

Processing Unfamiliar Metaphors in a Self-Paced Reading Task

Frank Brisard, Steven Frisson, and Dominiek Sandra

*Department of Germanic Languages
University of Antwerp*

In 2 self-paced reading experiments, we investigate the processing characteristics of unfamiliar metaphorical subject–predicate structures. The literal first hypothesis predicts that processing metaphorical expressions of the type “*an x is a y*” will proceed more slowly than in the case of literal statements of the same type. This prediction is confirmed: At the position of the metaphorical term, reaction times were indeed higher for the metaphorical conditions than for the literal ones. This result was obtained both without (Experiment 1) and with a supportive context sentence (Experiment 2). In Experiment 2, a distinction also emerges between apt and nonapt instances, such that reaction times for apt metaphors are no longer significantly higher toward the end of the clause containing them. This suggests that, when embedded in a rich context, the interpretation of unfamiliar apt metaphors can be completed by the end of a fragment that can serve as a clause.

The model that has probably had the strongest effect on the literature concerning the time course involved in processing metaphorical language starts from the so-called *literal first hypothesis*, which observes that the interpretation of metaphors needs to pass through a stage in which the literal meaning of an utterance is processed before its figurative meaning can be computed. The hypothesis is derived from a *stage model* of metaphor comprehension that originated in contemporary philosophy of language. In their semantic theories, scholars like Searle (1979) and Grice (1975) distinguished between sentence meaning and utterance or speaker’s meaning, reflecting the distinction between what is said through an utterance (i.e., the conventional, literal meanings of words and how they are syntactically combined) and the ulterior meaning the speaker wishes to express, which can only be

implicated in the case of metaphor. According to this model, language users encountering metaphorical statements first determine the sentence meaning, then discover that this meaning cannot be what the speaker intended (because it is typically false, literally), only to reject the literal meaning afterward in favor of a derived, contextually computed figurative meaning. Strictly speaking, this model also implies that the search for a figurative interpretation cannot be initiated if a literal interpretation is successfully integrated in the sentence context. That is, any type of literal (sentence) meaning that can be ascribed to an utterance and that is in some way compatible with the interpretive context in which the utterance appears will automatically block the process of finding an alternative reading. Thus, the position of literal meaning in this model is clearly an absolute one and always prevails on nonliteral derivations, whether of a metaphorical nature or otherwise related to the computation of the speaker's meaning.

In the original formulations of the literal first hypothesis, a figurative interpretation can only be computed at the end of a sentence. In what follows, however, we adopt a version of the model that is more in line with current theories of incremental processing. To verify the claims of literal first as a stage model, the hypothesis is put forward that, if a metaphorical meaning is derived from a previously determined literal one, metaphors must take longer to process than (matched) literal propositions. To test this, it is necessary to tap the processing of metaphors online (i.e., during the word-for-word presentation of the metaphorical stimulus sentence). If reaction times (RTs) are measured for complete metaphorical sentences only, other components of metaphorical interpretation, like the actual appreciation of the metaphor in question (Gibbs, 1992), will have already had the chance to exert an influence on the course of processing. A genuine online measuring technique, then, can be implemented effectively with experimental stimuli that are limited to the fairly simple structure of categorization statements, like "*An X is a Y*," because the position of the literal and metaphorical term *Y* remains constant in such expressions (in contrast with referring metaphors, where this type of positional variation is much harder to control for). Now, instances of metaphorical language, especially when they are of the predicative type distinguished here, invite language users to make classifications that do not fit any literal taxonomy. Thus, metaphor is a device that enables the language user to redeploy a category scheme that characterizes one domain to effect a reorganization of another. If somebody says "*Friends are trees*," he or she is asking us to consider that some items are not only people but also trees, a taxonomic error, unless of course only the relevant similarities are sorted out. Theoretically, this sorting and the resulting nonliteral interpretation need not occur after a literal interpretation is attempted. But how could this be empirically demonstrated?

Many experiments have been carried out with exactly the type of predicative stimuli included in the present series. Fairly few of these studies, however, address the issue of unfamiliar metaphors, among them Blasko and Connine (1993) and

Gerrig (1989). Glucksberg, Gildea, and Bookin (1982) used predicative sentence structures to investigate whether the literal meaning of a metaphorical expression can be responded to before its metaphorical meaning is available. In the experiment, participants had to decide whether sentences were literally false or not; that is, they had to monitor for the literal meaning of the sentence only and react to that. RTs and error rates were compared between two categories of literally false items: metaphors and nonmetaphorical false statements. The authors demonstrated that it is more difficult for participants in a speeded response task to answer “no” to sentences like “*All jobs are jails*” (as opposed to blatantly false sentences without a possible metaphorical interpretation), with longer RTs and higher error rates. This shows that the availability of a true metaphorical meaning interferes with the execution of a negative response. However, as the authors themselves remarked, these findings cannot really reject the literal first hypothesis. The construction of the metaphorical meaning in a second stage may be so fast and automatic as to interfere with the processing of the literal meaning regardless of its secondary status. Alternatively, participants may simply be unable to monitor for early processing stages, as these may be part of the processing machinery of a modular system (which by definition cannot be penetrated by conscious attention processes). Hence, although Glucksberg and associates reported findings that seem to argue against a literal first model, the nature of the experimental tasks employed does not make the interpretation of their data compelling in this respect. The only thing these experiments show is that metaphor processing is highly automatic; it cannot be brought under the conscious control of participants (to facilitate task compliance), yet they remain neutral as to the involvement of metaphorical meanings in an initial stage of processing.

