Psychonomic Bulletin & Review
2006, 13 (5), 848-853

Time course of retrieving conceptual information: A speed–accuracy trade-off study

BRIAN McELREE, GREGORY L. MURPHY, and TAMARA OCHOA
New York University, New York, New York

Words carry considerable information, but much of that information is not relevant in context. Research has shown that readers selectively activate and remember relevant information associated with words in different contexts, but it is not known when in processing this selection occurs. This experiment investigated whether context can change which properties are initially retrieved, using a speed–accuracy trade-off paradigm. Readers had to verify a property of a modifier–noun phrase (e.g., in the sentence *Boiled celery is soft*) within a specified interval, from 300–3,000 msec after presentation. Results revealed that properties associated with the noun alone were activated sooner than were properties that required integration of the modifier with the noun. Thus, context did not serve to influence the initial retrieval of properties but only to activate or suppress properties later in processing.

Content words communicate not only what might be thought of as "linguistic" or definitional information, but also conceptual information about the entities referred to. If you expressed surprise when someone told you, "Chris licked me," the speaker might explain, "Chris is my dog." This response perfectly explains Chris's surprising behavior, even though no one could claim that the word *dog* includes in its definition or purely linguistic representation the information that dogs lick people. Essentially, anything known about dogs can be referred to by *dog*. Given that words provide access to vast amounts of conceptual information (Murphy, 2002, chap. 11), how do comprehenders know what information is relevant in a given context? Clearly, one cannot retrieve all known facts about a word every time the word is encountered. Even if one could, most of the facts would be irrelevant to the present purposes and so would impair rather than aid comprehension. Not surprisingly, researchers have found that people tend to selectively access conceptual information, depending on the context.

Tabossi and Johnson-Laird (1980) presented words within context sentences such as (1) and (2) that emphasized one of two different aspects of the word's meaning.

(1) The goldsmith cut the glass with the diamond.

(2) The mirror dispersed the light from the diamond.

After reading one of these sentences, the participants then judged a sentence that tested an aspect of the meaning of *diamond* evoked by (1) or (2), for example, *Diamonds are hard* or *Diamonds are brilliant*. Verification was faster when the test sentence queried an aspect of meaning emphasized in the context sentence (see also McKoon & Ratcliff, 1988; Tabossi, 1982).

Research on text memory has also emphasized contextual influences on word interpretation. Barclay, Bransford, Franks, McCarrell, and Nitsch (1974) presented readers with sentences such as (3) or (4).

(3) The man lifted the piano.

(4) The man tuned the piano.

On a later memory test, they showed that "something heavy" was a better memory cue for (3) than was "something with a nice sound." However, the reverse was true for (4), suggesting that one sentence caused readers to retrieve the weight of pianos and the other their sound.

Context-sensitive retrieval of information is important to comprehension: One need not retrieve everything one knows about diamonds to understand sentence (1), but the sentence would be difficult to understand if one failed to retrieve the fact that diamonds are hard. Although the context sensitivity of lexical information is well agreed upon, it is not known how retrieval activates the relevant information associated with a word and not contextually irrelevant information.

One possibility is that the other words in the sentence independently prime the relevant aspects of meaning, and so the priming from the context and the target word sum to result in an active property. However, independent priming cannot explain the results: Barclay et al. (1974), for example, showed that the sentential context alone could not produce their effects (and see McKoon & Ratcliff, 1988).

In short, lexical activation alone apparently cannot explain the selection of relevant noun properties. Rather, these properties must be retrieved as a result of interpreting the sentence or passage meaning. How does this

The order of authorship is alphabetical. This research was supported by NSF Grant 0236732 (awarded to B.M.) and NIMH Grant MH41704 (awarded to G.L.M.). Correspondence concerning this article should be addressed to B. McElree, Department of Psychology, New York University, 6 Washington Place, 8th Floor, New York, NY 10003 (e-mail: brian.mcelree@nyu.edu).

occur? One possibility is that constructing a discourse representation leads to the activation of relevant information and suppression of irrelevant information. This proposal suggests that selection of relevant properties is a fairly late process, occurring during the formation of a discourse representation. Indeed, an important commonality of the studies reviewed above is that sentences or paragraphs were read and interpreted first, with the dependent measures taken offline—measured after comprehension was complete, sometimes minutes later.

