Journal of Memory and Language **41,** 469–495 (1999)
Article ID jmla.1999.2660, available online at http://www.idealibrary.com on IDEAL®

A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing

Kara D. Federmeier

Department of Cognitive Science, University of California, San Diego

and

Marta Kutas

Departments of Cognitive Science and Neuroscience, University of California, San Diego

The effects of sentential context and semantic memory structure during on-line sentence processing were examined by recording event-related brain potentials as individuals read pairs of sentences for comprehension. The first sentence established an expectation for a particular exemplar of a semantic category, while the second ended with (1) that expected exemplar, (2) an unexpected exemplar from the same (expected) category, or (3) an unexpected item from a different (unexpected) category. Expected endings elicited a positivity between 250 and 550 ms while all unexpected endings elicited an N400, which was significantly smaller to items from the expected category. This N400 reduction varied with the strength of the contextually induced expectation: unexpected, categorically related endings elicited smaller N400s in more constraining contexts, despite their poorer fit to context (lower plausibility). This pattern of effects is best explained as reflecting the impact of context-independent long-term memory structure on sentence processing. The results thus suggest that physical and functional similarities that hold between objects in the world—i.e., category structure—influence neural organization and, in turn, routine language comprehension processes. © 1999 Academic Press

Key Words: sentence processing; categorization; event-related potentials; N400.

At its heart, language comprehension involves the recruitment and integration of world knowledge stored in long-term memory. Consider, for example, the following pair of sentences:

> "Getting himself and his car to work on the neighboring island was time consuming. Every morning he drove for a few minutes and then boarded the . . . "

When asked, most individuals report that they expect the missing final word of this sentence pair to be "ferry." How do they come to that expectation? None of the individual words in this sentence pair is strongly associated with the word *ferry.* There are, in fact, a number of different vehicle names that could plausibly complete the sentence. Yet, the expectation is remarkably consistent across individuals. In order to create their expectations, readers must use information in the sentence context to build a cognitive model involving vehicles that can transport both people and cars across water and that are likely to be used habitually. It is their store of world knowledge, combined with this model, that allows readers to then determine that the vehicle in question is likely to be a ferry and not an ocean liner, barge, airplane, or helicopter. Given how crucial long-term memory is for language processing, it is somewhat surprising that so little is known about how information from memory is accessed and used during on-line language processing.

Context Effects in Language

Off-line tasks make it clear that there is often sufficient information in a sentence context to

This work was supported by a Howard Hughes Predoctoral Fellowship to K.F. and grants HD22614, AG08313, and MH52893 to M.K. We thank Elizabeth Bates, Seana Coulson, Phillip Holcomb, Jonathan King, David Kirsh, Marty Sereno, and two anonymous reviewers for helpful comments and discussion.

Address reprint requests to Kara D. Federmeier, Department of Cognitive Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0515. E-mail: kfederme@cogsci.ucsd.edu; Fax: 619-534-1128.

0749-596X/99 $30.00

significantly constrain guesses about what concepts or even words are likely to be next encountered. However, whether—and, if so, when and how—this information affects processing on-line remains a hotly debated issue. Much psycholinguistic research suggests that words that are predictable in a sentence context are perceived and processed more rapidly and accurately than the same words when they occur out of context or in incongruent contexts. For example, contextual information decreases the duration of readers' eye fixations (Ehrlich & Rayner, 1981; Morris, 1994; Zola, 1984). Congruent contexts also facilitate the time to pronounce sentence-final or phrase-final words (Duffy, Henderson, & Morris, 1989; Hess, Foss, & Carroll, 1995; McClelland & O'Regan, 1981; Stanovich & West, 1983) and the speed of word/nonword judgments (lexical decision) on them (Fischler & Bloom, 1985; Kleiman, 1980; Schuberth, Spoehr, & Lane, 1981). This facilitation occurs even when the relatedness of lexical items within congruent and incongruent sentences is matched, suggesting that the observed increase in processing fluency cannot be attributed solely to lexical priming, but involves information provided by the sentence as a whole (e.g., Duffy et al., 1989; Morris, 1994; Ratcliff, 1987).

Electrophysiological results support these findings and suggest that contextual information is used early and builds continuously over the course of processing a sentence. The event-related brain potential (ERP) technique involves recording at the scalp neural activity that is time-locked to a particular event. The neural activity recorded is known to reflect the summation of graded postsynaptic potentials, predominantly from pyramidal cells of the cerebral cortex (for review, see Kutas & Dale, 1997). The ERP technique provides a continuous, multidimensional measure that can be recorded during natural language processing (without the imposition of an additional task). It provides millisecond-level temporal resolution and information about the number and, in some cases, location of the neural sources contributing to a given task or condition (e.g., Rugg & Coles, 1995).

An ERP component that has proven especially useful for the study of contextual influences in language processing is the N400, a negative-going potential peaking around 400 ms after stimulus onset. Kutas and Hillyard (1980b) first observed the N400 during a task in which individuals read sentences word by word for comprehension. Sentence final words that were semantically anomalous with respect to the sentence context were associated with a significantly larger negativity 250 to 600 ms post-stimulus-onset than were words that fit the sentence context. Subsequent investigations have revealed that each word in a sentence elicits an N400 and that the amplitude of this component is highly correlated with individuals' off-line expectations as measured by "cloze probability[1]" (Kutas & Hillyard, 1984) and decreases as contextual information builds over the course of a sentence (Van Petten & Kutas, 1990).

The influence of contextual information on word processing has been most clearly demonstrated for words that are highly predictable in their sentence contexts ("best completions"; i.e., words with the highest cloze probability in the context). However, to a more limited degree, contextual information has also been found to affect the processing of less predictable words. With behavioral techniques, for example, some researchers find facilitation for unexpected but contextually congruous words (e.g., Schwantes, 1985; Stanovich & West, 1983); others, however, do not (e.g., Fischler & Bloom, 1979; Kleiman, 1980; Schwanenflugel & LaCount, 1988). In electrophysiological investigations, these congruent but low cloze probability items elicit N400 responses that are larger than those to higher cloze probability items but smaller than those to contextually incongruent items (e.g., Kutas & Hillyard, 1984). Both behavioral and electrophysiological studies have also observed facilitation for unexpected items that are semantically related to the best completion (Kleiman, 1980; Kutas & Hillyard, 1984; Kutas,

[1] The cloze probability of a word in a given context refers to the proportion of people who would choose to complete that particular sentence fragment with that particular word (Taylor, 1953).

Lindamood, & Hillyard, 1984; Schwanenflugel & LaCount, 1988); these effects can be observed even for words that do not form acceptable sentence completions (Kleiman, 1980; Kutas & Hillyard, 1984; Kutas et al., 1984).

The types of words facilitated by a context and the degree of facilitation for each seem to vary with the nature of the context itself. For example, highly constraining contexts seem to provide greater facilitation of "best completions" than do less constraining contexts (Fischler & Bloom, 1979; McClelland & O'Regan, 1981). At the same time, however, highly constraining contexts have a narrower "scope" of facilitation that does not extend to less predictable items (e.g., Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985). Less constraining contexts, on the other hand, facilitate a wider range of items and provide enhanced facilitation for less predictable items. Research has also shown that with greater semantic-associative information (i.e., more words that are lexically associated with a target) in a context, one observes greater facilitation of contextually congruent words (Duffy et al., 1989) and more elaborative inference-drawing (McKoon & Ratcliff, 1989); this factor has not always been controlled for in other studies looking at, for example, effects of context on contextually incongruent, semantically associated targets.

Taken together, this body of work suggests that sentence contexts facilitate, in a graded manner, the processing of a set of concepts and/or words. Moreover, the nature and strength of the sentence context affects what items/concepts are facilitated and to what extent. However, it cannot be information in the sentence context alone[2] that determines what is or is not facilitated, as at times facilitation has been observed for contextually inappropriate (but semantically related) items but not observed for unexpected but contextually congruent ones (Schwanenflugel & LaCount, 1988). Because language comprehension crucially relies on information stored in long-term memory, we hypothesized that the structured nature of this memory is another significant—but relatively unexplored—variable likely to be affecting how words are processed during reading.

The N400 and Long-Term Memory

The hypothesis that the organization of semantic memory plays an integral role in determining how information in a sentence context will affect word processing receives support from the observation that N400 effects, while insensitive to nonsemantic manipulations of context (e.g., changes in the physical attributes of words (Kutas & Hillyard, 1980a) or grammatical and morphological violations (Kutas & Hillyard, 1983)) or deviations in nonlinguistic stimuli (e.g., anomalous notes in melodies (Besson & Macar, 1987)), do seem to be sensitive to long-term memory processes. N400 components have been recorded during investigations of recognition memory for both words (Neville, Kutas, Chesney, & Schmidt, 1986; Smith, Stapleton, & Halgren, 1986) and pictures (Friedman, 1990). Some studies have claimed to observe N400-like components during memory tasks involving stimuli that are not particularly semantic in nature. For example, Stuss et al. (1986) reported an N400-like component whose amplitude varied with the number of pictures to be remembered in a continuous recognition–memory task. Chao et al. (1995) also reported what they describe as an N400 effect to environmental noise stimuli, but only during conditions involving repetitions after long delays. Both groups suggest that their findings implicate the N400 component in search through long-term memory.

Studies into the neurobiological basis of the N400 effect also support a link between this component and long-term memory processes. McCarthy et al. (1995; also Nobre & McCarthy, 1995) recorded field potentials from intracranial electrodes implanted in humans undergoing treatment for epilepsy as they read sentences for comprehension. Anomalous sentence endings were associated with large field potentials in the

[2] Of course, no sentential context effects are wholly independent of long-term memory, as context effects necessarily derive from information stored in memory. When we speak of the influence of sentence context alone, we refer to the kinds of sentence context effects that would be expected even if memory were unstructured.

left and right anterior medial temporal lobes. The authors suggested that these potentials were generated in anterior fusiform and parahippocampal gyri and perhaps the hippocampus proper. Grunwald et al. (1995) also showed that the presence or absence and amplitude of field potentials recorded from the left anterior medial temporal lobe were correlated with performance on a delayed word recognition task. As medial temporal lobe structures are considered critical for successful performance on declarative memory tasks (e.g., Squire, 1987), a medial temporal lobe source for at least some of the N400 activity at the scalp lends credence to the idea that the N400 may index semantic memory involvement as a word is integrated with previous context.

Insofar as context effects—and associated N400 effects—derive from perceivers' knowledge about the world and the access of that knowledge from memory, on-line language processing should be influenced by the structure of that memory. The structure of semantic memory may be based on many factors, but it seems likely that an important part of its organization could involve the kind of featural similarity structure that has been observed to underlie human categorization (and information representation in the brain; see, e.g., Tanaka, 1996, for a striking example from higher order visual representation). Categorization research suggests that many human categories are *taxonomic*: items are grouped together on the basis of shared perceptual and functional attributes (e.g., Kay, 1971; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and these groupings occur at multiple levels of generality, similar to biological taxonomies. In this scheme, category membership is graded, determined by whether —and how many—attributes an item shares with other members of a category (e.g., Rosch, 1975; Rosch & Mervis, 1975; Rosch, 1973). In turn, there also seems to be a structured, graded organization of categories themselves. For example, a particular item may be called "a plant," "a flower," or "a rose"—each a category itself, albeit with successively decreasing inclusiveness.

