with no bugs (i.e., it is always able to find some candidate interpretation while processing the sentence) and reaches a final interpretation, INP answers *true*. Thus, it considers that a probe is true if it matches some fact in the prior text. On the other hand, if it produces a bug while comprehending the sentence, INP considers the sentence potentially false because, at some point, the information conveyed by it did not match anything in the prior passage. Later on, the model may recover from that local comprehension failure and still find a final interpretation for the probe. If that is the case, at the end of the sentence, before rejecting it because of the bug, INP may occasionally attempt to resolve the bug using the final interpretation. Thus, having an interpretation for the sentence can help INP get the right meaning of a metaphor.

Only some of the bugs are reevaluated - specifically, those bugs that correspond to potential metaphors in text. When it creates a bug, INP records information about the current word and, also, about whether the current word occurred in the passage. If the word did not occur in the passage, the bug is called a **metaphor bug**. Note that words with no referent in text are more likely to be metaphors than words that have already occurred in the text.

Let us look at how INP comprehends the metaphoric target *The bear cracked the nuts*, after reading the story in Example 1 from Table 1. After processing the metaphoric word *bear*, the model searches for a fact in the passage that matches that word. Due to the low similarity between *bear* and other words in the text, INP fails to find such an interpretation, so it forms a bug to record this local comprehension failure. The bug contains information about the current word (*bear*) and is marked as a metaphor bug, because *bear* does not occur in the passage. Next, INP processes the word *cracked*, which spreads positive activation to other passage facts that involve that concept (e.g., *John Quinch cracked the nuts*), enabling the model to select one of those facts as a candidate interpretation. Let us assume that the selected interpretation is *John Quinch cracked the nuts*. Finally, INP reads the word *nuts*. That word matches the current candidate interpretation, which becomes the final interpretation of the sentence. Thus, based on the literal words in the sentence, INP is able to find a final interpretation; however, the bug created early in the processing signals that the sentence has some anomaly in it and, thus, may be false. At the end of the sentence, when it must judge the truth of the sentence, the model can either reject the target because it had generated bugs or it can check whether the comprehension failure(s) can be reconciled with the final interpretation. INP selects the second alternative with a probability of 0.51; in that case, it retrieves the metaphor bug (i.e., the one corresponding to *bear*) and processes again its corresponding meaning as if it was a word phrase at the end of the input sentence. In our

example, the model checks whether *bear* matches the agent *John Quinch* of the final interpretation. In INP, matching a concept to an interpretation is somewhat easier than finding an interpretation based on that concept, so this reevaluation process succeeds very often.

The processing of literal targets (e.g., *The athlete cracked the nuts*) is identical to that of metaphoric sentences. As in the case of the metaphoric targets, the word *athlete* does not necessarily occur in the previous passage; however, because it is more similar than *bear* to other words in the passage, the model often finds an interpretation involving *athlete* when it processes this word, and, thus, may generate no bug. Thus, most of the time the model produces no bugs on literal targets; this behavior has two important consequences: (1) the model is correct most of the time on literal targets; and (2) it is faster most of the time on literal targets, because it does not need to go through the time-expensive process of bug reevaluation.

Let us now look at INP's behavior on foils. Because foils, by definition, never match perfectly a proposition in the discourse, the model either forms a bug at some point while processing them or ends up with no interpretation.

Both people and INP tend to find metaphoric easy foils less difficult than literal easy foils. The model shows this tendency because most words in both the subject and the rest of the sentence do not match the prior material, so metaphoric easy foils can be easily rejected. Thus, for the easy foil *The bear worked hard in his city residence* from Example 2 in Table 1, INP fails to find an interpretation on *bear*, because *bear* is not similar enough to other concepts in the context. Moreover, although the subsequent two words *worked hard* do match previous information and may lead to the model selecting some candidate interpretation (e.g., *John Quinch worked hard at the chalet*), the last part of the sentence, *city residence*, does not match that candidate interpretation (or anything else in the context), so INP ends up with no final interpretation. This lack of interpretation makes the model answer *false*.

INP's behavior on literal easy foils such as *The athlete worked hard in his city residence* is comparable to its behavior on metaphoric easy foils, although the model may find a candidate interpretation from the very beginning of the sentence, on the literal word (due to the greater similarity between that literal and its antecedent) and, thus, may avoid forming a bug on the literal. For Example 2 in Table 1, a possible candidate interpretation after reading *athlete* is *John Quinch worked hard at the chalet*. That candidate interpretation, even if it can be maintained for some time (for instance, on the words *worked hard* in our example), will be rejected at some later point (because the easy foil does not match perfectly anything in the context). In our example, this rejection happens on the concept *city residence*. Thus, for literal easy foils, the model will end up with no interpretation and will tend to answer false.

INP is faster on literal easy foils than on metaphoric easy foils because it forms fewer bugs (at least one less) and the process of bug creation is time expensive. (The data in Table 2 show a nonsignificant effect in the opposite direction; however, based on other results in Budiu and Anderson, 2001, we believe that this tendency reflects noise in the data[2].) Occasionally, if the comprehension is smooth enough (i.e., an interpretation, even if not final, was found during reading and there were not too many bugs), INP may make an error and answer *true* on an easy foil. We estimated a probability of 0.49 of answering *true* even in the case when there is no final interpretation, provided that a candidate interpretation was found at some point and that the model formed only a single bug during the comprehension of the sentence.