However, evidence from the interpretation of noun phrases (NPs) suggests that this process might not be very late. Potter and Faulconer (1979) aurally presented sentences such as *It was already getting late when the man saw the burning house ahead of him*. In the critical condition, a picture appeared immediately after the word *house*, and participants had to verify whether the picture referred to something in the sentence. The picture could either be of the noun, unmodified (e.g., a normal house) or the noun as modified by the preceding adjective (a burning house). The participants were instructed to respond positively to either kind of picture. With an unmodified noun, people responded faster to the unmodified picture; when the adjective was present, they responded slightly faster to the modified picture. Thus, the presence of the adjective right before the noun seemed to cause an immediate change in the noun's interpretation. The investigators suggested that the modifier engendered a selective search process, in which only information relevant to the NP was retrieved. That is, some normal noun properties were not retrieved when the modifier was present, and some novel properties not strongly associated with the noun *were* retrieved.

Similarly, Springer and Murphy (1992) found that people were faster to verify sentences such as *Boiled celery is soft* than *Boiled celery is green*, even though the first requires combining the adjective and noun (because celery per se is not soft), and the second does not (because celery per se is green). Thus, selective retrieval of information seemed to operate quickly.

This conclusion may appear to conflict with results from studies of homonyms, which have often concluded that initial selection of a meaning is insensitive to context (e.g., Swinney, 1979). However, those studies concern a theoretically very different process, selection of a lexical entry, than the question of semantic information retrieval. Furthermore, their measures are often subject to the same concerns we raise in the next section about prior studies of semantic retrieval.

Evidence for Selective Retrieval Processes

All these findings argue that the representation of a word's meaning in discourse is context sensitive. The Potter and Faulconer (1979) and Springer and Murphy (1992) studies went farther, in suggesting that contextually relevant information is not merely emphasized in the final sentence representation, but that the retrieval process itself is context sensitive in being biased toward relevant properties. Irrelevant properties are *never* retrieved, even if they are strongly associated with the concept. If true, this would suggest an extremely "smart" retrieval process in which semantic memory was addressable not only by words but also by word combinations and sentences.

However, the results so far do not justify this conclusion. The memory studies used a dependent measure that was taken after too great a delay to tell us what information was initially retrieved. In the Springer and Murphy (1992) paradigm, participants were allowed to take as long as they wanted to answer the questions, and it is not clear when the information was retrieved in the process of comprehension. Also, it is possible that the effect was due in part to sentence pragmatics: Verification could have been influenced by the fact that it is more pragmatically appropriate to ask about relevant properties than apparently irrelevant ones (Gagné & Murphy, 1996; Glucksberg & Estes, 2000).

Potter and Faulconer (1979) argued most specifically for a selective retrieval process, given that their picture probe appeared immediately after the critical noun. However, responses to this probe were as slow as 930 msec, and so participants may not have relied on information that was retrieved early. More importantly, their critical comparison did not independently sample relevant and irrelevant material. They assumed that the normal picture revealed whether information about the bare noun was retrieved, whereas the modified picture revealed whether retrieval was selective (because it matched the meaning of the modified noun but poorly matched the unmodified noun). The problem is that the normal picture not only matches the noun, it *mismatches* the modified NP. That is, it does not simply measure whether people retrieved information about houses, it could also show that people realized that the picture did not match the phrase *burning house*. Even if readers did initially retrieve information about (unmodified) houses, as interpretation of the modified phrase developed, it may have interfered with verification of the normal picture. In short, the unmodified picture does reflect retrieval of the noun, but it also could reflect interference from the phrase, and so it cannot be used as a neutral measure of retrieval of noun information.