The interpretive ambiguity in Glucksberg et al. (1982) derives from the fact that they used an indirect method to compare literal and metaphorical processing by focusing on the processing of the literal meanings of metaphorical statements. Indeed, the indirectness of comparisons between literal and metaphorical conditions constitutes one of the more important difficulties in the interpretation of experimental results concerning the time course of metaphor comprehension. A second methodological problem in experimental studies of metaphor is exemplified in Gibbs’s repeated attempts to falsify literal first. Concretely, Gibbs questioned the psychological validity of the literal first hypothesis on the basis of experimental work on indirect requests, idioms, and sarcastic utterances (for overviews, see Gibbs, 1984, 1994). Gibbs also referred to experimental work of his own in which he showed that these kinds of expression (which can be subsumed, together with metaphors, under the general heading of figurative language) are processed as fast as literal sentences. However, the experimental paradigms reported by Gibbs may not be the best way to assess literal first, as global RTs measured for complete sentences, as the standard technique applied in these experiments, cannot tease apart immediate from additional processing.

In sum, if there is research indicating that literal and figurative meanings may be processed equally fast, the methodology that is generally used does not allow us to conclude that this is due to the use of an identical processing routine (a single stage for both types of language use) or, for that matter, to the parallel activation of two different routines (one for literal and one for figurative language). In particular, when experimental paradigms do not employ a genuine online method to measure RTs within sentences (word-for-word measurements), small effects may simply be undetectable. Consequently, when measuring global RTs (for complete sentences), the presence of an effect does not allow its exact localization (i.e., where it begins to emerge and how long it persists), and its absence (the null effect) may be due to the fact that the effect has been drowned in the sum of all individual word RTs. To make statements on the processing routine itself, one must therefore track the course of interpretation more meticulously, as also argued by Dascal (1989). This is what can be achieved by using a self-paced reading task. We ran two experiments in which this technique was applied.

Only unfamiliar metaphors are investigated, because they provide the most obvious point of entry for an investigation into the creative function of figurative language and its online characteristics. Much of the existing experimental literature on the interpretation and processing of figurative language either does not systematically control for the distinction between conventional and novel metaphors in the design of the stimuli, making the reported results hard to interpret, or explicitly chooses to concentrate on (more or less highly) conventionalized instances. However, the chances of finding effects that would confirm the literal first model in the processing of conventional metaphors will be considerably lower given that the meanings involved are likely to be represented in the mental lexicon, in which case factors (like frequency and saliency; cf. Giora, 1997) enter the picture that do not strictly take part in the frame proposed by literal first. This is a theoretical problem for literal first, but not one that should prevent us from seeing the model as generating general predictions that hold for both types of metaphor, conventional and new. The decision to focus on unconventional metaphors in this series of experiments is a strategic one, in that it should enable us to examine the processing behavior of meanings that are, by definition, not represented in the lexicon and that can therefore not be affected by such extra variables.

GENERAL METHOD

In the experiments reported here, we make a direct comparison between the RTs of metaphorical and literal expressions. We create the best possible matching between material types, as we use the same predicative structures (with differing subjects) for literal and metaphorical conditions. To achieve better online accuracy, we also make use of a technique, self-paced reading, that stays close to the language user's

real-time processing behavior. In a self-paced reading experiment, the reader has to move gradually through a sentence at his or her own pace. A timing device measures the time during which each word within a stimulus sentence remains on the computer screen.

The purpose of these experiments is (a) to study differences and similarities in the online (word-for-word) processing of literal and metaphorical sentences and (b) to do so over two different conditions: with supportive preceding context (providing the ground for the subsequent metaphors and a comparably suitable context for the literal sentences) and with no preceding context. The second concern is also included in the design of the experiments because it has been experimentally demonstrated that a preceding context with a strong supportive value for metaphorical readings will generally facilitate the comprehension of metaphors. In the experiments, literal and metaphorical sentences are of the categorization type “*An x is a y,*” followed by additional linguistic material (relative clause, prepositional phrase, etc.) modifying the category name *y*. Thus, for each target word two sentences are produced that are identical from the predicate slot onward (up until the end of the sentence), differing only in the kinds of subject assigned to the predicate. This differentiation in subject assignment, then, is the sole factor distinguishing between a literal and a metaphorical reading of the resulting categorization statement.

Literal: “*An oak is a tree ...*”

Metaphorical: “*A friend is a tree ...*”

Within the set of literal sentences, a further distinction is made between prototypical and peripheral members of the category at issue (e.g., for tree, *oak* would be prototypical and *maple* peripheral). For metaphorical sentences, we differentiate between apt and nonapt metaphors—that is, between sentence types in which the assignment of a metaphorical subject will result in metaphors of a fairly high quality and those in which this is not the case. With respect to metaphor aptness, a number of cross-modal priming experiments (Blasko & Connine, 1993) have shown that metaphors of low familiarity (the type considered here) do not trigger figurative meanings unless the metaphors in question have been rated highly or moderately apt (i.e., of high or moderate quality). That is, aptness seems to play a role in processes of comprehension when participants are faced with metaphorical statements they have not encountered before. Again, however, the specific locus of the reported effects turns out to be highly volatile and cannot be pinned down on lexical activation patterns for the topic or vehicle of the metaphor. In fact, the authors themselves suggested that the construction of figurative meanings for these metaphors was caused by a set of emergent properties for each metaphorical phrase. The present experimental paradigm, which makes use of self-paced reading, is particularly well suited to deal with issues such as these, as the possibilities for locating the source of effects for figurative meanings follow

automatically from applying a technique that records RTs for each individual word of a stimulus sentence (and not just for topics or vehicles).