Consistent with a relationship between taxonomic categories, long-term memory, and context effects, electrophysiological studies have shown that N400 amplitude is sensitive to category membership. For example, in studies by Polich (1985) and Harbin et al. (1984) volunteers were shown a series of words belonging to a particular taxonomic category. The word series ended either with another member of the category or with a word from a different category. Both groups found that a final word that was not a category member generated more N400 activity than did a category member. Category effects on the N400 have also been observed during the performance of sentence verification tasks (in which individuals are asked to judge the truth of statements of the form "An *X* is a *Y*"—for example, "A robin is a bird") (e.g., Fischler, Bloom, Childers, Roucos, & Perry, 1983; Kounios, 1996; Kounios & Holcomb, 1992). In these tasks, an enlarged N400 response is observed to the final word of false statements such as "A carrot is a fruit." Revealingly, in some cases large N400 responses are also observed at the end of true statements such as "A carrot is not a fruit" (e.g., Fischler et al., 1983). In other words, at least in some contexts, the categorical relationship between the subject and object actually seems to be a more reliable predictor of the N400 response than the fit of the item in the sentence context itself (though effects of propositional truth on the N400 have also been reported (Fischler, Bloom, Childers, Arroyo, & Perry, 1984; Fischler, Childers, Achariyapaopan, & Perry, 1985)).

The Present Study

Taken together, behavioral and electrophysiological evidence suggest that the impact of a sentence context on a word's processing may be influenced by and interact with the structure of knowledge in long-term memory—a structure that is likely based, at least in part, on the perceptual and functional similarity captured by semantic categories. In fact, the first influences of both semantic context and category membership on lexical processing are manifest as a change in the amplitude of the same ERP component—the N400. Therefore, we can use the N400 to examine the extent to which long-term memory structure interacts with contextual information during on-line sentence processing. In particular, in this study we examined whether readers' processing would be affected by memory structure even when that structure was irrelevant to the language comprehension task. In addition, we aimed at getting a better understanding of the role of memory structure in reading by comparing its influence when sentence contexts are strong versus when they are weaker.

We addressed these issues by comparing the effects of two types of contextual violations: (1) those that come from the same semantic category as the contextually predicted item and thus share many features in common with it ("within-category violations") and (2) those that come from different semantic categories and thus share far fewer features in common with the predicted item ("between-category violations"). ERPs were recorded as volunteers read pairs of sentences. Each sentence pair was designed to create an expectation for a specific exemplar of a specific category (e.g., "They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of . . . "). The second sentence of the pair ended with either (1) the expected exemplar ("palms"), (2) an unexpected exemplar from the same category as the expected exemplar ("pines"), termed the within-category violation, or (3) an unexpected exemplar from a different category than the expected exemplar ("tulips"), termed the between-category violation. It is important to note that exemplars for the between-category violations were still members of a shared higher-level category (e.g., "plants") and therefore match the other two items on most general dimensions. Items rotated roles across sentences such that they served as each type of ending once across the stimulus set, as the following example illustrates:

1. They wanted to make the hotel look more like a tropical resort. So along the driveway, they planted rows of *palms/pines/tulips.*

2. The air smelled like a Christmas wreath and the ground was littered with needles. The land in this part of the country was just covered with *pines/palms/roses.*

3. The gardener really impressed his wife on Valentine's Day. To surprise her, he had secretly grown some *roses/tulips/palms.*

4. The tourist in Holland stared in awe at the rows and rows of color. She wished she lived in a place where they grew *tulips/roses/pines.*

The sentence contexts varied in their constraint, the degree to which they led to a strong (consistent) expectation for the best completion.

Comparing the pattern of ERP results obtained when individuals read sentences with and without violations of the two types should help unravel the importance of sentence context information and semantic memory structure for language comprehension. Previous work suggests that the best completions (i.e., the expected exemplars) will elicit a positivity between 300 and 500 ms. By contrast, the between-category violations, which are contextually unexpected, difficult to integrate, and share few features in common with the best completions, will likely elicit an N400 in the same time window (e.g., Kutas & Hillyard, 1980b). What is unknown from previous work is how the response to the within-category violations will compare with the response to expected exemplars and between-category violations.

Within-category violations are similar to expected exemplars in that they share many semantic features in common. Therefore, if (at the level of processing indexed by the N400) the system is sensitive only to a fairly general feature match between an item and a sentence context, one might expect a similar amplitude to expected exemplars and within-category violations. A difference between expected exemplars and within-category violations would suggest that the system is sensitive to more specific contextual information (the kind that allows individuals, off-line, to predict the expected exemplar but not the within-category violation).

Alternatively, if the system is sensitive to specific contextual information—and that alone—one would expect an N400 of similar amplitude to both within- and between-category violations. Neither within- nor between-category violations are expected. Moreover, both are relatively implausible in their sentence contexts. There is thus no reason for them to differ based on their fit to context alone. A smaller N400 amplitude to the within- relative to between-category violations would therefore suggest that long-term memory is structured by feature similarity as reflected in semantic categories and that, independent of sentence context, this structure impacts on-line sentence processing. If within-category violations elicit an N400 of intermediate amplitude (greater than expected but less than between), it would suggest that the system is sensitive both to specific contextual information and to the relationship (feature overlap) between concepts in long-term memory.

Finally, if, in fact, long-term memory structure affects on-line language processing, one can examine its interaction with sentence context information by comparing the impact of memory structure when sentential context information is strong (in highly constraining sentences) with its impact when context is weaker (as in less constraining sentences).

METHODS

Materials

Stimulus material consisted of 132 pairs of sentences, each ending with three target words: (1) the expected exemplar, the highest cloze probability ending for a given sentence pair, (2) the within-category violation, an unexpected (cloze probability < 0.05) exemplar from the same taxonomic category as the expected exemplar[3], and (3) the between-category violation, an unexpected (cloze probability < 0.05) exemplar from a different category than the expected exemplar. The first sentence of each sentence pair established the expectation for item and category[4]. In contrast, the second sentence, when separated from the first, could be plausibly completed by any of the three possible targets. There were no lexical associates of any of the possible endings within the sentence containing the target word.

Target items were pictureable objects from 66 categories (two items from each). Categories were chosen to be those at the lowest level of inclusion for which the average undergraduate student could be expected to readily differentiate several exemplars. For approximately half the categories used, this level was basic as determined by Rosch et al. (1976) or by analogy (e.g., tree, fish, bird, cat, dog, pants, shoes, shirt, lamp, and car were all determined to be basic level by Rosch et al. and flower, rodent, bear, boat, insect, dinosaur, cheese, bread, etc. would seem to be so as well). Other categories were based at what Rosch et al. would have defined as the next highest level (a superordinate of the basic level) because it was unclear that the average participant could clearly and consistently differentiate below this level (e.g., vegetable (different types of carrots?), sports equipment (different types of bats?))[5]. Between-

[3] While these items came from the same category, they were generally not lexical associates (only 10/132 had a lexical association greater than 0.1 according to the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973)).

[4] Forty-two out of 132 of these first-sentence contexts contained a word lexically associated at a level of 0.1 or greater (Edinburgh Associative Thesaurus (Kiss et al., 1973)) with the expected exemplar.

[5] Rosch et al. (1976) used four criteria to determine the basic level: the basic level is the most inclusive level whose members (1) possess significant numbers of attributes in common, (2) have similar motor programs, (3) have similar shapes, and (4) can be identified from the average shape. However, these criteria can be difficult to apply to some categories and may yield conflicting results for others. For example, “country” is a category for which individuals can name a large number of exemplars and for which there does not seem to be a lower level other than exemplar. But it is doubtful that the category “country” can be identified from its average shape or what it would mean to say that its members are interacted with via similar motor programs. While “book” may fulfill the definition for basic, it is also the case that the category “reading material” (books, magazines, newspapers) can likely be identified from its average shape and that its members have similar shapes and are interacted with via similar motor programs. Other items, such as “hammer” clearly seem to be basic by these criteria; however, the average undergraduate probably cannot actually make (at least verbal) differentiations below this level. For the purposes of this study, therefore, to have enough categories we were forced to go to the next highest level

category targets for each sentence pair were chosen from a related category that shared key features (e.g., animacy, size, general function) with that from which the expected exemplar and within-category violation were derived. Appendix A lists the categories and their pairings[6].

Target items were rotated across the stimulus set such that each item appeared three times, once as each kind of ending. Thus, across the experiment target conditions were perfectly controlled for length, frequency, imageability, and concreteness; context sentences in each condition were also perfectly controlled for length and grammatical complexity. The experimental sentences were divided into three lists of 132 sentences each. Sentence contexts and items were used only once per list; each list consisted of 44 of each type of target (expected exemplars, within-category violations, between-category violations). Within each list, the three target conditions were matched for mean word length and frequency. To balance the number of plausible and implausible sentences read by each participant, 44 plausible filler sentence pairs were added to each list. Stimuli were randomized once within each list and then presented in the same order for each participant. Appendix B gives examples of the stimuli.

Cloze Procedure

Cloze probabilities were obtained for the 132 sentence pair contexts (sentence pairs missing the final word of the second sentence). These were divided into two lists so that the two sentence contexts presumed to be predictive of items coming from the same category did not both appear on the same list. Student volunteers were asked to complete each sentence pair with “the first word that comes to mind.” List one was completed by 56 students and list two was completed by a different set of 59 students. A subset of the original stimuli was rewritten and clozed separately by a third group of 55 students. Cloze probability for a given word in a given context was calculated as the proportion of individuals choosing to complete that particular context with that particular word. Expected exemplars were always the item with the highest cloze probability for a given context. Mean cloze probability for the expected exemplars was 0.74. Within category violations and between category violations always had cloze probabilities of less than 0.05. Mean cloze probability was 0.004 for the within-category violations and 0.001 for the between-category violations.

Constraint

Although all expected exemplars were items with the highest cloze probability for their sentence contexts, the actual cloze probability of these items ranged from 0.17 to 1.0. In other words, the sentence contexts differed in their constraint, or the degree to which they led individuals to strongly expect one particular item versus a number of different items. To examine the effects of sentential constraint on the ERP response to target items, we divided the sentences into two groups, “high constraint” and “low constraint,” by a median split on the cloze probability of the expected exemplar. For the high constraint sentences, the cloze probability of the expected exemplars had a range of 0.784 to 1.0 and an average value of 0.896 (median = 0.904). For the low constraint sentences, the cloze probability of the expected exemplars had a range of 0.17 to 0.784 and an average value of 0.588 (median = 0.608). High constraint sentences are thus those in which there is a single, highly preferred ending, while low constraint sentences are those that are compatible with a larger range of ending types and in which the expected exemplar has at least one, and generally several, close competitors. Word frequency and word length were controlled across all constraint and ending type conditions[7].