The present research attempted to provide more conclusive evidence on this issue by two means. First, we used the Springer and Murphy (1992) approach in which modified NPs were tested on properties relevant to the noun only or to the entire phrase. But in contrast to the Potter and Faulconer (1979) design, the noun features were also consistent with the phrase. *Boiled celery is green* is true because of celery being green (thus, a noun feature), and boiling celery does not change its color. Therefore, this feature serves as a measure of when information is retrieved from the noun concept, without being interfered with by the phrase as a whole. In contrast, the sentence *Boiled celery is soft* includes information that is not found in the noun but is found in the phrase. If Springer and Murphy and Potter and Faulconer were right in concluding that retrieval of properties is itself context sensitive, then such properties ought to be retrieved faster than (or at least no slower than) noun properties that are not contextually relevant.

Second, we used a speed–accuracy trade-off (SAT) variant of the sentence verification task (see McElree, Foraker, & Dyer, 2003). The participants were required to respond within six predefined time windows, and the dependent measure was response accuracy within each window. The windows were chosen so that responses were initially at chance and eventually reached maximum accuracy, so that the accumulation of information could be traced as a function of time. A critical advantage of this technique is that people cannot themselves decide when to respond. For example, if Springer and Murphy's (1992) participants waited until they felt that they had fully understood the sentence before responding, late processes of sentence interpretation would have influenced their judgments rather than only initial memory retrieval. By forcing readers to respond both early and late, we could measure processes operating at each time window.

Thus, we tested people's verification of sentences containing modified NPs, examining the differences between features associated with the noun versus features of the phrase that were not strongly associated with the noun. We also tested corresponding false properties, such as *Boiled celery is blue* (false of the noun) and *Boiled celery is crisp* (true of the noun but false of the phrase). These sentences may also provide a test of the selective retrieval process, because one might expect them to be equally easy to reject, if only relevant properties are retrieved.

METHOD

Participants

The participants were 92 NYU undergraduates, of which 60 performed norming, 20 served in the SAT experiment, and 12 participated in a supplemental reaction time (RT) experiment. All were native English speakers.

Materials

We generated 113 modifier–noun combination phrases. Each phrase was paired with a predicate from each of four conditions (see Table 1): true of the noun (TN), false of the noun (FN), true of the phrase (TP), and false of the phrase (FP), yielding a total of 456 experimental sentences. The TN and FN sentences contained predicates whose truth value could be determined by virtue of the noun alone; the modifier was irrelevant. The TP and FP sentences contained predicates whose truth value depended on the entire phrase. For example, the TP sentence *Water pistols are harmless* is true by virtue of the whole phrase, but false by virtue of *pistol* alone.

Norming. Two tests ensured that properties were equally typical or atypical in the noun and phrase conditions. The first measured how typical true properties were and how atypical false properties were of each phrase. The experimental sentences were divided into two lists, and 15 participants judged sentences on each list using a scale of 1 (*not at all typical*) to 7 (*very typical*). Phrases were discarded if their scores on the TP and FP conditions differed by less than 2.0.

A second test measured the typicality of the properties with respect to only the noun of the combination. For example, the participants rated *Pistols are harmless*. Fifteen participants rated each list of these items. We retained 95 items having differences between the two tests greater than 1 in the TP and FP conditions and less than 1 in the TN and FN conditions. Overall, the selected TP predicates were typical of the phrase but not the noun alone (Ms = 5.74 and 3.05). In contrast, the TN predicates were equally typical of phrase and noun (Ms = 6.01 and 6.05). Similarly, the FP items were not typical of the phrase but were of the noun (Ms = 2.16 and 5.51), but the FN items were atypical of both (Ms = 1.70 and 1.78).

Table 1
Sample Sentences Used in This Study

Sentence Type	Sample Sentence
TN	Water pistols have triggers.
TP	Water pistols are harmless.
FN	Water pistols have string.
FP	Water pistols are dangerous.