When measured on the category name *y* in a self-paced reading task, targets appearing in literal sentences should, on the whole, be read faster than those that occur in metaphorical sentences. This is motivated by the contention, within the literal first hypothesis, that literal meanings need to be (at least partially) processed before a figurative one can start being computed at all. For the experiments reported here, this means that the problematic category status of metaphorical targets needs to be established first. Only then can participants begin to look for possible alternative interpretations (which they typically have to do if they are to understand the meaning of the sentence as a whole). In addition, we predict that differences in RTs on the predicate position of literal sentences will also reflect the degree of membership that can be attributed to the subjects of these sentences, so that targets (e.g., *tree*) will be read more slowly when preceded by peripheral members (e.g., *maple*) than when they appear in sentences containing prototypical subjects (e.g., *oak*). This particular prediction should fall out of the standard results reported in the prototype literature. For unconventional metaphors, the latter qualification relates not so much to the status of the subject as a category member (because no actual membership is assumed in these metaphorical statements), but rather to their contribution to the aptness of the resulting metaphor. Thus, the prototypical versus peripheral status of category members in literal statements is complemented by the distinction in aptness between two types of metaphorical statement. Targets in nonapt metaphors should be read more slowly than those in apt ones, because aptness determines the ease of comprehension for unconventional metaphors.

EXPERIMENT 1

Method

Materials and design. The items in the experiment gave rise to four conditions, presenting two literal and two metaphorical terms per category name. They were selected on the basis of several pretests. Participants throughout the pretests and the experiments were native speakers of Dutch, and the sentences presented to them were in Dutch. No participants were used twice.

In Pretest 1 (see Table 1), a production task featuring 51 category names, participants had to write down, for each category, as many members as they could within a limited period of time (20 min on average, although individual participants who could not finish the task in time were allowed to complete the questionnaire within an additional 5 min). The category labels were distributed over three lists of 17 items. Each of these lists was presented in two random orders to 20 participants (i.e., there were 60 participants in total). On this basis, 24 semantic categories,

TABLE 1
 Pretest 1: Selected Category Names With Prototypical and Peripheral Items

<i>Category</i>	<i>Prototype</i>	<i>Positions 1 + 2</i>	<i>Peripheral</i>
Monster	Dragon [10]	70%	Werewolf [2]
Amusement park (Dutch: <i>pretpark</i>)	Walibi [20]	79%	Phantasieland [3]
Insect	Fly [19]	89%	Beetle [2]
Flower	Rose [20]	85%	Carnation [2]
Tree	Oak [16]	69%	Maple [2]
Artist	Painter [11]	50%	Poet [2]
House	Villa [17]	73%	Farm [2]
Color	Red [17]	88%	Violet [2]
Genius	Einstein [16]	65%	Edison [2]
Organ	Liver [19]	100%	Pancreas [2]
Medication (Dutch: <i>medicijn</i>)	Aspirin [11]	63%	Penicillin [2]
Bird	Sparrow [18]	55%	Owl [2]

Note. The number of tokens generated for each listed category member is indicated in square brackets. Percentages indicating the relative frequencies of prototypical items mentioned in first and second position are included in the third column.

each with its prototypical and peripheral members, could be selected for further use in the experiments.

Only categories for which participants produced more than 10 members were considered. Prototypical items were selected on the basis of their absolute production frequencies (they had to occur at least 10 times; i.e., half of the participants should have mentioned them), as well as a weighted frequency that favored first and second mentions. Peripheral category members, although obviously showing a very low production frequency, had to be mentioned by more than one participant. Concretely, all but one of the selected peripheral category members were mentioned each time by exactly two participants.

An overview of the results obtained for the following two pretests is provided in Table 2, each time limited to those items that have been retained on the basis of earlier selection procedures.

In Pretest 2, we focused on metaphorical combinations with the purpose of discriminating between apt and nonapt terms. To each of the remaining 24 categories, six nonmember terms were added as subjects in a subject–predicate structure of the type “*An x is a y.*” Thus, 144 unfamiliar metaphors were created, distributed over three lists of 48 items (with each predicate or category name appearing not more than twice per list). With two random orders per list, a total of 60 participants had to assess the quality of the metaphorical subject–predicate relation on a 7-point scale ranging from 1 (*nonapt*) to 7 (*apt*). A second group of 60 participants was asked to judge the conventionality of these metaphors on a 7-point scale, as we are only interested in metaphors with a fairly low degree of conventionality and familiarity. As can

TABLE 2
Mean Ratings and Standard Deviations for Selected Items in Pretests 2 and 3

<i>Pretest 2</i>				
<i>Rating Level</i>	<i>Aptness</i>		<i>Conventionality</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
High	4.50	.70	3.66	1.08
Low	1.76	.23	1.53	.34

<i>Pretest 3</i>		
<i>Rating Level</i>	<i>Aptness</i>	
	<i>M</i>	<i>SD</i>
High	3.72	1.18
Low	3.18	1.27

be gathered from the first section in Table 2, we obtained reliable differences between two sets of metaphorical items, which were called apt and nonapt, respectively. In addition, all metaphors scored around or below average for conventionality (i.e., they were generally considered fairly unconventional). When considering the mean rating scores of the items that got selected, aptness and conventionality were highly correlated ($r^2 = .84, p < .001$). On the basis of these results, the 12 categories appearing in Table 1 were selected, each of them giving rise to two metaphors. Only metaphorical items with a low degree of conventionality were marked for selection, and the aptness distinction was distributed equally over these items, so that each category produced one apt and one nonapt metaphor.