(carpentry tools) in order to have clearly differentiable items.

[6] Items of a category vary in their “typicality”—that is, the degree to which they are judged to be representative of the category to which they belong. Typicality was not manipulated in this study, and the set of items used contains both typical and atypical exemplars.

[7] Average values for word frequency/word length split by ending type and constraint were: high constraint, expected (18.6/6.0); low constraint, expected (21.2/6.0); high con-

Plausibility Ratings

The two violation types were both kept below a cloze probability of 0.05. While this indicates that none of the violations was considered a good ending for the sentence context, it leaves open the question of whether one of the violation types might have been, on average, a more plausible ending. To determine this, a different group of student volunteers was asked to rate the plausibility of all of the endings within their experimental sentence contexts. The sentences were split into the same three lists used in the actual ERP experiment so that no item or context was repeated within a list. Volunteers were asked to rate how much "sense" each sentence pair made on a percentage scale (where 0% meant the sentence pair "makes no sense at all (is very implausible)" and 100% meant the sentence pair "makes perfect sense (is very plausible)"). Lists one, two, and three were completed, respectively, by 18, 21, and 18 student volunteers; none of these individuals participated in either the cloze probability ratings or the ERP experiment.

Mean rated plausibility was calculated by averaging the plausibility ratings for all items of a given condition within each participant and then averaging the scores across participants. Expected exemplars had a mean plausibility rating of 95.6%, within-category violations had a mean plausibility rating of 28.3%, and between-category violations had a mean plausibility rating of 15.3%. These plausibility measures were subjected to an omnibus analysis of variance (ANOVA) with repeated measures on three levels of Ending Type (expected exemplars vs within-category violations vs between-category violations), revealing a significant effect of Ending Type [($F(2,112) = 2738.25$; $p < .001$]. Planned comparisons showed that expected exemplars were rated as significantly more plausible than within-category violations [$t = 46.06$; $p < .001$] and within-category violations as more plausible than between-category violations [$t = 15.75$; $p < .001$].

Plausibility ratings for the violation types after splitting the sentences by constraint are shown in Table 1.

TABLE 1

	High constraint	Low constraint
Expected exemplars	97.7%	93.5%
Within-category violations	23.6%	30.2%
Between-category violations	11.9%	18.7%

An omnibus ANOVA with repeated measures performed on two levels of Constraint (high vs low) and three levels of Ending Type (expected exemplars vs within-category violations vs between-category violations) revealed significant main effects of both Constraint [$F(1,56) = 3472.05$; $p < .001$] and Ending Type [$F(2,112) = 1369.32$; $p < .001$] and a significant Constraint by Ending Type interaction [$F(2,112) = 994.49$; $p < .001$]. Rated plausibility significantly increased for expected exemplars in high versus low constraint contexts [$t = 5.00$; $p < .001$] but decreased for both within-category violations [$t = 3.54$; $p < .001$] and between-category violations [$t = 8.21$; $p < .001$] in high versus low constraint contexts. In other words, the pattern of plausibility ratings was congruent with claims from the behavioral literature (Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985) that more highly constraining contexts allow greater integration of best completions but reduced integration of improbable completions.

Participants

Eighteen UCSD undergraduate volunteers (10 men and 8 women, 18 to 24 years of age, mean age 20) participated in the experiment for credit and/or cash (none of these took part in any of the norming procedures). All were right-handed (as assessed by the Edinburgh Inventory (Oldfield, 1971)), monolingual English speakers with no history of reading difficulties or neurological/psychiatric disorders; five of the volunteers reported having a left-handed family

straint, within (21.0/5.9); low constraint, within (18.8/6.2); high constraint, between (18.7/6.0); low constraint, between (21.1, 6.0). Word frequency information was obtained from Francis and Kucera (1982).

member. Six participants were randomly assigned to each of the three stimulus lists.

Experimental Procedure

Volunteers were tested in a single experimental session conducted in a soundproof, electrically shielded chamber. They were seated in a comfortable chair approximately 60 cm in front of a monitor and instructed to read the stimulus sentences for comprehension. They were informed at the start of the experiment that they would be given a recognition memory test over the stimuli at the conclusion of recording. The session began with a short practice trial designed to reiterate the experimental instructions and to acclimate volunteers to the experimental conditions and the task. Each trial began with the first sentence of a sentence pair appearing in full on a CRT. Volunteers read this sentence at their own pace and pushed a button to view the second sentence. Presentation of the second sentence was preceded by a series of crosses to orient the volunteer toward the center of the screen. The second sentence was then presented one word at a time in the center of the screen. Nonsentence final words were presented for a duration of 200 ms with a stimulus-onset asynchrony of 500 ms. Sentence final words were presented for a duration of 500 ms. Volunteers were asked not to blink or move their eyes during the second sentence. The final, target word was followed by a blank screen for 3000 ms, after which the next sentence appeared automatically. Volunteers were given a short break after every 17 pairs of sentences. At the conclusion of the recording session, participants were given a recognition memory test consisting of 50 sets of sentence pairs: 10 new ones, 20 unchanged experimental pairs (of which 10 ended with expected exemplars, 5 ended with within-category violations, and 5 ended with between-category violations), and 20 modified sentence pairs in which the final word had been changed from that originally viewed by the volunteer (10 in which violations had been changed to expected exemplars and 10 in which expected exemplars had been changed to violations). Volunteers were instructed to classify the sentences as new, old, or similar (changed).

EEG Recording Parameters

The electroencephalogram (EEG) was recorded from 26 tin electrodes embedded in an Electro-cap, referenced to the left mastoid. Electrode sites are shown on Fig. 1. These sites included midline prefrontal (MiPf), left and right medial prefrontal (LMPf and RMPf), left and right lateral prefrontal (LLPf and RLPf), left and right medial frontal (LMFr and RMFr), left and right mediolateral frontal (LDFr and RDFr), left and right lateral frontal (LLFr and RLFr), midline central (MiCe), left and right medial central (LMCe and RMCe), left and right mediolateral central (LDCe and RDCe), midline parietal (MiPa), left and right mediolateral parietal (LDPa and RDPa), left and right lateral temporal (LLTe and RLTe), midline occipital (MiOc), left and right medial occipital (LMOc and RMOc), and left and right lateral occipital (LLOc and RLOc). Blinks and eye movements were monitored via electrodes placed on the outer canthus (left electrode serving as reference) and infraorbital ridge of each eye (referenced to the left mastoid). Electrode impedances were kept below 5 KΩ. EEG was processed through Grass amplifiers set at a bandpass of 0.01–100 Hz. EEG was continuously digitized at 250 Hz and stored on hard disk for later analysis.

Data Analysis

Data was rereferenced off-line to the algebraic sum of the left and right mastoids. Trials contaminated by eye movements, excessive muscle activity, or amplifier blocking were rejected off-line before averaging; approximately 10% of trials were lost due to such artifacts. Blinks were corrected via a spatial filter algorithm devised by Dale (1994). ERPs were computed for epochs extending from 100 ms before stimulus onset to 920 ms after stimulus onset. Averages of artifact-free ERP trials were calculated for each type of target word (expected exemplars, within-category violations, between-category violations) after subtraction of the 100 ms prestimulus baseline.

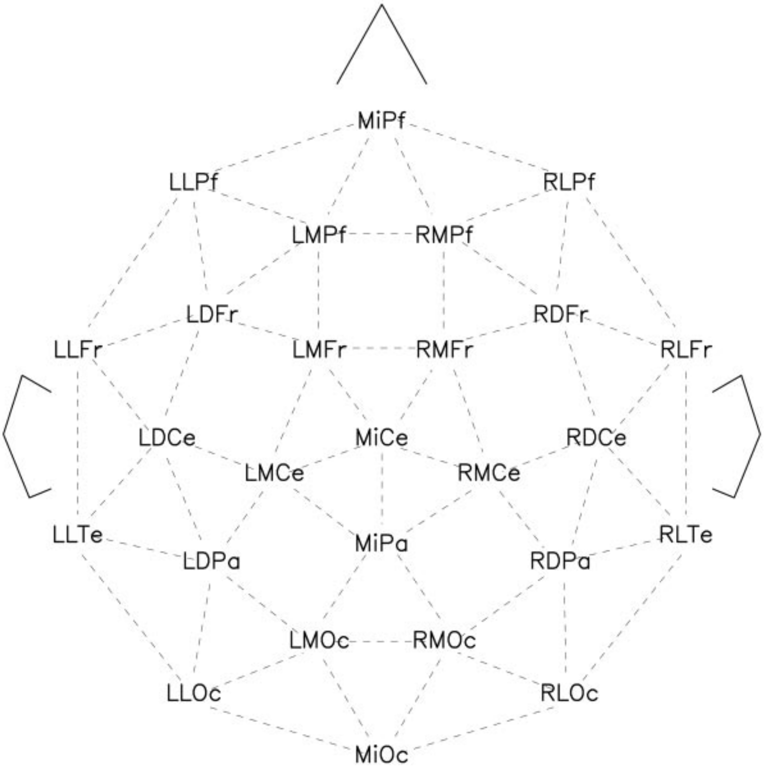

FIG. 1. Schematic of the electrode array used in the experiment. In all, 26 scalp electrodes were employed, arranged in a series of four equally spaced concentric rings.

RESULTS

Behavior

On average, volunteers correctly classified 88% (range 74–100%) of the items on the recognition memory test. The most common type of error was a misclassification of "similar" sentences (those in which only the final word had been altered from that actually shown in the experiment) as "old," followed by a misclassification of "old" sentences (those seen in the same form during the recording session) as "similar." Together, these two error types account for 73% of all errors observed. Most of the remainder of the errors consisted in volunteers classifying "old" or "similar" sentences as "new" (average number of either of these type of errors was less than one). Only two errors in which "new" sentences were classified as "old" or "similar" (one each) were observed across the 18 participants. The behavioral results indicate that the experimental sentences were attended during the recording session.