Note—TN, true of the noun; TP, true of the phrase; FN, false of the noun; FP, false of the phrase.

Procedure

The participants served in two 1-h sessions. The first was preceded by 10 min of practice with sentences similar to those in the experimental blocks. The two sessions used the same stimulus sentences, but the order of sentences and processing intervals varied. Each session consisted of four blocks of 96 sentences. Each sentence was presented only once per session.

The SAT procedure was used to measure changes in accuracy over time. The participants were told that they would judge the truth of the sentences presented on a computer screen and that they needed to respond within 100–300 msec of hearing a tone. The practice trained the participants to respond within this interval.

Each trial began with a 1-sec fixation asterisk presented at the center left of the computer screen. A combination phrase such as *water pistols* then appeared for 600 msec, beginning where the asterisk had been. Then a property such as *are harmless* appeared at the same location and remained on the screen for a varying amount of time, followed by a tone that cued the participants to respond. This tone occurred at 300, 500, 700, 900, 1,500, or 3,000 msec after the onset of the property, randomly intermixed within a block. The participants responded "yes" or "no" to the sentence by pressing 1 or 3 on the number pad of the keyboard. The latency to respond to the tone was presented after each response. If the response was not within the time window, an error message appeared. There were a total of 32 trials/condition for each participant at each of the six interruption times.

RESULTS

Positive Conditions

Table 2 presents the acceptance rates of each condition at each latency. We first examined differences between the TP and TN conditions. A d' measure was constructed for each participant by scaling the z score of the hit rates of each condition against the z score of the average false alarm rate of the FP and FN conditions. This d' scaling isolated the differences in hit rates between the TP and TN conditions, scaling them against the average proportion of false alarms, to correct for possible response biases. Figure 1A shows these d' scalings for the average data as a function of processing time (the lag of the interruption tone plus the latency to respond to the tone).

Figure 1A illustrates that there were no reliable differences in asymptotic performance at the longest interruption point [times $>$ 3 sec; $t(20) = 0.62$, $p > .25$]. However, at early interruption times there were substantial