The participants in Pretest 3 had to assess the quality of the metaphors remaining from the previous pretest. This time, however, full sentences were presented, consisting of a categorization statement (“*An x is a y*”) plus additional material following the target (e.g., “*A friend is a tree with very firm roots and thick branches*”). Each stimulus was rated by 6 participants. Here, the rating distance dividing apt and nonapt metaphors proved to be smaller than in the previous pretest. This indicates that in an offline task, the addition of lexical material elaborating the ground of the presented metaphors affects their interpretability. We expect such an effect of interpretation to show up again in online tasks toward the end of the stimulus sentence—that is, by measuring RTs on target words that occur fairly late in the course of processing (cf. the Results section). When analyzing the two mean scores for aptness, taken together over Pretests 2 and 3, a highly significant effect was obtained, $F_t(1, 11) = 75.42, p < .001$. Also, the interaction between conditions with and without additional lexical material (the difference between Pretest 2 and

3) and aptness turned out significant, $F_t(1, 11) = 71.75, p < .001$, which indicates that there is a significant effect of aptness between the stimulus types presented in the two pretests; that is, bad metaphors occurring in full-blown sentences (not just bare subject–predicate structures) are generally considered more apt.

Finally, Pretest 4 investigated, for the same 12 categories, whether the contrast in their literal counterparts between the selected prototypical and peripheral items was large enough to be measurable in an RT experiment. To check this, we ran a verification task of the type performed by McCloskey and Glucksberg (1979), using counterbalanced lists. (The experiment was run over 2 nonconsecutive days.) Eight participants had to verify under time pressure whether the prime, the first item presented (during 1 sec), was a member of the category whose name was given immediately afterward (the target). The results showed that the literal prototypical items, as derived from Pretest 1, were indeed verified more rapidly than the peripheral ones (with significant differences in subject and item analyses, both $ps < .05$). This allowed us to use these two (literal) item sets in the actual experiments with their status as prototypical or peripheral category members empirically verified.

In sum, 12 category names were selected, each giving rise to four different conditions depending on the status of the subject term (examples are translated from Dutch):

- Literal and prototypical: “*An oak is a tree with very firm roots and thick branches.*”
- Literal and peripheral: “*A maple is a tree with very firm roots and thick branches.*”
- Metaphorical and apt: “*A friend is a tree with very firm roots and thick branches.*”
- Metaphorical and nonapt: “*A racist is a tree with very firm roots and thick branches.*”

Considering the way the materials have been constructed, they constitute 12 sets of matched quartets. Additional filler material was created with sentences that had the same initial structure (“*An x is a y*”) as the critical items. Of these fillers (24 in total), half were metaphorical in the sense of not providing an established categorization of the grammatical subject. The other half presented literal statements. In turn, half of the metaphorical fillers were considered apt metaphors (as established on the basis of extra material used in Pretest 2), whereas the other half yielded nonapt items. For literal sentences, half of the filler set presented prototypical members and the other half contained peripheral ones. Neither literal nor metaphorical filler items were analyzed in the experiments, as they had not been subjected to Pretests 3 or 4, respectively.

All of the critical items were distributed over four lists, with each of the lists yielding two randomized orders. To each list, the same filler items were added. The

critical items were distributed across the lists according to the following criteria: Each list contained all of the 12 category names (no one category appeared twice in a list), with three instances of each of the four types distinguished earlier. There were 15 participants per list ($N = 60$).

Procedure. Before the experiment started, participants were instructed, orally and in writing, about relevant aspects of the experimental procedure. During the experiment, they were sitting in front of a computer screen in a darkened room. The experimental sentences appeared on the screen one by one. For each sentence, a number of dashes represented the words contained in that sentence without revealing the actual words themselves beforehand. Participants could thus assess the length of the sentences without anticipating the exact nature of their contents. The participants' task was to proceed through the sentence one word at a time, making use of the middle button on a button box. Each time this button was pressed, a new word would appear (and the previous one would disappear). Participants were told to go through the entire sentence this way, maintaining a reasonable reading speed and making sure they saw and understood each of the words making up the sentence. The time-out for individual words was set at 2 sec.

After the sentence was read, the same question always appeared ("Do you agree with this statement?"). At that point, participants had to indicate their answer by pressing the left or right button on a button box. This question was inserted to ensure that participants were motivated to attend to the content of the sentences instead of to their formal characteristics. They were asked to answer the question in a fully personal and subjective way, stressing the focus on content even more. All of the filler and experimental sentences were designed so that their contents made for more or less informative, nontrivial statements, making the task varied enough to hold the participants' attention.

Each experiment contained two pauses of 10 sec, which participants could abort if they wanted to proceed faster.

Participants. For Experiment 1, 60 undergraduate foreign-language students volunteered to participate. Nobody participated more than once. All students were native speakers of Dutch. No volunteers were paid for their participation.