ERPs

Grand average ERPs (across all 18 volunteers) to sentence final words from all recording sites are shown in Fig. 2. Early components in all conditions include, at posterior sites, a positivity peaking around 110 ms (P1), a negativity peaking around 180 ms (N1), and a positivity peaking around 280 ms (P2), and, at frontal sites, a negativity peaking around 130 ms (N1) and a positivity peaking around 230 ms (P2). Early components are followed, in the expected exemplar condition, by a broad late positivity, largest over central and posterior sites, and, in the two violation conditions, by a negativity peaking around

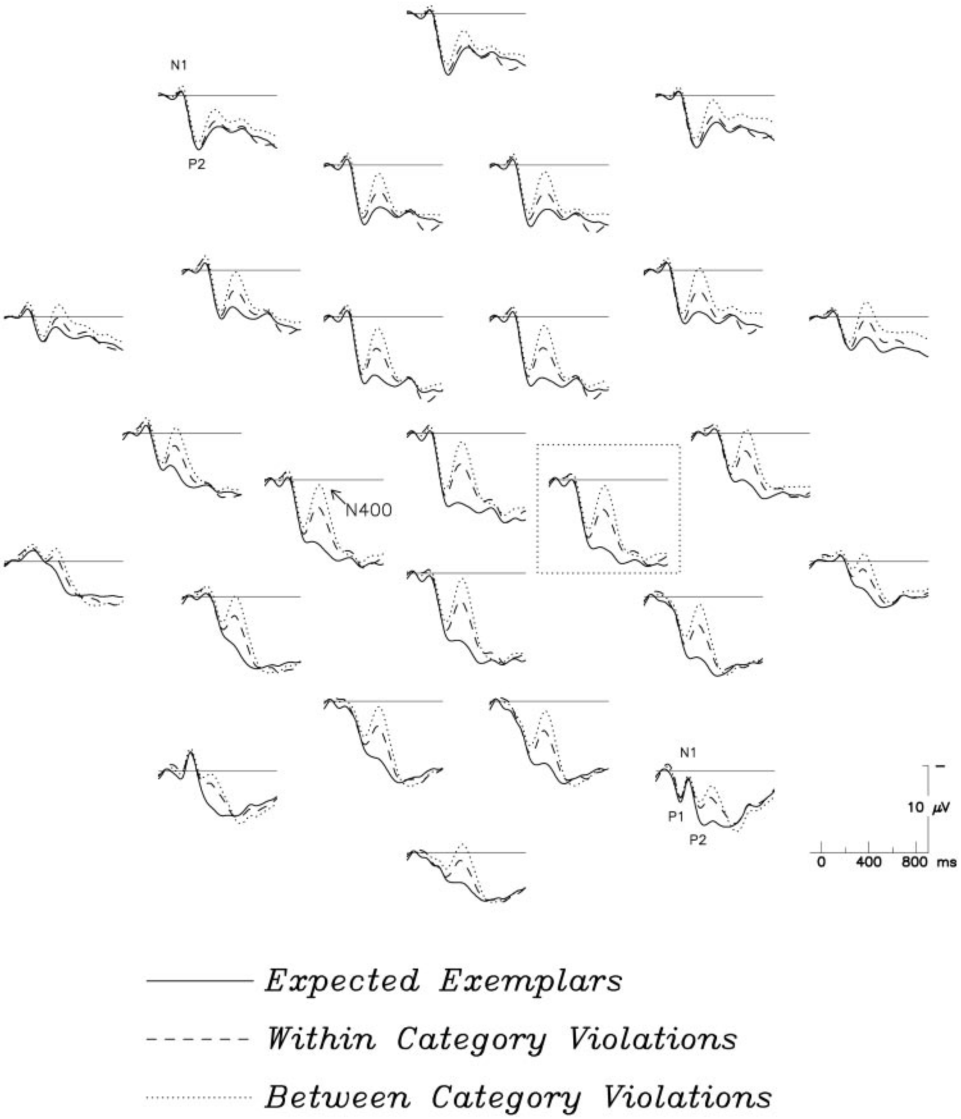

FIG. 2. Grand average ($N = 18$) ERP waveforms for the three ending types shown at all 26 electrode sites. Negative is plotted up. The ending types are characterized by the same set of early components. In the 350- to 450-ms time window, expected exemplars (solid line) showed a sustained positivity while both within-category violations (dashed line) and between-category violations (dotted line) showed a negativity, the N400, which was larger for the between category violations. A box around the right medial central site indicates the electrode that will be used in subsequent figures.

400 ms (N400), also largest over central and parietal sites. The N400 in the two violation conditions is followed by an extended late positivity of similar amplitude to that observed for the expected exemplars.

Peak Latency of the N400 Response

In order to determine the appropriate window for mean amplitude analyses and to ascertain that the latency of the N400 did not differ across conditions, latency of the largest negative peak between 350 and 450 ms was measured for each condition in each participant and subjected to an omnibus ANOVA. Repeated measures included three levels of Ending Types (expected exemplars vs within-category violations vs between-category violations) and 26 levels of Electrode. All p-values in this and all subsequent analyses are reported after Epsilon correction (Greenhouse–Geisser) for repeated measures with greater than one degree of freedom.

Mean peak latency (in milliseconds) is 385 for the expected exemplars (though, in fact almost no N400 is observed in this condition), 377 for the within category violations, and 375 for the between category violations. The effect of Ending Type was not statistically significant [$F(2,34) = 2.15$; $p = .157$] and did not interact with the effect of electrode [$F(50,850) = 0.92$; $p = .511$].

Mean Amplitude Analyses

Based on the peak latency analysis, mean voltage measures were taken in a 50-ms window around 375 ms (i.e., 350–400 ms post-stimulus-onset). These measures were subjected to an omnibus ANOVA. List (three levels) was a between-subjects variable while repeated measures included three levels of Ending Type (expected exemplar vs within-category violation vs between-category violation) and 26 levels of Electrode.

Effects of list. While there was no main effect of List [$F(2,15) = 0.67$; $p = .524$], a significant List × Ending Type interaction was observed [$F(4,30) = 4.59$; $p = .005$]. Examination of the means revealed no difference in the qualitative pattern of ending type effects; in all cases expected exemplars have the most positive mean voltage in this time window and between-category violations have the least positive mean voltage. The interaction seemed to be caused by differences in the range of these difference, which are largest in list 3 (7.26, 3.66, and −0.22 μV for expected exemplars, within-category violations, and between-category violations, respectively), and smallest in list 2 (list 1: 5.52, 3.10, 0.88 μV; list 2: 3.48, 1.44, 0.80 μV). List was also found to interact with Electrode [$F(50,375) = 4.14$, $p = .002$], indicating that the distribution of the N400 effect is slightly different across lists and thus across individuals. Given that individuals' brainwaves vary and that there were a relatively small number of participants per list ($n = 6$), these variations in the overall size and distribution of effects did not seem important for the questions addressed by this study; lists were therefore grouped together for subsequent analyses.

Effects of ending type. A main effect of Ending Type was observed [$F(2,30) = 55.03$, $p < .001$], as was an Ending Type × Electrode interaction [$F(50,750) = 7.04$, $p = .001$][8]. Figure 3 shows this effect of Ending Type, in which the smallest N400 amplitude is observed to the expected exemplars, which are very positive in this time window, and the largest negativity is observed to the between-category violations. The N400 to the two violation types tended to be largest over central and posterior sites, larger medially than laterally, and slightly larger over the right than the left hemisphere sites. Planned comparisons were conducted via an omnibus ANOVA on two levels of Ending Type (between-category violation vs within-category violation and within-category violation vs expected exemplar) and 26 levels of Electrode. Between-category violations were significantly more negative than within-category violations [$F(1,17) = 17.06$, $p = .001$] across the scalp. Additionally, within-category violations were significantly more negative than expected exemplars [$F(1,17) = 32.75$, $p < .001$]; this effect showed a significant interaction with Electrode [$F(25,45) = 6.50$, $p = .004$], indicating that the broad positivity to the expected exemplars has a different scalp distribution than the N400 response to the within-category items.

[8] Analyses were done in a 50-ms window around 375 ms (the peak of the N400 effect). The beginning of the effect could be observed from about 250 ms post-stimulus-onset and ended about 250 ms later. We chose to analyze a subset of that time interval to minimize the effects of overlapping components and because some of the effects (such as the constraint effects) were more temporally specific. The basic ending type effect, however, was statistically significant even when analyzed over large time windows (e.g., 250–500 ms: [$F(2,34) = 35.16$; $p < .001$]).

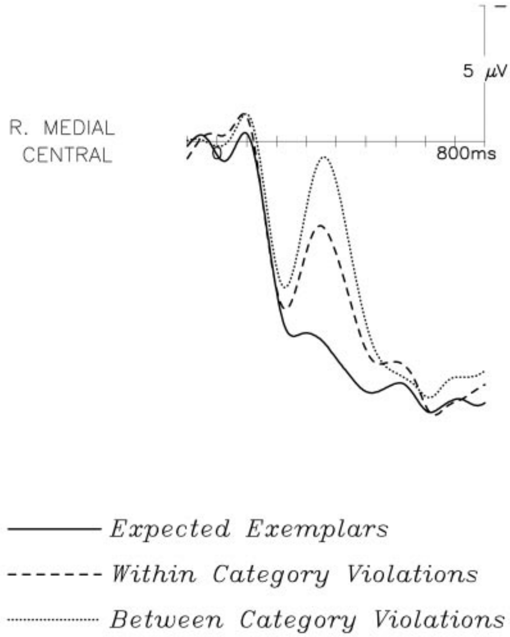

FIG. 3. Effect of ending type, shown at the right medial central site. A three-way split can be observed in the amplitude of the N400 response. N400 amplitude was significantly larger for between-category violations (dotted line) than for within-category violations (dashed line) and significantly larger for within-category violations than for expected exemplars (solid line).

Distribution of the N400 Effect

The distribution of the N400 effect for each violation type was examined by looking at the mean amplitude ERP difference between the violation type and the expected exemplar in the 350- to 400-ms time window. The difference waves (within-category violation ERP minus expected exemplar ERP and between-category violation ERP minus expected exemplar ERP) were normalized according to the procedure described in McCarthy and Wood (1985) and were then subjected to an ANOVA on four repeated measures, two levels of Ending Type Difference, two levels of Hemisphere (left vs right), two levels of Laterality (lateral vs medial), and four levels of Anterior/Posterior (prefrontal vs frontal vs parietal vs occipital).

A main effect of Laterality was observed [$F(1,17) = 34.57$; $p < .001$] as was a main effect of Anteriority [$F(3,51) = 5.08$; $p = .036$]; there was a nonsignificant trend toward a main effect of Hemisphere [$F(1,17) = 3.87$; $p = .066$]. In short, N400 effects were biggest over medial relative to lateral sites and over central/posterior relative to more anterior sites and tended to be bigger on the right than on the left. A Laterality by Anteriority interaction [$F(3,51) = 11.13$; $p < .001$] indicates that N400 effects at lateral sites are biggest over the occiput while N400 effects at medial sites are biggest parietally. A trend toward a Hemisphere by Laterality effect [$F(1,17) = 3.61$; $p = .075$] suggests that the difference between medial and lateral electrode sites is greater over the left scalp than over the right. N400 effects thus tended to be largest over medial, parietal sites and bigger over the right than over the left hemisphere. This distribution can be seen in the ERPs in Fig. 4; it is the one that is most typically reported for N400 effects during word by word sentential reading (Kutas & Van Petten, 1994).

A main effect of Ending Type was again observed [$F(1,17) = 4.52$; $p = .048$]. Ending Type did not interact with any of the distributional factors. Thus, the N400 response to within-category violations and the N400 response to between-category violations relative to expected exemplars are very similar in distribution.