For metaphoric hard foils the processing is similar to that of metaphoric targets. Consider the story from Example 1 in Table 1. When it reads the word *bear* in the hard foil *The bear baked the cake*, INP produces a bug, as for the other types of metaphoric sentences. However, because the hard foil contains a predicate that matches a proposition in the context (e.g., *Mrs. Jones baked the cake*), that proposition may be the final interpretation. Therefore, for a hard foil, as for metaphoric targets, INP has two options: either to answer *false* in virtue of the metaphor bug or to reevaluate. Unlike for true sentences, the metaphor reevaluation rarely succeeds (because the subject of the final interpretation - *Mrs. Jones* - is not at all similar to the metaphor *bear*). Therefore, whether or not INP chooses to reevaluate the metaphor tends not to make a difference with respect to the final answer, which frequently is *false*.

As for metaphoric hard foils, for literal hard foils such as *The athlete baked the cake* (see Example 1 in Table 1), the model can potentially form a bug on the initial concept (*athlete*), due to the fact that this literal does not typically occur per se in the passage and, thus, may have a low similarity to the other nouns in the preceding discourse. However, the central assumption of the model is that the similarity between literals and their antecedents is higher than between metaphors and their antecedents. Thus, on average, the model will form fewer bugs on the initial word for literals than for metaphors. Later on, as for metaphoric hard foils, the other words in the sentence may lead INP to a final interpretation. Note that even when the model does not form a bug on the literal, it will still form bugs subsequently. Indeed, suppose that after reading the word *athlete*, the model settles for the candidate interpretation *The athlete helped Mrs. Jones*. When it reads the word *baked*, the model matches it against the

[2] Experiment 1 in Budiu and Anderson (2001) also looked at latency differences between literal and metaphoric easy foils and found that participants were about 700 ms slower on metaphors.

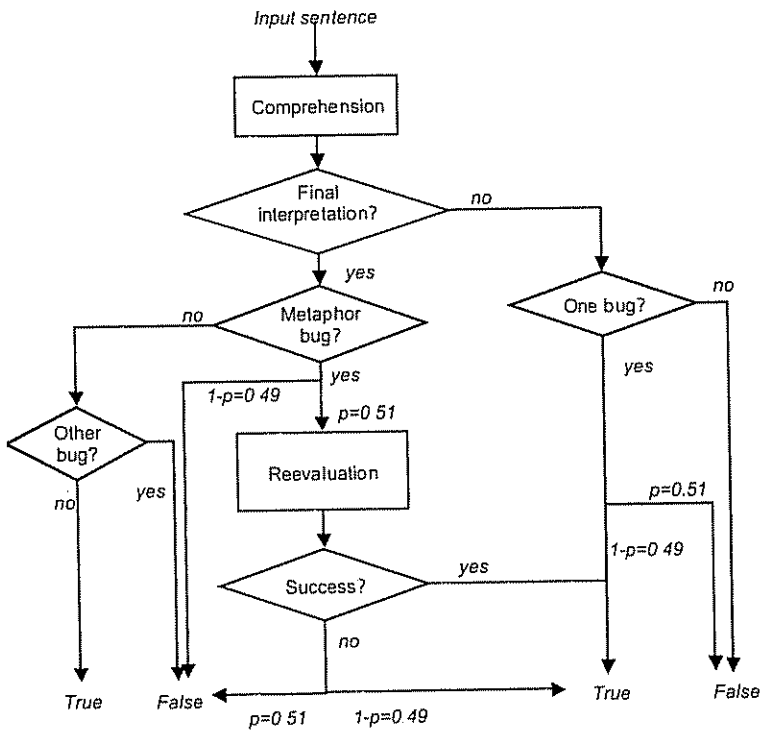


Figure 1. Summary of the model.

current candidate interpretation; because *helped* and *baked* do not match, the current interpretation is rejected and another one is sought. A possible candidate is *Mrs Jones baked the cake*; this fact receives positive activation spreading from *baked*, which is countervailed by the negative activation from *athlete* (because *athlete* and *Mrs Jones* are not at all similar). Thus, the total activation spreading to the fact *Mrs Jones baked the cake* is not enough to raise it over the threshold and make the model select it as a candidate interpretation. In this situation, the model will end up forming a bug on the word *baked*. When the final word *cake* is read, the total activation boost from *baked* and *cake* is normally enough to make *Mrs Jones baked the cake* a candidate interpretation. That interpretation is in fact the final interpretation of the sentence. However, the bugs (such as the one formed on the word *baked*) will enable the model to answer false on literal hard foils, but they are non-metaphor bugs and do not stand reevaluation. Hence, the model tends to be faster on literal hard foils than on metaphoric hard foils.