Results. Average RTs were calculated on the target word (i.e., the category name y , which is the point in the sentence where its literal or metaphorical status becomes clear) and on the following word (target + 1) to check for spillover effects. The results for target + 1 are not reported here, as they completely match those for the target word itself. RTs were also measured on target2 and on target2

+ 1. The second target occurs at the syntactic end of a clause. It indicates the first point in the sentence following the actual target at which a complete grammatical clause can be construed (in the example “*An X is a tree with very firm roots and thick branches,*” target2 will coincide with the word *roots*). This second target thus indicates the point at which participants have enough sentential material to wrap up their interpretation of (part of) the proposition. It is experimentally demonstrated (as discussed in Frazier, 1999; see also Abrams & Bever, 1969) that additional processing can be assumed to go on at this particular point in the processing of a sentence. For the metaphorical statements, this means that something of a stable metaphorical interpretation becomes available due to the sufficient amount of (incrementally processed) preceding information. Therefore, in case metaphorical sentences of the present type cause participants to put their interpretation on hold until more material is available for interpretation, result patterns for target2 should differ from those for the first target. In addition, RTs on the word immediately following this second target (*and* in the preceding example) are recorded to check whether the additional processing occurring on target2 spills over to the following region target2 + 1. Target2 + 1 is always a function word, a grammatical expression that adds little or no lexical information to the proposition at issue. In the experiments reported here, this position is typically filled by simple conjunctions (*and, but*), relative pronouns, or prepositions. Given the lexically impoverished nature of this class, it is to be expected that processing properties for such words will show little or no significant variation.

In discussing the results for this and the following experiment, we perform analyses of variance (ANOVAs) by subjects and items. These tests indicate the probability that subjects (or some related procedure) and items can be treated as random effects. If a level of statistical significance is reached, it will be implied that the results obtained are justifiably generalizable over subjects or stimulus materials; that is, the probability of their random nature is negligible. For by-subjects analyses, this means that the sample subjected to the experiment is representative of an entire population (typically to be interpreted as the average native speaker). For by-items analyses, the failure to find a significant effect would suggest that an effect is restricted to (part of) the set of materials used in the experiment; that is, this set is not representative of the experimental topic at hand. In the latter case, it is likely that the items that produce an effect have some properties that the experimenter has not noticed.

Table 3 shows the RTs for the four conditions in Experiment 1 on target, target2, and target2 + 1. An ANOVA for a 2 (sense: literal or metaphorical) \times 2 (quality: prototypical–peripheral and apt–nonapt) design was performed on the RTs of target words. On the first target word, literal sentences were read significantly faster than their metaphorical counterparts, resulting in an overall sense effect for the subject and the item analysis, $F_s(1, 59) = 6.20, p < .05$; $F_i(1, 11) = 6.60, p < .05$. Pairwise analyses between and within conditions only indicate significant differences be-

TABLE 3
 Experiment 1: Mean Reaction Times (Msec) and Standard Deviations on Target,
 Target2, and Target2 + 1

Condition	Target (= Category Name)		Target2		Target2 + 1	
	M	SD	M	SD	M	SD
Literal, prototype	514	212	564	238	525	169
Literal, peripheral	476	168	607	280	491	126
Metaphor, apt	554	248	631	327	518	176
Metaphor, nonapt	545	244	650	305	539	162

tween the literal–peripheral condition and the two metaphorical conditions (both $ps < .05$). The comparisons between literal–prototypical and the two metaphorical conditions are not significant (both $ps > .08$).

On target2, the overall sense effect between literal and metaphorical sentences remains for the subject analysis, $F_s(1, 59) = 4.71, p < .05$, and is marginal in the item analysis, $F_i(1, 11) = 3.76, p < .08$. At this point in the processing of the sentence, the prototypical condition stands out as the fastest one by about 40 msec, when compared to the literal peripheral condition (as opposed to the 31 msec in the opposite direction on target). Still, no significance can be found between them, $F_s(1, 59) = 1.89, p > .1; F_i < 1$. The aptness distinction was nonsignificant on this position as well (both $F_s < 1$). Pairwise analyses between conditions only yield significance for comparisons between literal–prototypical and the two metaphorical conditions: prototypical versus apt, $F_s(1, 59) = 4.57, p < .05; F_i(1, 11) = 2.30, p > .1$; prototypical versus nonapt, $F_s(1, 59) = 7.51, p < .01; F_i(1, 11) = 3.78, p < .08$. The two other comparisons are nonsignificant (both $ps > .11$).

No sense effect was found on target2 + 1, $F_s(1, 59) = 1.84, p > .1; F_i < 1$. All pairwise analyses yield nonsignificant effects (all $ps > .1$), except the by-subject comparison between literal–peripheral versus nonapt metaphors, $F_s(1, 59) = 5.31, p < .05; F_i(1, 11) = 2.27, p > .15$. The absence of a sense effect here, and the fact that this is the main point of distinction with respect to the preceding data points, suggest that our selection of target2 as a point of syntactic closure is well motivated. It might be argued that no differences between metaphorical and literal conditions surfaced at this point because of a possible floor effect. However, given the observed effects reported for the same position in the following experiment, this seems very unlikely.

In general, Experiment 1 shows that it takes longer to process the same set of words when a predicate assignment is interpreted metaphorically than when it is interpreted literally. This may not be all too surprising, of course, as participants in the experiment received no preceding context sentence. They may have been slow on the metaphors simply because they missed the relevant informa-

tion to make the intended metaphorical interpretation. From this perspective, Experiment 1 can be considered a benchmark for Experiment 2, which introduced a context motivating in advance the sentence that followed it, whether metaphorical or literal. In the following experiment, we also pay particular attention to the absence of an overall sense effect that was found on target2 + 1 in Experiment 1.

EXPERIMENT 2

Method

Materials and design. The experimental materials and design were the same as those in Experiment 1. Each sentence was now preceded by a context sentence. Items in this experiment were of the following type.

Example 1

Context:

“Great deeds don’t need large audiences.”

Target:

“A painter/poet/spider/bear is an artist [target] who lives on his talents [target2], in silence.”

Example 2

Context:

“The smallest seed can grow into something big.”

Target:

“A(n) oak/maple/friend/racist is a tree [target] with very firm roots [target2] and thick branches.”