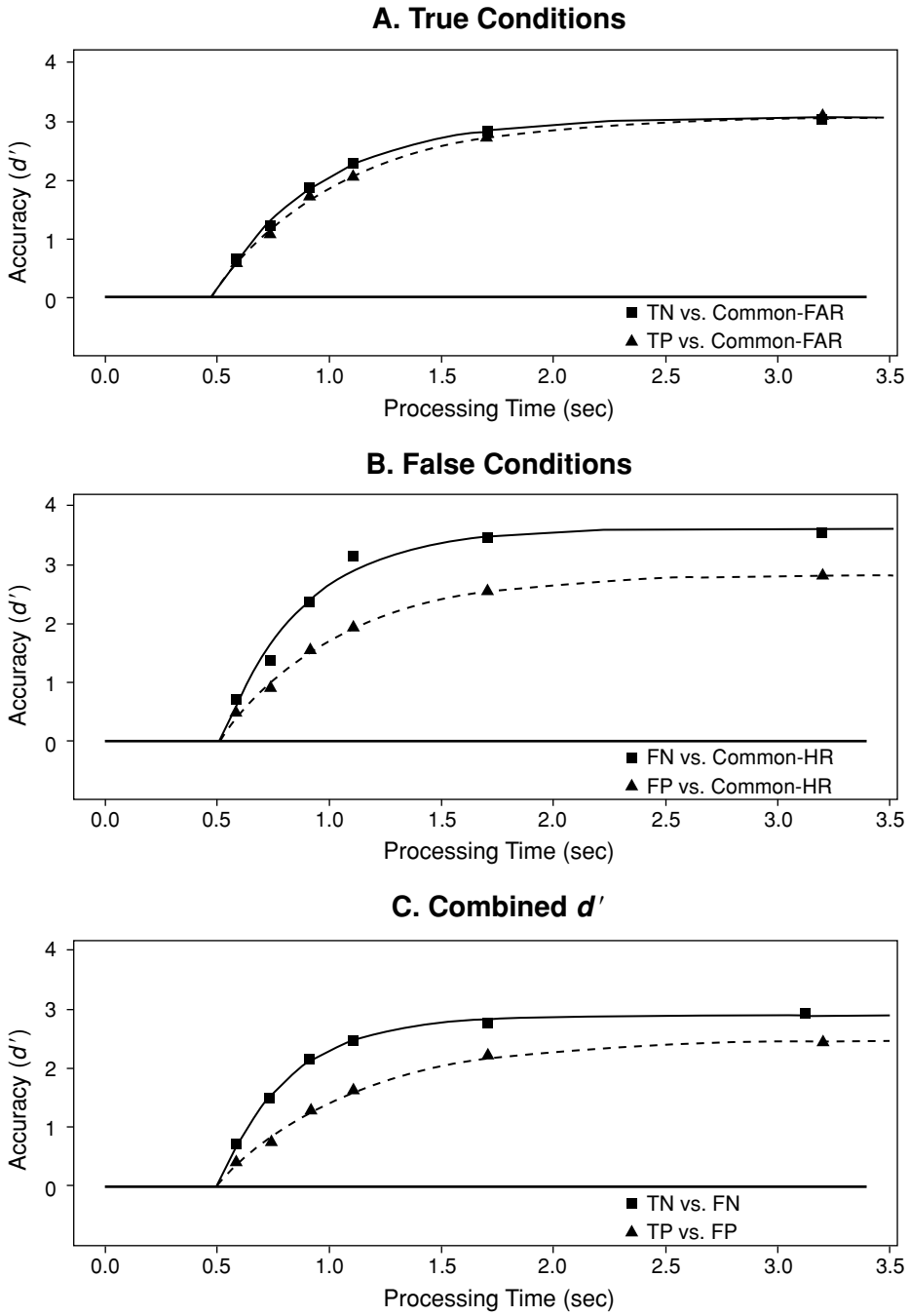

Figure 1. Average *d′* accuracy (symbols) as a function of processing time (lag of the response tone plus latency to respond to the tone). Panel A shows judgments of the TN (true of the noun; squares) and TP (true of the phrase; triangles) conditions. The solid and dashed lines show the best-fitting (1λ–2β–1δ) exponential (Equation 1) model (see text). Panel B shows judgments of the FN (false of the noun; squares) and FP (false of the phrase; triangles) conditions. The lines show the best-fitting (2λ–2β–1δ) exponential model (see text). Panel C shows the combined hit and false alarm differences between all conditions with standard *d′* scaling of the hit rate for the TN condition against the false alarm rate for the FN condition (squares) and hit rate for the TP condition against the false alarm rate for the FP condition (triangles). The lines show the best-fitting (2λ–2β–1δ) exponential model (see text). FAR, false alarm rate; HR, hit rate.

differences, with higher levels of accuracy in the TN condition than in the TP condition, suggesting that processing dynamics were faster in the TN than in the TP condition.

To quantify processing time-course, the data were fit with an exponential function:

$$d'(t) = \lambda(1-e^{-\beta(t-\delta)}), \text{ for } t > \delta, \text{ else } 0, \quad (1)$$

where λ is the asymptotic parameter reflecting discriminability at maximum processing time and β and δ are the parameters estimating function dynamics. The δ parameter describes the intercept at which performance begins to depart from chance, and the β parameter indexes the rate at which performance increases from chance to asymptote.

Following standard SAT procedures (see McElree et al., 2003, for details), differences between conditions were quantified by fitting the data with hierarchically nested models, which ranged from a null model (all conditions fit with single λ, β, and δ) to a fully saturated model (unique parameters for each condition), using an iterative hill-climbing least-squares algorithm. The quality of the fits was assessed by an adjusted R^2 statistic and by evaluating the consistency of the parameter estimates across participants. The analyses were performed on individual participant data; we report averaged data to summarize patterns across participants.