Occasionally the model makes errors on both metaphoric and literal hard foils. These are due to a probability (estimated as 0.49) that the reevaluation process goes wrong and that INP answers *true* where *false* would have been appropriate. The higher error rate on metaphoric hard foils (compared with literal hard foils) is because a larger percent of the metaphoric foils lead to metaphor bugs and thus need reevaluation.

The results obtained by the model are given in Table 2. The model succeeds in capturing the main result - namely, the difference between metaphoric and literal targets[3]. The latency difference between the metaphoric and literal sentences is due mainly to reevaluation and to extra bugs on metaphors. The accuracy difference is also the result of more bugs on metaphoric sentences. We estimated a number of ACT-R parameters: (1) the **retrieval threshold,** $\tau$, was estimated as -2; it indicates the minimum activation that a proposition (or other declarative structure) must have to be considered as a candidate interpretation; (2) the **latency factor,** $F$, was estimated as 0.035; the latency factor is a scaling factor in the ACT-R equation that relates the time to retrieve a declarative structure to the activation of that structure: $RT = Fe^{-A}$. For these parameters, the ACT-R theory does not stipulate any values and they vary widely among different ACT-R models. However, the value $Fe^{-\tau}$, which corresponds to the retrieval latency when the activation is at the threshold, tends to less variable across models; for our model, this value is 0.26s, which is in the range of the values used by other models (Anderson & Lebiere, 1998, Budiu & Anderson, in preparation)

## Conclusions

We have presented a computational model that explains processing-time differences between metaphors and literals as due to different semantic similarities to their

[3] The correlation between data and model is r = 0.995 for accuracies and 0.752 for latencies. Whereas reproducing quantitative results was important, we strive mostly for capturing the qualitative results and for consistency with other INP simulations.

antecedents in the passage. The model predicts those differences using the same processing mechanism for both metaphors and literals. Figure 1 summarizes the behavior of the model. The model answers *true* to sentences for which it is able to reach a final interpretation without encountering any local failures of comprehension (i.e., bugs). Occasionally, when it finds a final interpretation, but has also formed a bug (as it is often the case for metaphoric sentences), INP spends some time at the end of the sentence to reevaluate the metaphor; if that reevaluation is successful it answers *true*. Note that most decisions are nondeterministic: in only about 50% of the times the model chooses to reevaluate a metaphor bug. Also, sometimes INP makes mistakes and answers *true* when the appropriate response would have been *false* (e.g., when it has no interpretation or when the reevaluation was unsuccessful). This sloppiness of the model reflects reports from participants in this study - they sometimes answered *true* when they noticed an inappropriate word such as *bear* in contexts such as those from Table 1, regardless of the other words in the sentence.

The metaphors are often not understood correctly (due to their low similarity to their antecedents), but further words in the sentence may help find some gist of the probe; in that case, reevaluation of the metaphors may happen at the end of the sentence. This reevaluation leads to slower comprehension for metaphors than for literals. Budiu (2001), Budiu and Anderson (in preparation) showed that when comprehension accuracy is of less importance, INP is capable of reading literal and metaphoric sentences as fast, at the expense of comprehension quality. To conclude, our simulation captures the main difference between metaphors and literals through a time-expensive process of reevaluating past failures of comprehension in the light of global sentence information. Because the similarity of the metaphor to its referent is small and because the metaphor occurs at the beginning of the sentence, INP acts as if it processed the initial metaphoric word literally.

## Acknowledgments

This research was supported by National Science Foundation Grant BCS997-5-220.

## References

Anderson, J.R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Budiu, R. (2001). *The role of background knowledge in sentence processing*. Doctoral dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Budiu, R., & Anderson, J.R. (2000). Integration of background knowledge in sentence processing: a unified theory of metaphor understanding, semantic illusions and text memory. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 50-57). Netherlands: Universal Press.

Budiu, R., & Anderson, J.R. (2001). *Word learning in context: Metaphors and neologisms* (Tech. Rep. CMU-CS-01-147). Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.

Budiu, R., & Anderson, J.R. (2002). Comprehending anaphoric metaphors. *Memory and Cognition, 30,* 158-165.

Budiu, R., & Anderson, J.R. (in preparation). *Interpretation-based processing: A unified theory of sentence comprehension.*

Gibbs, R. (1990). Comprehending figurative referential descriptions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16,* 56-66.

Haviland, S., & Clark, H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior, 13,* 512-521.

Inhoff, A., Lima, S., & Carroll, P. (1984). Contextual effects on metaphor comprehension in reading. *Memory and Cognition, 2,* 558-567.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review, 7,* 257-266.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In M. Shafto & M. Johnson (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 105,* 221-240.

Onishi, K., & Murphy, G. (1993). Metaphoric reference: When metaphors are not understood as easily as literal comprehension. *Memory and Cognition, 21,* 763-772.

Ortony, A., Schallert, D., Reynolds, R., & Antos, S. (1978). Interpreting metaphors and idioms: Some effects on comprehension. *Journal of Verbal Learning and Verbal Behavior, 17,* 465-477.

Shinjo, M., & Myers, J. (1987). The role of context in metaphor comprehension. Journal of Memory and Language, 26, 226-241.