Mean Amplitude Analyses of Constraint

Effects of contextual constraint could be observed in the grand average ERP waveforms. These effects were most prominent over medial, central-parietal sites where the N400 effect was biggest. Therefore, constraint was analyzed in the same time window at the four medio-central electrode sites (MiCe, LMCe, RMCe, MiPa). Mean voltage measures were subjected to an omnibus ANOVA with repeated measures on two levels of Constraint (high vs low), three levels of Ending Type (expected exemplars vs within-category violations vs between-category violations), and four levels of Electrode.

Main effects of both Ending Type [$F(2,34) = 38.43$; $p < .001$] and Constraint [$F(1,17) = 6.47$; $p = .021$] were observed. High constraint sentences are associated with overall higher

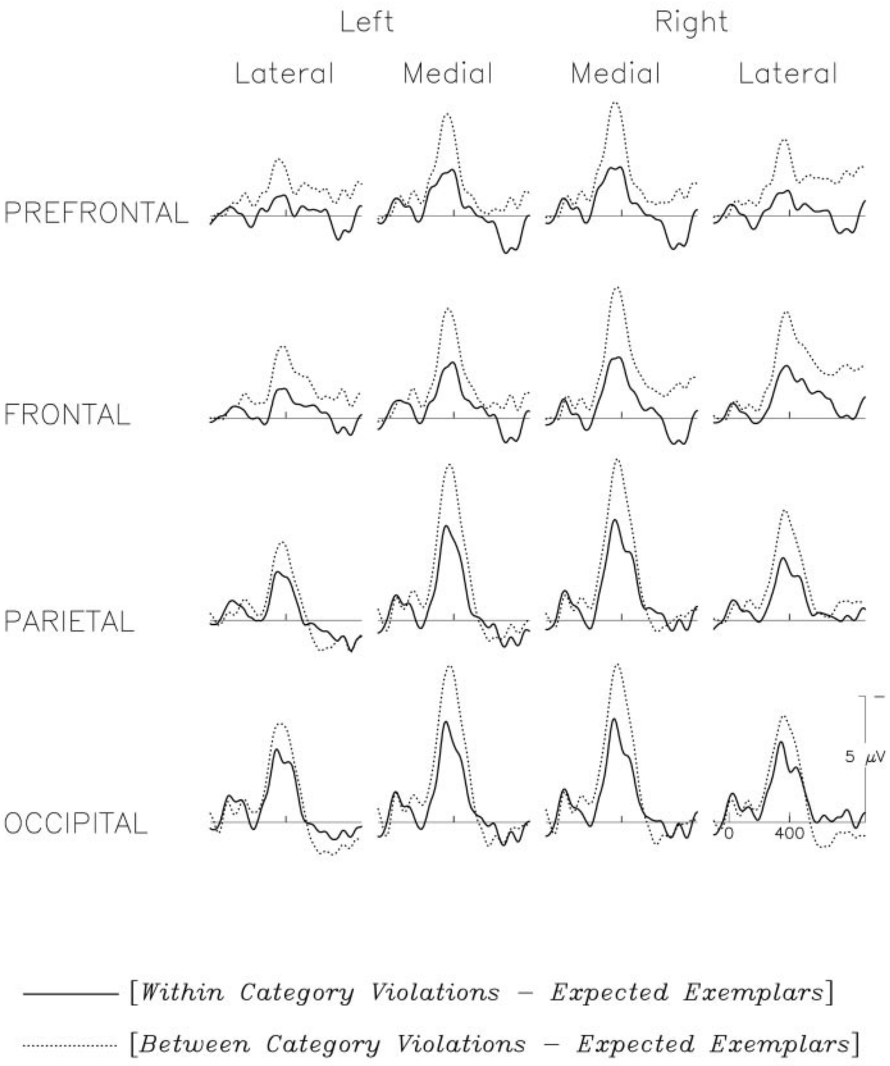

FIG. 4. Difference waves showing the N400 effect to within-category violations (solid line) and between-category violations (dotted lines). The waveforms at the 16 electrode sites (LLPf, LLFr, LLTe, LLOc, LMPf, LDFr, LDPa, LMOc, RMPf, RDFr, RDPa, RMOc, RLPf, RLFr, RLTe, RLOc) illustrate the distribution of the N400 effect. For both conditions, the N400 effect was larger over medial posterior sites and slightly larger on the right than on the left.

(more positive) mean amplitudes than are low constraint sentences. In addition, there was a Constraint by Ending Type interaction [$F(2,34) = 3.45$; $p = .043$], as can be seen in Fig. 5. Examination of the means (high constraint: 8.43, 5.10, and 1.23 μV for expected exemplars, within-category violations, and between-category violations, respectively; low constraint: 7.72, 2.38, and 0.85 μV) revealed that, while mean amplitudes are slightly larger for high than low constraint sentences for all ending types, most of the amplitude difference is accounted for by the difference between within-category violations in high versus low constraint sentences as shown in Fig. 6. Planned comparisons were performed between high and low constraint sentences for each ending type using *t*-tests ($df = 17$; $\alpha = 0.05$). These con-

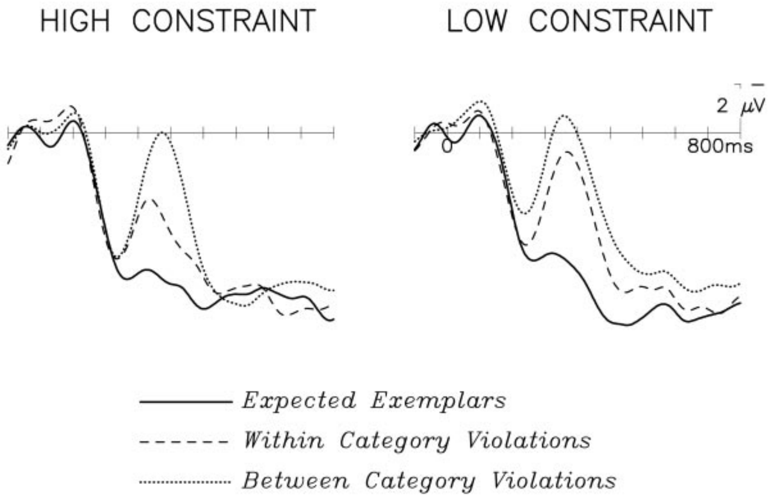

FIG. 5. Effect of constraint on the N400 response, shown at the right medial central site. Constraint did not affect the response to expected exemplars (solid line) or between-category violations (dotted line). Within-category violations (dashed line) in high constraint sentences (left) elicited smaller amplitude N400s than within-category violations in low constraint sentences (right).

firmed that the difference in the N400 to expected exemplars in high versus low constraint sentences ($t = 0.91$; $p = .373$) and to between-category violations in high versus low constraint sentences ($t = 0.50$; $p = .625$) are both minimal. By contrast, the N400 to within-category violations is significantly smaller when these are in high constraint rather than low constraint sentences ($t = 3.91$; $p = .001$).

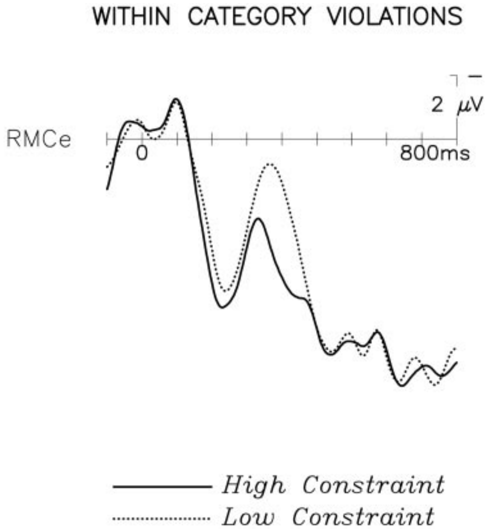

FIG. 6. Effect of constraint on within-category violations, shown at the right medial central site. Larger N400s were elicited by within-category violations in low constraint sentences (dotted line) than in high constraint sentences (solid line).

Stability of Effects over Items

Because of the low signal to noise ratios in electrophysiological data, it is usually not possible to perform item analyses, as would generally be done for behavioral data. However, it is possible to provide some indications that the observed effects are reasonably stable over items. The list and constraint analyses above provide some evidence for stability, as significant and qualitatively similar ending type effects were observed for all three lists and under both constraint conditions (consisting of different contexts and different items). To provide additional evidence, we made another split of the data for each participant into bins consisting of the first 22 items of each ending type condition (expected exemplar, within-category violation, between-category violation) and the second 22 items of each condition. Mean voltage measures were then taken for each bin in a 50-ms window around 375 ms (i.e., 350–400 ms post-stimulus-onset). These measures were subjected to an omnibus ANOVA with repeated measures on two levels of Experimental Half (first half vs second half), three levels of Ending Type (expected exemplar vs within-category violation vs between-category violation), and 26 levels of Electrode.

We again observed a main effect of Ending Type [$F(2,34) = 33.65$; $p < .001$] and an Ending Type × Electrode interaction [$F(50,850) = 7.61$; $p < .001$]. However, neither the main effect of Experimental Half [$F(1,17) = 2.82$; $p = .111$] nor its interaction with Ending Type [$F(2,34) = 0.48$; $p = .625$] was significant. Thus, there was no significant difference between the ending type effect observed for the (random) selection of items that consistently appeared in the first half of each list as compared with that observed for the items that appeared in the second half. This, combined with the findings from the list and constraint analyses, suggests that our effects are reasonably stable over items and suggests in addition that the effects are stable over time/practice.

Summary of Main Results

While expected exemplars elicit a late positivity in the 350- to 400-ms time window, both within-category violations and between-category violations elicit a qualitatively similar N400 response with a medial, posterior-central, right hemisphere distribution. The amplitude of the N400 is bigger for between- than for within-category violations and is bigger for within-category violations in low than in high constraint sentences.

DISCUSSION

At a general level, this experiment sought to determine the extent to which the processing of the final word in a sentence is affected not only by specific information directly activated by prior words in the sentence (i.e., context) but also by more general, context-independent information (semantic feature overlap) indirectly deriving from the structure of real-world knowledge in long-term memory. We addressed these issues by examining (1) whether the on-line processing of two items sharing significant numbers of semantic features in common (i.e., both members of the same category) differed when the preceding context was more consistent with one than the other, and (2) whether the on-line processing of two items, neither of which is especially consistent with the context (i.e., contextually expected), nonetheless differed as a function of their semantic similarity (categorical relationship) to the most probable or expected ending (i.e., best completion). We also examined whether or not the impact of either of these variables would be modulated by the degree to which the context anticipated a particular exemplar (high contextual constraint) versus several possibilities (low contextual constraint).

Our results show that fairly specific information from the sentential context is available by at least 375 ms after a word's presentation to affect its processing. Specifically, we observed that the brain's response to two category members (items sharing many semantic features in common) does differ inasmuch as one of them is consistent with the context (the expected exemplar) and the other is not (the within-category violation). The expected exemplar elicited a late positivity whereas the within-category violation elicited a moderate N400 between 300 and 600 ms. If processing the context serves to increase the availability of only fairly general feature information, then two members of a category which share these features should elicit a very similar brain response. However, this was not what we observed, indicating instead that context serves to increase the availability of specific feature information; that is, as we will argue, context leads to specific predictions or expectations.