The context sentences were the same for literal and metaphorical conditions. They had been carefully selected on the basis of pretests. In one pretest, participants had to choose between two possible context sentences on the basis of considerations of semantic integration. These context sentences preceded all four conditions distinguished in the experiments (literal: prototypical vs. peripheral; metaphorical: apt vs. nonapt), as equally distributed over different lists; that is, with each list containing the same number of sentences for each of these conditions. In another pretest, participants had to assess the degree of semantic congruity between context and target sentences on a 7-point scale ranging from 1 (*low integration*) to 7 (*high integration*). In both pretests, although critical items were obviously preceded by meaningful context sentences, half of the filler items followed nonsense context sentences (i.e., grammatical sentences unrelated to the topic of the target sen-

ences), included to counterbalance the high degree of semantic integration for appropriate context sentences. The four conditions were matched on contextual fit.

In the experiment, the relation between a target sentence and its preceding context is such that one or several words appearing in the target belong to the same domain as, or are otherwise semantically linked to (part of), the content evoked in the context. For instance, the word *artist* is closely related to the word *audiences* in the target sentence (and even more so in the original Dutch example, which made use of an expression, *publiek*, that typically occurs in the context of artistic stage productions). Any possibility of interpreting other words in the target sentence as themselves metaphorical (in addition to the dominant metaphor elaborated by the target sentence) was excluded, except if such terms appeared after the final data point, target2. When preceding metaphorical target sentences, the context sentence is seen here as providing the ground for the metaphorical interpretation of the target, because it is supposed to motivate the particular similarities that are evoked by the metaphorical description. This is usually done by simply highlighting one of the more salient properties of the so-called vehicle, which can be done directly (e.g., *trees* grow from small *seeds*) or through inference (e.g., an *artist* shows or performs his or her work in relation to an *audience*).

The context sentences presented in this experiment are limited compared to the use of more elaborate preceding context in other, similar experiments. Again, this is a strategic consideration, in that finding effects (or the absence thereof) that would differ from the ones in Experiment 1 (i.e., obtaining an effect of context) with such a small context will constitute harder proof of the influence of a preceding context on metaphor comprehension. Previous research (Inhoff, Lima, & Carroll, 1984; Ortony, Schallert, Reynolds, & Antos, 1978) has shown that, typically, a preceding context needs to be fairly long (in any case, longer than the 4- to 10-word range given as short context by Inhoff et al., 1984) for metaphorical interpretations to be available as fast as literal ones. However, it should be pointed out in this respect that the context used in these studies did not systematically exploit a (conceptual) relation between its own content and that of the metaphorical target sentence, in contrast to the context type employed in the present experiment. In fact, Inhoff et al. (1984) correctly pointed out that metaphorical sentences following the type of context used by Ortony et al. (1978) behaved similarly to literal sentences preceded by semantically unrelated (literal) contexts, which shows that the processing problems involved were not necessarily due to the distinction between literal and figurative conditions, but rather to issues of conceptual integration and what might be called discourse coherence (on a local scale). In contrast, when contexts were used that did contain metaphorically interpretable elements, RTs for metaphorical sentences dropped considerably. In a similar vein, Gildea and Glucksberg's (1983) search for a minimal appropriate context leading to immediate metaphor comprehension suggests that preceding context sentences, if they are to facilitate metaphor comprehension, should at least contain references to (sa-

lient) properties of the metaphorical vehicle (here and in Gildea & Glucksberg's case, the sentence predicate). Again, this is true for the context sentences that were used in this experiment as well. As a consequence, the length of the preceding context, when appropriately constructed, is of secondary importance. What is important is the interaction between length and content; the preceding context should be long enough to allow participants to construe a workable ground for the interpretation of the metaphor that is to follow.

Procedure. The procedure for Experiment 2 was the same as that for Experiment 1, except that a context sentence preceded each target sentence. The entire context sentence appeared on the screen at once (i.e., no self-paced reading was required for this part of the materials). Having read this context sentence, participants pushed the middle button on a button box to make the sentence disappear, at which time they could proceed with the actual target sentence in the manner described for Experiment 1. This time, the time-out for individual words in the target sentences was set at 2.5 sec. Because target sentences were preceded by context, we had to grant a minimal amount of extra reading time to allow participants to successfully integrate the lexical content of each word with this previous information. Also, given the higher processing load placed on participants because of the inclusion of context material, we wanted to make sure that not too many time-outs occurred that would be due to effects of fatigue.

Participants. In Experiment 2, 60 undergraduate foreign-language and business students volunteered. Nobody participated more than once in any of the pretests. All students were native speakers of Dutch. No volunteers were paid for their participation.

Results. Figure 1 compares results from Experiment 2 with those from Experiment 1, for the target position.

Globally, all RTs are much shorter than those in Experiment 1 (a difference of about 100 msec on average). As in Experiment 1, the overall sense effect on the target position, with shorter RTs for the literal predicates (see Table 4), is significant in the subject analysis, $F_s(1, 59) = 4.34, p < .05$, and marginal in the item analysis, $F_i(1, 11) = 3.94, p < .08$. The pairwise analyses are nonsignificant (all $ps > .1$), except for the comparison between literal-peripheral versus apt metaphors, $F_s(1, 59) = 3.47, p < .07$.