Models that allocated separate λ parameters to the TP and TN conditions did not improve the fit, nor were the λ estimates reliably different for the TP and TN conditions. However, accuracy differences at earlier processing times gave rise to clear differences in the SAT dynamics parameters, either in rate (β) or intercept (δ). We present here the better-fitting model, differing in rates. Varying the rate increased adjusted R^2 from .989 for a null 1λ–2β–1δ model to .997 for a 1λ–2β–1δ model (separate rates). Importantly, this model resulted in consistent and reliably ordered parameter estimates across participants. The rate was estimated at 2.08 in the TN condition versus 1.77 in the TP condition, a difference of 84 msec in $(1/\beta)$ units. Sixteen participants showed a rate advantage for the TN condition [$t(19) = 2.52$, $p = .02$].

The parameter estimates from both the 1λ–1β–2δ and the 1λ–2β–1δ models provide clear evidence that processing speed was faster in the TN than in the TP condition. The smooth functions in Figure 1A show the estimated functions for the 1λ–2β–1δ model.

Negative Conditions

To isolate differences in the false conditions for each participant, the z scores of the false alarm rates for the FP and FN conditions were scaled against a common hit rate derived from the average hit rate between the TP condition and the TN condition, to correct for possible response biases. Figure 1B shows these d' scalings for the average data, revealing large differences in asymptotic performance at the longest interruption point (times $>$ 3 sec). Consequently, model fits assigned separate λ parameters to the FP and FN conditions; 3.43 versus 3.12 in the average data [$t(19) = 9.17$, $p < .001$].

Table 2
Proportion of Yes Responses as a Function of the Latency of the Response Tone

	Latency of the Response Tone					
Condition	0.3	0.5	0.7	0.9	1.5	3.0
TP	.47	.63	.76	.81	.89	.94
TN	.55	.67	.78	.86	.91	.93
FP	.38	.33	.25	.21	.14	.12
FN	.28	.17	.11	.09	.04	.03

Note—TP, true of the phrase; TN, true of the noun; FP, false of the phrase; FN, false of the noun.

Additionally, we found clear evidence for differences in processing dynamics. A 2λ–2β–1δ model yielded the best description of the full time-course data, with an adjusted R^2 value of .978. β was estimated at 2.71 and 1.91 in the FP and FN conditions, a difference of 154 msec in $(1/\beta)$ units. Eighteen participants showed a rate advantage for the FN condition [$t(19) = 3.14$, $p < .01$]. The curves in Figure 1B show the estimated functions for the 2λ–2β–1δ model.

Alternative Scaling

To further highlight the processing disadvantage for properties that require the whole phrase to evaluate, Figure 1C presents d' scalings of the hit rate for each true condition against its respective false condition. The TN function clearly rises faster and reaches a higher asymptote than the TP function. A 2λ–2β–1δ model yielded the best description of this time-course data, with an adjusted R^2 value of .994. The asymptote was lower by 0.4 d' units and the rate slower by 270 msec for phrasal properties than for those that could be evaluated against the noun.

RT Study

We repeated the experiment in a simple RT paradigm in which participants judged the truth of the entire sentence with no response cues. The results mirrored the asymptotic results of the SAT experiment (see Table 3). Time to judge the TN and TP sentences did not differ [$t_1(11) = 1.27$, $p = .20$; $t_2(95) = 1.13$, $p = .13$], but the FN condition was faster than FP [$t_1(11) = 7.13$, $p < .001$; $t_2(95) = 3.36$, $p < .001$]. The percentage correct followed the same pattern, with higher accuracies for FN than for FP sentences [$t_1(11) = 7.71$, $p < .001$; $t_2(95) = 7.08$, $p < .001$] and no difference between TP and TN sentences [$t_1(11) = 1.79$, $p = .63$; $t_2(95) = .92$, $p = .36$].