While contextual information specific enough to distinguish between two semantically similar items affects word processing, it is also the case that the language processing system is sensitive to the categorical relationship between them. Specifically, there is a significant difference in the brain's response to two words, both of which are inconsistent with the context and thus unexpected, but one of which shares many semantic features with the expected ending and one of which does not. The N400 to an unexpected word that is a member of the same category as the expected ending (within-category violation) is significantly smaller than the N400 to a word that is similarly unexpected but shares fewer semantic features with the expected ending (between-category violation). This pattern of results is in accord with reports of behavioral facilitation for items semantically associated with the most expected endings or so-called best completions (e.g., Kleiman, 1980; Schwanenflugel & LaCount, 1988). However, since context and semantic similarity both modulate the same ERP component, the N400, the electrophysiological data allow two additional inferences to be drawn. First, this finding reveals a significant temporal overlap in the influence of both these factors on a word's processing. This has been previously shown, although not with as much care for the stimulus materials (see also Kutas & Hillyard, 1984; Kutas et al., 1984). Second, we can infer that the influences of context and semantic feature overlap on a word's processing are of the same kind. Insofar as the effects of these variables differ, it seems to be a matter of degree.

What is the basis for the processing benefit gained by items that are related to the expected exemplars, but which themselves are not expected? A straightforward explanation for this would have emerged if it were the case that only category-level (as opposed to more specific) contextual information was available to affect a word's processing in the N400 time window. In that case, we would expect an equivalent processing benefit to accrue to members of a category regardless of their specific fit to the context; in this experiment, we would expect best completions and within category violations to elicit similar ERPs. In fact, as in this study, both reaction time and ERP studies have previously observed that unexpected, related items generally yield a response intermediate in value between a best completion and an unexpected, unrelated ending. In other words, best completions have generally enjoyed greater processing benefits than semantically related but contextually unexpected endings, although both typically show facilitated processing relative to contextually unexpected and semantically unrelated endings. It is important to note, however, that in these previous studies "related" items were actually related in several different ways. Some were related because they shared feature information (i.e., were in the same taxonomic category), as in the present study, but others were only associatively or thematically related (e.g., "umbrella" and "rain") and thus shared few, if any, semantic features in common (e.g., Kutas & Hillyard, 1984; Kutas et al., 1984). Under the circumstances, the intermediate response observed for semantically related endings could have been a spurious artifact of averaging together a subset of related items that shared general features with the context (and therefore were facilitated) and another subset that did not share any semantic features (and therefore were not facilitated). In our experiment, however, as the nature of the relation between the within category violations and the expected exemplars was controlled such that they always shared many semantic features, this particular account does not apply. Our results thus not only demonstrate that specific contextual information is available to distinguish between such semantically similar items in real time but also call for an alternative, viable explanation for the intermediate response to unexpected but related items.

We believe that the smaller N400 to within-category violations compared to between-category violations reflects an influence of semantic memory structure, built of real-world experience, on on-line language processing. Although these two ending types are both equally incompatible with the specific information in the context that leads to the prediction for the expected exemplar, one of them has substantial semantic feature overlap with the ending most predicted in the context while the other does not. We suggest that it is the processor's sensitivity to this featural overlap that affords within-category violations a processing benefit relative to the between-category violations. Moreover, we are impressed by the sentence processing system's sensitivity to this categorical relationship between the expected ending and the within-category item even though it is irrelevant to the processing of the particular context and thus to making sense of the sentence (e.g., the fact that pines and palms are categorically related as trees is in no way relevant for determining that palms are an appropriate adornment for a tropical resort).

Before making this case strongly, however, we need to consider possible alternative explanations for the N400 difference between these two violation types. As it happens, we can readily rule out the possibility that our results simply reflect lexical associative priming from a word in the sentence context to the expected exemplar and, by extension, to the within-category violation (via some form of mediated priming). Our sentence pairs were constructed so that the sentence context ending with the target was equally compatible with all three ending types (e.g., “He asked his friend if he could borrow the *magazine/book/pencil*”), so none of these target sentences contained any lexical associate to any ending type. The few lexical associates that did exist were limited to the first sentence of the pair and were thus separated from the target word by at least five words or more. This distance is too great to sustain typical lexical associative priming effects, which are known to dissipate with even a single intervening word (Gough, Alford, & Holley-Wilcox, 1981; Masson, 1991; Ratcliff & McKoon, 1988). Furthermore, only about one third of our context (first) sentences actually contained a word lexically associated with the expected ending. Thus, we consider it highly unlikely that the lack of an N400 to the expected endings reflects lexical associative priming. More importantly, our expected exemplars were not lexically associated with the within category violations overall (less than 10% contained any degree of association). Thus (mediated) lexical associative priming cannot be invoked to account for the smaller N400s to the within-category violations relative to the between-category violations.

Another, more likely, explanation for the observed N400 difference between our violation types might be in terms of plausibility. Perhaps within category violations elicited smaller N400s than between category violations because they were more plausible, i.e., actually did fit the context better. The relationship between N400 amplitude and plausibility is poorly understood. On the one hand, N400 amplitude has been found to vary with the predictability of a word in its context (Kutas & Hillyard, 1984), which can arise from discourse as well as sentence level factors (e.g., Van Berkum, Hagoort, & Brown, in press). On the other hand, it has long been known that N400 amplitude cannot serve as a pure index of global level plausibility; Fischler et al. (1983) showed in an early series of studies that the N400 is not sensitive to negation. As plausibility and expectancy (as measured by cloze probability) are undoubtedly linked, and as the N400’s sensitivity to cloze probability is well established, plausibility certainly must make some contribution to our N400 effects. The real question, however, is the extent to which plausibility alone can explain the observed reduction in N400 amplitude to the within-category violations. As will soon become evident, plausibility alone will not suffice.

If plausibility alone were driving the pattern of results observed, then at minimum N400 amplitudes should be monotonically related to rated plausibility. That is, N400 amplitude should decrease when plausibility increases and increase when plausibility decreases, whatever the exact relation in the rate of change of the two. At the most general level, we do observe a monotonic relation between plausibility and N400 amplitude: best completions elicit the smallest N400s and the highest plausibility ratings, between-category violations elicit the largest N400s and the lowest plausibility ratings, and within-category violations are intermediate on both variables. However, we find that this monotonic relation between N400 amplitude and plausibility does not hold when the data are broken down by contextual constraint. Although the rated plausibility is significantly higher for best completions in high versus low constraint sentence contexts, the associated ERPs do not differ. Likewise, although the rated plausibility is higher for between category violations in low than high constraint sentences, their N400s also do not differ. Finally, and most damning for the plausibility hypothesis, is our finding that while the rated plausibility for within category violations is also significantly higher under low than high contextual constraint, N400 amplitudes are significantly different in the opposite direction. That is, the more plausible within-category violations in the low constraint sentences are associated with a larger N400 than are the less plausible within-category violations in the high constraint sentences.

If we loosen the montonicity criterion, we can maintain the plausibility hypothesis despite the absence of a significant N400 amplitude effect in the presence of a significant plausibility effect. However, no explanation in terms of plausibility will be able to account for our finding that among the within-category violation endings, the more plausible endings elicit larger N400s than do the more implausible endings (e.g., "baseball," in "'Checkmate!' Rosaline announced with glee. She was getting to be really good at baseball." has a smaller N400 than "earring," in "She keeps twirling it around and around under her collar. Stephanie seems really happy that Dan gave her that earring."). This deviance from monotonicity forces us to reject the plausibility hypothesis as an explanation for the current pattern of results.

While lexical associative priming and plausibility clearly play significant roles in on-line language comprehension, even in combination they cannot account for the smaller N400s to within- than between-category violation endings. We now return to consider our original hypothesis that the explanation is inherent in the structure of information in long-term memory. The greater reduction in N400 amplitude to within category violations in more than less constraining contexts is counter to their rated plausibilities. Instead, it appears to pattern with the expectancy and plausibility ratings for the expected items: the very contexts that set up very specific expectations for the expected exemplar also provide the within-category violations with the greatest facilitation (i.e., the greatest N400 reduction). Thus, there seems to be a functional link between the expected endings and the within-category violations—a link which we argue reflects memory structure.

This functional link has implications for how information in a sentence context is used during language comprehension as well as for the organization of long-term memory and its impact on sentence processing in real time. First, this link suggests that in the course of processing a sentence, the comprehension system is involved in some process tantamount to prediction. By using the word "prediction," we do not necessarily mean to imply that the process is either conscious or strategic. Rather, prediction here refers to activation of the semantic features of upcoming words prior to their occurrence. In this specific case, we mean that semantic features of the category exemplar (not necessarily a specific lexical item) most likely to complete the target sentence are activated prior to the presentation of the actual sentence-final word. When the prediction is incorrect and the expectancy is not met, the data are characterized by increased N400 activity relative to when the prediction is correct. Note that such activation of the semantic features of the expected ending occurs above and beyond the activation of the semantic features of the context words themselves, although naturally it must be contingent on their presence.

Of course, some might argue against any form of prediction, opting instead for a matching process initiated by the final word. On this view, as a sentence is processed, the set of active features includes those of the current word and of the preceding context words, but none of the features of upcoming words. As each word is presented, the comprehension system presumably checks the degree of (mis) match between context features and the current word's features, and responds accordingly. The greater the mismatch, the larger the N400 elicited. This account would seem to predict that the more the information in a context constrains the possible exemplars, the greater the mismatch (and the associated N400) when that preferred exemplar does not occur. While this is a reasonable hypothesis, it fails to account for our finding smaller N400s to within-category violations in high than in low contextual constraint.

In contrast, this outcome easily falls out of our "prediction" account. The N400 to within-category violations is reduced because many of its features, namely those it shares with the expected (but not presented) exemplar, are already active prior to its appearance. Between-category violations share fewer features in common with expected exemplars and therefore cannot benefit from this type of prediction, as reflected in a larger N400. It is thus the semantic relationship between the within-category violation and the expected exemplar—the item that would be predicted in the context but that is never actually seen—that results in an N400 reduction (e.g., although pine trees themselves are not very compatible with tropical resorts, they share features in common with something that is, namely palm trees). Strikingly, this reduction increases when the contextual information allows a strong as opposed to a weak prediction of that expected item (and a correspondingly weaker prediction of the within-category violation, the item actually seen). Thus, as a result of processing the context, the semantic features of the expected exemplar must become activated, and it is the overlap between the within category violation and that prediction that determines the size of the observed N400 response (see McKoon & Ratcliff, 1989, for a similar conclusion in work on elaborative inferences).