On target2, the overall sense effect is marginal for the subject analysis, $F_s(1, 59) = 3.60, p = .06$, but not significant for the item analysis, $F_i(1, 11) = 3.53, p = .09$. Pairwise analyses do not yield significance (all $ps > .1$), except for the comparison

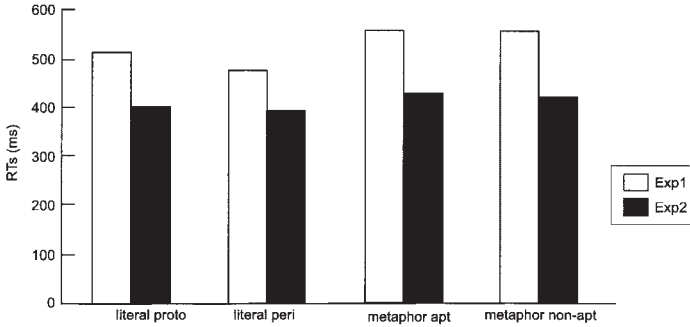


FIGURE 1 Reaction times (RTs) for Experiments 1 and 2, on target.

TABLE 4
Experiment 2: Mean Reaction Times (Msec) and Standard Deviations on Target,
Target2, and Target2 + 1

Condition	Target (= Category Name)		Target2		Target2 + 1	
	M	SD	M	SD	M	SD
Literal, prototype	396	106	449	75	436	151
Literal, peripheral	389	110	503	102	433	109
Metaphor, apt	422	160	525	129	439	151
Metaphor, nonapt	415	152	498	80	480	168

between literal prototypes and apt metaphors, $F_s(1, 59) = 5.36, p < .05$; $F_i(1, 11) = 5.78, p < .05$.

On target2 + 1, no sense effect can be distinguished between literal and metaphorical conditions, $F_s(1, 59) = 2.99, p < .09$; $F_i(1, 11) = 1.59, p > .2$. This absence of a sense effect at a rather advanced stage in the processing of the sentence corresponds to what has been found for the same position in Experiment 1. In addition, target2 + 1 in Experiment 2 offers no significant prototype effect within the literal conditions (both $F_s < 1$). Importantly, the effect of aptness was significant at this point, $F_s(1, 59) = 4.93, p < .05$; $F_i(1, 11) = 4.46, p < .06$. At target2 + 1, apt items were processed as fast as prototypical and peripheral items (no significance, $ps > .10$). In contrast, all comparisons between conditions with nonapt metaphors are significant or marginal; literal–prototypical versus nonapt metaphors: $F_s(1, 59) = 5.47, p < .05$; $F_i(1, 11) = 4.96, p < .05$; literal–peripheral versus nonapt metaphors: $F_s(1, 59) = 6.35, p < .05$; $F_i(1, 11) = 5.75, p < .05$. Experiment 2 differs from the first experiment, then, mainly in the emergence of an aptness effect toward the end of a clause, which is even more striking when we consider the item analysis, given the limited number of metaphori-

cal items (12) used in the experiment. Figure 2 summarizes the results for Experiment 2 over the three data points distinguished.

In comparison with Experiment 1, we notice, first of all, the systematically lower RTs for all data points concerned, due to the effect of a preceding context for both literal and metaphorical conditions. Experiment 2 shows an overall sense effect between literal and metaphorical conditions on target and target2. Interestingly, on target2 + 1 the sense effect that disappeared in Experiment 1 is also absent in Experiment 2, but nonapt metaphors perform ostensibly worse at this point (compared to the other conditions for Experiment 2) than in Experiment 1. As noted earlier, it is also at this point that a significant effect could be observed between apt and nonapt metaphors. In fact, we see that it is the metaphorical condition containing apt items that practically coincides with the two literal conditions. In other words, on target2 + 1 apt metaphors behave as if they have been more or less fully interpreted (at least as fully as their literal counterparts), whereas nonapt metaphors continue to cause interpretive difficulties. On the whole, RTs on target2 + 1 for apt metaphors and both types of literal condition are considerably lower than in Experiment 1 (about 75 msec on average).

DISCUSSION

The purpose of our experiments was to provide an online measurement of the processing characteristics involved in the interpretation of unfamiliar metaphors. To this effect, we made a comparison with matched literal sentences, with and without preceding context. In addition, we wanted to find out whether processing speed is determined by degree of prototypicality in literal classifications and degree of aptness in metaphorical ones. RTs were measured within the sentence, instead of calculating global RTs for whole sentences, because possible sense effects might show up in the course of incremental processing and disappear again toward the end, when an inter-

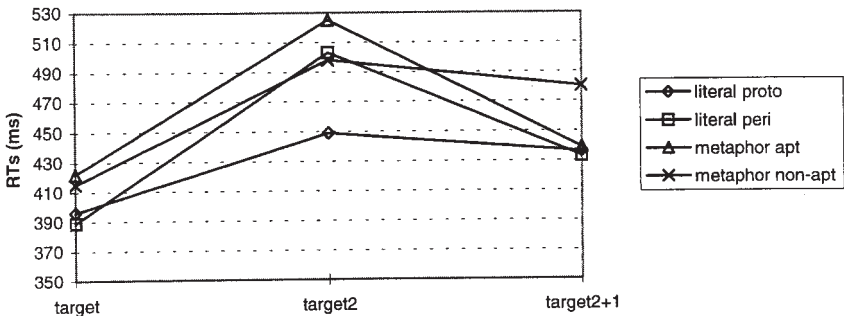


FIGURE 2 Reaction times (RTs) for target, target2, and target2 + 1 (Experiment 2).