Table 3
Reaction Time (in Seconds) and Percent Correct in Reaction Time Study

	TP	TN	FP	FN
Reaction time	1.407	1.335	1.485	1.257
Percent correct	88	87	92	97

Note—TP, true of the phrase; TN, true of the noun; FP, false of the phrase; FN, false of the noun.

DISCUSSION

By 2 sec after the presentation of the predicate, the participants were extremely accurate in verifying true properties, whether they were based on the noun alone or required integration with the modifier. This confirms the typicality ratings showing that the properties in the two conditions were equally true of the phrase. However, earlier in processing, the participants judged properties from the noun more accurately than properties requiring integration of the modifier and noun. This faster activation rate is contrary to the conclusions of Potter and Faulconer (1979) and Springer and Murphy (1992). That relevant phrase properties were not activated faster than less relevant noun properties indicates that further computation is required to derive emergent properties, and that properties of the noun that have nothing to do with the modifier are not suppressed early in processing. But although the emergent properties are slower to be activated, they eventually do reach the same level of activation as the noun properties.

For the false items, the difference was even more pronounced, in that the FP features were less accurate than the FN features at the asymptote and were processed at a slower rate. However, false features are less diagnostic, because it is difficult to be sure when something is "more false" than something else, and the asymptotic differences suggest that the FP features may have been less related to the phrase than the FN ones were. Nevertheless, the time-course difference between the FN and FP conditions suggests that noun properties are retrieved quickly even when they are not contextually relevant.

Why were noun properties activated faster than phrase properties, when previous studies found faster responses in the opposite direction? (Actually, Potter & Faulconer [1979] found only a 25-msec difference, and did not report its significance.) As noted earlier, Potter and Faulconer used noun pictures that were inconsistent with the phrase, possibly causing competition. In sentence verification tasks, pragmatic factors may have been involved (Gagné & Murphy, 1996; Glucksberg & Estes, 2000), but our results at the short delays were opposite to the pragmatic explanation, which claims that phrase properties are most relevant and are therefore processed faster. However, it is important to note that if we had tested only the longest delays, we would have concluded that noun and phrase features are equally activated, thereby giving evidence that selective retrieval is just as fast as nonselective retrieval. Indeed, that is just what we found in the RT version of our experiment. With the SAT technique, we were able to document that selective retrieval occurs later than retrieval of information from the noun alone.

One limitation of this work is the use of many isolated sentences. Selective retrieval may be faster in situations that provide stronger contextual support for a specific property. However, it is also possible that the context itself would activate this property prior to the critical word, and so at this moment, there is no strong evidence that readers selectively retrieve relevant information from semantic memory over irrelevant information. Rather, it seems more likely that the sorting-out of relevant information occurs at a later, integrative stage of sentence or discourse interpretation.

REFERENCES

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning & Verbal Behavior*, **13**, 471-481.

Gagné, C. L., & Murphy, G. L. (1996). Influence of discourse context on feature availability in conceptual combination. *Discourse Processes*, **22**, 79-101.

Glucksberg, S., & Estes, Z. (2000). Feature accessibility in conceptual combination: Effects of context-induced relevance. *Psychonomic Bulletin & Review*, **7**, 510-515.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory & Language*, **48**, 67-91.

McKoon, G., & Ratcliff, R. (1988). Contextually relevant aspects of meaning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 331-343.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Potter, M. C., & Faulconer, B. A. (1979). Understanding noun phrases. *Journal of Verbal Learning & Verbal Behavior*, **18**, 509-521.

Springer, K., & Murphy, G. L. (1992). Feature availability in conceptual combination. *Psychological Science*, **3**, 111-117.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behavior*, **18**, 645-659.

Tabossi, P. (1982). Sentential context and the interpretation of unambiguous words. *Quarterly Journal of Experimental Psychology*, **34A**, 79-90.

Tabossi, P., & Johnson-Laird, P. N. (1980). Linguistic context and the priming of semantic information. *Quarterly Journal of Experimental Psychology*, **32**, 595-603.

(Manuscript received July 27, 2005;
revision accepted for publication March 6, 2006.)