Second, the functional link between unexpected but categorically related endings and expected exemplars not only supports the view held by many researchers that long-term memory is structured but also our specific proposal that this structure has an inherent effect on sentence processing in real time. Our finding that information is routinely retrieved from long-term memory during language comprehension is not news. How could it be otherwise? Likewise, we are not the first to suggest that semantic memory has a categorical structural component. The results of categorization research have long been used to infer that experience with the world structures long-term memory, so that items that share perceptual or functional traits come to be grouped together and treated as similar to one another, i.e., as a category. However, such evidence for the categorical structure of semantic memory has usually been obtained in some kind of categorization task. Participants in these studies are either explicitly asked to categorize or shown items grouped in a way that renders their categorical relationship quite apparent. It has therefore proven difficult to draw unequivocal conclusions from such studies. How can we be sure that the category-based effects arise because individuals tap into existing structured representations and not because individuals create structure as needed by the categorization task itself (for review, see Kounios, 1996)?

Our results are thus important for resolving this issue because they are not subject to these concerns. Our participants were not asked to perform any explicit categorization or comparison; their only task was to read the sentences for comprehension. Nevertheless, we observed a reliable category-based effect during the processing of the sentence final word. It is important to note that we observed a category-based effect on the N400, even though the categorical relationship was neither obvious nor relevant to the comprehension task at hand. Our participants did not even see the two categorically related items in close proximity, because one of them—the expected exemplar—was not even presented, but merely implied by the context. Thus, our results show that robust category-based effects can be obtained within a few hundred milliseconds of an event's occurrence even outside of a categorization task. Another, more novel, finding is that this category-based structure of long term memory seems to routinely influence language processing, even when it is irrelevant and perhaps detrimental to the comprehension process.

We observed that the N400 to within-category violations is reduced by virtue of its categorical relationship to the word that is reportedly most expected in the two-sentence context. Previously we argued that neither lexical priming nor plausibility could explain this particular result, nor could any view that assumes that long-term memory exists without any inherent structure or that there are no categories except those that are dynamically generated de novo as a function of context (e.g., Barsalou & Medin, 1986). Our data thus provide one of the first clear demonstrations that the experientially imposed structure of long-term memory has a significant and measurable impact on contextually driven language processes.

Yet another novel finding emerging from this study is that the influence of the categorical structure of memory on sentence processing is modulated by the degree to which context constrains the expected exemplar. Contrary to a plausibility or matching account, the influence of this structure actually increases in highly constraining contexts. This observation suggests that the semantic memory structure is not simply a factor that becomes relevant when other cues are absent, weak, or less available, but rather that its influence is an inherent consequence of the way the brain processes linguistic input[9]. Studies using lexical decision tasks have previously shown that weakly constraining contexts typically provide a wider scope of facilitation than do more constraining contexts (Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1985). While highly constraining contexts are quite specific in facilitating only the best completion, weaker contexts can apparently facilitate a congruent ending semantically related to the best completion as well. We found a similar pattern in the plausibility ratings we obtained off-line: participants were indeed much more willing to accept several different endings as plausible in weakly than in strongly constraining contexts. The ERP data, however, indicate a different pattern of context effects earlier in the processing stream; around 300 ms, the apparent facilitation of irrelevant, semantically related items is actually greater in stronger as opposed to weaker contexts. On the one hand, high constraint sentences seem to provide more specific contextual information. On the other hand, the effect of memory structure on processing, as indexed by the N400, is greater in these contexts than in weaker ones. We take this to mean that information about semantic feature overlap is automatically used in language processing, in proportion to the system's ability to predict the semantic features of items that will come next.

What does it mean to say that long-term memory structure (as captured by semantic similarity) has an inherent effect on language processing? By this we do not mean to imply that the brain contains discrete categories or bins corresponding to “fruits” or “trees,” and so on. As detailed in the introduction, there is ample evidence suggesting that categories are graded, and overlapping, that category membership cannot be strictly similarity-based, and that what constitutes a category will vary as a function of both context and the level of abstraction at which categorization is performed (Barsalou & Medin, 1986; see also review by Komatsu 1992; Lassaline, Wisniewski, & Medin, 1992; e.g., Rosch, 1975; Rosch & Mervis, 1975; Rosch et al., 1976; Roth & Shoben, 1983). What we do mean, however, is that the kind of perceptual and functional similarity captured by semantic categories like those we used as sentence endings, but also possibly other types, has an impact on neural and linguistic processing. The precise role similarity plays in categorization is unknown and there remains much debate over how best to define similarity or to calculate it (e.g., Goldstone, 1995; Medin, Goldstone, & Gentner, 1993; Murphy & Spalding, 1995). However, we do know that the brain is sensitive to various input features and that the representations of this feature information are often structured, for example, into cortical maps wherein cells responding to similar features are physically close to one another, clustered in columnar regions (e.g., Brugge & Merzenich, 1973; Hubel & Wiesel, 1972; Tanaka, 1996).

We can thus view the neural representation of the objects that words refer to as a set of features[10]. Whenever two words refer to two things that look alike or sound alike or invoke similar motor programs for interaction, then we suppose that there is also a similarity in how they

[9] By this we do not mean to imply that it is language specific.

[10] These features need not be simple; they can include higher order relations. There is also no necessity to assume a context-independent or one-to-one mapping between a word and the set of features or neurons that represents it. The only critical assumption we make here is that at any given moment some concept is represented by a set of features and that a closely related concept sharing many of those features and activated under similar conditions would involve activity in a partially overlapping set of neurons.

are represented in the brain, i.e., in the neural activity they elicit. A neural system structured in this way would likely find it easier to transition between the pattern of activation corresponding to one thing and that corresponding to a different but related thing. More specifically, if the comprehension system predicts the features of "palms" in the way we proposed earlier, then a structured representation based on feature overlap of the sort we just described would leave it better prepared to activate the features of "pines" than of "tulips." Moreover, our ERP data suggest that the more strongly the features of "palms" are activated, the better prepared the system is to deal with the "pines," even if these features are a poor fit to the activated context features.

In conclusion, our data suggest that the language comprehension system is sensitive to specific contextual information and to the consistency between that specific information and the meaning of a target word by around 375 ms into word processing. This information is specific enough to distinguish two words whose referents share many semantic features in common. However, the fit between a given word and specific contextual information alone does not completely predict the brain's response. In particular, in the same time window as we first observe the influence of contextual information on word processing, we also observe an influence of semantic feature overlap (as reflected in taxonomic semantic categories) that is independent of the fit of that word to the specific sentence context. That is, we observe an inherent influence of long-term memory structure on language processing, at least that aspect indexed by the N400. This suggests that the processing of a sentence context results in the activation of a set of semantic features associated with the word or words that are likely to come next. When a word is actually encountered, it is the degree of semantic feature match or mismatch between it and the prediction derived from context that determines the difficulty of processing, at least initially. Stronger contexts allow better predictions and greater facilitation for items that share features with the predicted word. Thus, context and long-term memory structure have a dynamic, mutually dependent relationship with one another and contribute jointly to the processes involved in making sense of what we read.

APPENDIX A

Categories Used in the Experiment

Sixty-six different categories were used in the experiment. These categories were paired such that for two-sentence pairs the expected and within-category target items were derived from one category of the pair while the between-category item was derived from the other; these roles reversed for a second set of two-sentence pairs. The categories used in the experiment and their pairings were as follows.

Biological categories
Plants
  trees, flowers
Animals
  crustaceans, fish
  marine mammals, marsupials
  dogs, equines
  insects, rodents
  birds, reptiles
  (wild)cats, bears
  dinosaurs, mythical beings
Human-related, human-like items
  body parts (external), internal organs
  superheroes, cartoon animals
Nonbiological categories
Foods
  fruits, vegetables
  meats, cheeses
  desserts, breads
  alcoholic, non-alcoholic beverages
Places and buildings
  land formations, celestial bodies
  countries, states
  native dwellings, religious buildings
Vehicles
  cars, public transportation
  aircraft, boats
  war vehicles, heavy machinery
Tools and household objects
  carpentry tools, gardening tools
  measuring instruments, optical instruments
  medical supplies, office supplies

dishes, utensils
small (kitchen) appliances, lighting
Garments and personal articles
tops, toiletries
pants, shoes
safety-wear, walking aides
jewelry, make-up
Leisure- and hobby-related items
sports, board games
sports equipment, toys
reading material, writing instruments
Miscellaneous
containers, fasteners

APPENDIX B

Examples of Stimuli Used in the Experiment

One hundred thirty-two sentence contexts were used in the experiment, each ending with one of each of the three possible ending types (expected exemplars, within-category violations, between-category violations). Below are given 40 representative examples of these stimuli. Ending types are expected exemplar, within-category violation, and between-category violation, respectively. High constraint sentences are marked with an "H" and low constraint sentences are marked with an "L."

(H) "Checkmate," Rosaline announced with glee.
She was getting to be really good at chess/monopoly/football.

(H) Justin put a second house on Park Place.
He and his sister often spent hours playing monopoly/chess/baseball.

(H) He caught the pass and scored another touchdown.
There was nothing he enjoyed more than a good game of football/baseball/monopoly.

(H) Rich couldn't count the number of Yankees games he had seen with his father.
They both shared a lifelong interest in baseball/football/chess.

(L) She felt that she couldn't leave Venice without the experience.
It might be a touristy thing to do, but she wanted to ride in a gondola/ferry/helicopter.

(L) Getting both himself and his car to work on the neighboring island was time-consuming.
Every morning he drove for a few minutes and then boarded the ferry/gondola/plane.

(L) The patient was in critical condition and the ambulance wouldn't be fast enough.
They decided they would have to use the helicopter/plane/ferry.

(L) Amy was very anxious about traveling abroad for the first time.
She felt surprisingly better, however, when she actually boarded the plane/helicopter/gondola.

(L) The day before the wedding, the kitchen was just covered with frosting.
Annette's sister was responsible for making the cake/cookies/toast.

(H) The little girl was happy that Santa Claus left nothing but crumbs on the plate.
She decided he must have really enjoyed the cookies/cake/bagel.

(H) Chris moped around all morning when he discovered there was no cream cheese.
He complained that without it he couldn't eat his bagel/toast/cake.

(H) He wanted to make his wife breakfast, but he burned piece after piece.
I couldn't believe he was ruining even the toast/bagel/cookies.

(H) I guess his girlfriend really encouraged him to get it pierced.
But his father sure blew up when he came home wearing that earring/necklace/lipstick.

(L) She keeps twirling it around and around under her collar.
Stephanie seems really happy that Dan gave her that necklace/earring/mascara.

(H) She wanted to make her eyelashes look really black and thick.
So she asked to borrow her older friend's mascara/lipstick/necklace.

(H) He complained that after she kissed him, he couldn't get the red color off his face.
He finally just asked her to stop wearing that lipstick/mascara/earring.

(L) Eleanor offered to fix her visitor some coffee.

Then she realized she didn't have a clean cup/bowl/spoon.

(L) My aunt fixed my brother some cereal using her best china.

Of course, the first thing he did was drop the bowl/cup/knife.

(H) At the dinner party, I wondered why my mother wasn't eating her soup.

Then I noticed that she didn't have a spoon/knife/bowl.

(L) In the dorms, cutting your steak can be a huge struggle.

They always give you such a poor quality knife/spoon/cup.

(H) He journeyed to the African plains, hoping to get a photograph of the king of the beasts.

Unfortunately, the whole time he was there he never saw a lion/tiger/panda.

(L) George was hiking in India when he saw the orange and black striped animal leap out at him.

He sustained serious injuries before he managed to kill the tiger/lion/polar bear.

(H) Hitting the huge animal with a tranquilizer dart was difficult in the Arctic winds.

Eventually, however, they were able to approach and tag the polar bear/panda/lion.

(L) Wendy wondered how they had managed to ship such a large animal all the way from China.

She waited in line to see the newly acquired panda/polar bear/tiger.

(H) Barb loved the feel of the waves on her feet, but she hated to walk barefoot.

As a compromise, she usually wore a pair of sandals/boots/shorts.

(L) By the end of the day, the hiker's feet were extremely cold and wet.

It was the last time he would ever buy a cheap pair of boots/sandals/jeans.

(H) Everyone agreed that the stone-washed kind were out of style.

But he continued to wear the same old pair of jeans/shorts/sandals.

(L) As the afternoon progressed, it became hotter and hotter.

Keith finally decided to put on a pair of shorts/jeans/boots.

(L) Pablo wanted to cut the lumber he had bought to make some shelves.

He asked his neighbor if he could borrow her saw/hammer/rake.

(H) Tina lined up where she thought the nail should go.

When she was satisfied, she asked Bruce to hand her the hammer/saw/shovel.

(H) The snow had piled up on the drive so high that they couldn't get the car out.

When Albert woke up, his father handed him a shovel/rake/saw.

(H) The yard was completely covered with a thick layer of dead leaves.

Erica decided it was time to get out the rake/shovel/hammer.

(L) Fred went to the pantry and got out the homemade jelly his grandmother had brought.

Fifteen minutes later, however, he was still struggling to open the jar/box/zipper.

(L) After they unpacked the new refrigerator, they let Billy have his fun.

He played for days afterwards with the big box/jar/button.

(H) It seemed to catch every time she opened or closed her backpack.

She decided she would have to replace the zipper/button/box.

(H) One fell off her blouse and got lost, and she didn't have any extras.

She ended up searching all over town to find a matching button/zipper/jar.

(L) The firefighters wanted to have a mascot to live with them at the firehouse.

Naturally, they decided it would have to be a dalmatian/poodle/zebra.

(L) Muffie, old Mrs. Smith's pet, wears a bow on the puff of fur on its head.

I don't know how anyone could want to own a poodle/dalmatian/donkey.

(L) "I'm an animal like Eeyore!" the child exclaimed.

His mother wondered why he was pretending to be a donkey/zebra/dalmatian.

(H) At the zoo, my sister asked if they painted the black and white stripes on the animal.

I explained to her that they were natural features of a zebra/donkey/poodle.

REFERENCES

Barsalou, L. W., & Medin, D. L. (1986). Concepts: Static definitions or context-dependent representations? Special Issue: Context and cognition. *Cahiers de Psychologie,* **6,** 187–202.

Besson, M., & Macar, F. (1987). An event-related potential analysis of incongruity in music and other non-linguistic contexts. *Psychophysiology,* **24,** 14–25.

Brugge, J. F., & Merzenich, M. M. (1973). Responses of neurons in auditory cortex of the macaque monkey to monaural and binaural stimulation. *Journal of Neurophysiology,* **36,** 1138–1158.

Chao, L. L., Nielsen-Bohlman, L., & Knight, R. T. (1995). Auditory event-related potentials dissociate early and late memory processes. *Electroencephalography and Clinical Neurophysiology,* **96,** 157–168.

Dale, A. M. (1994). *Source localization and spatial discriminant analysis of event-related potentials: linear approaches.* La Jolla, CA: University of California San Diego.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **15,** 791–801.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior,* **20,** 641–655.

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior,* **18,** 1–20.

Fischler, I., Bloom, P. A., Childers, D. G., Arroyo, A. A., & Perry, N. W. J. (1984). Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia,* **22,** 559–568.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology,* **20,** 400–409.

Fischler, I., Childers, D. G., Achariyapaopan, T., & Perry, N. W. (1985). Brain potentials during sentence verification: Automatic aspects of comprehension. *Biological Psychology,* **21,** 83–105.

Fischler, I. S., & Bloom, P. A. (1985). Effects of constraint and validity of sentence contexts on lexical decisions. *Memory & Cognition,* **13,** 128–139.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Friedman, D. (1990). Cognitive event-related potential components during continuous recognition memory for pictures. *Psychophysiology,* **27,** 136–148.

Goldstone, R. L. (1995). Mainstream and avant-garde similarity. Special Issue: Similarity and categorization. *Psychologica Belgica,* **35,** 145–165.

Gough, P. B., Alford, J. A., Jr., & Holley-Wilcox, P. (1981). Words and contexts. In O. J. L. Tzeng & H. Singer (Eds.), *Perception of print: Reading research in experimental psychology* (pp. 85–102). Hillsdale, NJ: Erlbaum.

Grunwald, T., Elger, C. E., Lehnertz, K., Van Roost, D., & Heinze, H. J. (1995). Alterations of intrahippocampal cognitive potentials in temporal lobe epilepsy. *Electroencephalography and Clinical Neurophysiology,* **95,** 53–62.

Harbin, T. J., Marsh, G. R., & Harvey, M. T. (1984). Differences in the late components of the event-related potential due to age and to semantic and non-semantic tasks. *Electroencephalography & Clinical Neurophysiology: Evoked Potentials,* **59,** 489–496.

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General,* **124,** 62–82.

Hubel, D. H., & Wiesel, T. N. (1972). Laminar and columnar distribution of geniculo-cortical fibers in the macaque monkey. *Journal of Comparative Neurology,* **146,** 421–450.

Kay, P. (1971). Taxonomy and semantic contrast. *Language,* **47,** 866–887.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associated thesaurus of English and it computer analysis. In A. J. Aitken, R. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies.* Edinburgh: Edinburgh University Press.

Kleiman, G. M. (1980). Sentence frame contexts and lexical decisions: Sentence-acceptability and word-relatedness effects. *Memory & Cognition,* **8,** 336–344.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin,* **112,** 500–526.

Kounios, J. (1996). On the continuity of thought and the representation of knowledge: Electrophysiological and behavioral time-course measures reveal levels of structure in semantic memory. *Psychonomic Bulletin & Review,* **3,** 265–286.

Kounios, J., & Holcomb, P. J. (1992). Structure and process in semantic memory: Evidence from event-related brain potentials and reaction times. *Journal of Experimental Psychology: General,* **121,** 459–479.

Kutas, M., & Dale, A. (1997). Electrical and magnetic readings of mental functions. In M. D. Rugg (Ed.), *Cognitive Neuroscience* (pp. 197–242). Hove, East Sussex: Psychology Press.

Kutas, M., & Hillyard, S. A. (1980a). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology,* **11,** 99–116.

Kutas, M., & Hillyard, S. A. (1980b). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science,* **207,** 203–205.

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition,* **11,** 539–550.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature,* **307,** 161–163.

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory States and Processes* (pp. 217–237). Hilldale, NJ: Erlbaum.

Kutas, M., & Van Petten, C. K. (1994). Psycholinguistics electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 83–143). San Diego: Academic Press.

Lassaline, M. E., Wisniewski, E. J., & Medin, D. L. (1992). Basic levels in artificial and natural categories: Are all basic levels created equal? In B. Barbara (Ed.), Percepts, concepts and categories: *The representation and processing of information.* (Vol. 93, pp. 327–378). Amsterdam: North-Holland.

Masson, M. E. J. (1991). A distributed memory model of context effects in word identification. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 233–263). Hillsdale, NJ: Erlbaum.

McCarthy, G., Nobre, A. C., Bentin, S., & Spencer, D. D. (1995). Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. *Journal of Neuroscience,* **15,** 1080–1089.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology,* **62,** 203–208.

McClelland, J. L., & O'Regan, J. K. (1981). Expectations increase the benefit derived from parafoveal visual information in reading words aloud. *Journal of Experimental Psychology: Human Perception & Performance,* **7,** 634–644.

McKoon, G., & Ratcliff, R. (1989). Semantic associations and elaborative inference. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **15,** 326–338.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review,* **100,** 254–278.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **20,** 92–103.

Murphy, G. L., & Spalding, T. L. (1995). Knowledge, similarity, and concept formation. Special Issue: Similarity and categorization. *Psychologica Belgica,* **35,** 127–144.

Neville, H. J., Kutas, M., Chesney, G., & Schmidt, A. L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language,* **25,** 75–92.

Nobre, A. C., & McCarthy, G. (1995). Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. *Journal of Neuroscience,* **15,** 1090–1098.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia,* **9,** 97–113.

Polich, J. (1985). N400s from sentences, semantic categories, number and letter strings? *Bulletin of the Psychonomic Society,* **23,** 361–364.

Ratcliff, J. E. (1987). The plausibility effect: Lexical priming or sentential processing? *Memory & Cognition,* **15,** 482–496.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review,* **95,** 385–408.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General,* **104,** 192–233.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology,* **7,** 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology,* **8,** 382–439.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (Vol. xii, p. 308). New York: Academic Press.

Roth, E. M., & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology,* **15,** 346–378.

Rugg, M. D., & Coles, M. G. H. (Eds.). (1995). *Electrophysiology of mind: Event-related brain potentials and cognition*. (Vol. 25). Oxford: Oxford University Press.

Schuberth, R. E., Spoehr, K. T., & Lane, D. M. (1981). Effects of stimulus and contextual information on the lexical decision process. *Memory & Cognition,* **9,** 68–77.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **14,** 344–354.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language,* **24,** 232–252.

Schwantes, F. M. (1985). Expectancy, integration, and interactional processes: Age differences in the nature of words affected by sentence context. *Journal of Experimental Child Psychology,* **39,** 212–229.

Smith, M. E., Stapleton, J. M., & Halgren, E. (1986). Human medial temporal lobe potentials evoked in memory and language tasks. *Electroencephalography and Clinical Neurophysiology,* **63,** 145–159.

Squire, L. R. (1987). *Memory and brain.* New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General,* **112,** 1–36.

Stuss, D. T., Picton, T. W., & Cerri, A. M. (1986). Searching for the names of pictures: An event-related potential study. *Psychophysiology,* **23,** 215–223.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience,* **19,** 109–139.

Taylor, W. L. (1953). "Cloze procedure": a new tool for measuring readability. *Journalism Quarterly,* **30,** 415–433.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (in press). Semantic integration in sentences and discourse. *Journal of Cognitive Neuroscience.*

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition,* **18,** 380–393.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics,* **36,** 277–284.

(Received March 5, 1998)
(Revision received April 13, 1999)