pretation has been arrived at. These experiments show that, in the early stages of processing, the comprehension of both apt and nonapt unconventional metaphors is slower than for matched literal sentences. Thus, the results obtained for the early stages of metaphor processing are in line with the types of experimental finding proposed by Gerrig and associates (e.g., Gerrig, 1989; Gerrig & Healy, 1983), who stressed the extra initial processing efforts caused by interpreting familiar words with unfamiliar meanings. However, these findings do not offer any conclusive evidence in favor of or against particular processing models (parallel vs. serial). Toward the end of a sentence (target2 + 1), the distinction in RTs between literal sentences and apt metaphors disappears, although nonapt metaphors continue to perform relatively worse when preceded by context. It is also suggested that, at target2 + 1, apt metaphors preceded by a relevant context are as fully interpreted as their literal counterparts. At this late point in the processing of a sentence, target2 + 1 itself does not instantiate or add to the difference between a literal and a metaphorical interpretation, so that we must attribute the disappearance (apt) or persistence (nonapt) of a sense effect to mechanisms of incremental processing that are at work in the build-up toward this point. The results reported here disconfirm the theoretical preferences advanced by Gibbs (e.g., 1994), who insisted that extra processing resources are not obligatorily devoted to understanding metaphors, conventional or new.

The data on the processing of literal and unfamiliar metaphorical sentences are compatible with a literal first model, in the sense that the initial activation of literal meaning in the course of processing is empirically demonstrated. On target position (i.e., the word at which the literal or metaphorical nature of the sentence becomes clear, always the predicate term), we found a significant sense effect. Metaphorical predicates took longer to read than literal ones. This effect is to be expected when participants have no preceding context sentence which they can use to interpret the metaphor (Experiment 1), but it remained present when a preceding sentence provided them with the ground of the following metaphorical statement (Experiment 2). In other words, on encountering a noun that calls for an unconventional (metaphorical) classification, participants must initially process the literal meaning of this word, whether contextual information is available or not. However, this does not necessarily mean that the interpretation of unfamiliar metaphors needs to wait until the literal interpretation has been rejected and, thus, that one is forced to adopt a sequential model like the literal first hypothesis to explain these experimental findings. Longer RTs for metaphorical sentences are compatible with the literal first model but do not by themselves confirm the model.

Because we are dealing with novel metaphors, in which new meanings are triggered for old words that occupy a predicate position, such new meanings need to be created online by the language user, who is reasonably assumed to rely on meaning representations that are already available for these predicate terms. Thus, the work of sense creation, typical of the interpretation process for unfamiliar metaphors, must operate to supplement ordinary sense selection (i.e., the mere re-

trieval of existing representations). To fully back a stage model of metaphor comprehension, such as the literal first model, it should be established whether the process of sense creation can only be initiated after sense selection fails (in that the latter process produces an erroneous interpretation). However, the only conclusion that is readily available from these data is that the literal meanings of metaphorically used predicates are indeed activated in the very early stages of processing, not that these literal meanings determine the full extent of the time course involved in the interpretation of this type of metaphor. In particular, a parallel processing model might be equally compatible with the results of both experiments (see Gerrig, 1989). In this scenario, sense selection (or retrieval, for literal meanings) and creation (for figurative ones) operate simultaneously, at least after the processor has established that a new meaning needs to be created in the first place. In short, longer RTs for metaphorical sentences do not automatically imply that the reader can only find a contextually appropriate nonliteral interpretation on the basis of the literal meanings that have been retrieved first.

Interestingly, the sense effect detected on the target position persists at target2, the last word of the syntactic phrase to which the predicate term belongs. It was indicated earlier that, on top of the immediate (incremental) processing that goes on in the interpretation of sentences, potential clause boundaries (e.g., our target2) function as loci of interpretation, or points at which participants attempt to integrate the meaning of the encountered clause material. The fact that the sense effect still turns up at this point in the sentence, even when there is contextual support for interpretation, demonstrates that the interpretation of unfamiliar metaphors is a time-consuming process that lasts well beyond the moment at which the metaphorical term itself is introduced.

One might want to reject this account and attribute the persistent sense effect to an absence of interpretation for the metaphorical conditions in the experiments. According to this line of reasoning, participants experienced the unfamiliar metaphors as instances of noninterpretable language use and did not start, or rapidly aborted, an interpretive routine. However, one particularly salient piece of evidence from Experiment 2 refutes such an interpretation. On target2 + 1, we found that, for the first time in the experimental series, the two groups of metaphors (apt and nonapt) behaved differently. At this position, immediately following the syntactic boundary marked by target2, RTs for apt metaphors were nondistinct from those for literal sentences (the three conditions not differing among each other), whereas RTs for the nonapt metaphors were significantly longer than those for the apt ones. We take this finding to indicate the operation of an interpretation process that is more successful for the apt metaphors than for the nonapt ones. This interpretation process has been initiated on target position and is more time-consuming for all metaphors (hence the sense effect there). It continues up to the final position in the clause, where both types of metaphor still take more processing time than both types of literal sentences (hence the sense effect on target2). On target2 + 1, the divergence between the apt and nonapt metaphors

indicates that processing up to target2 has resulted in an acceptable interpretation for the apt metaphors (or that their interpretation has at least been as successful as that for the literal sentences), whereas such an interpretation has not been arrived at for the nonapt metaphors. In Experiment 1, we did not observe the same divergence between the two types of metaphor. Because in this experiment there was no preceding context, the initiated interpretive process on the target position, which was still going on at target2 (hence the sense effect at both positions again), is not likely to be successful for both types of metaphor. Participants may fail to come up with a metaphorical interpretation for lack of sufficient contextual support. If the interpretation runs aground on all metaphorical sentence types, it could be argued that participants may not attempt any further processing and abort the interpretive process to continue with the rest of the sentence. The end of a syntactic clause is a good point for abandoning unsuccessful processing attempts and resetting the semantic processor. This would account for the finding in Experiment 1 that, on target2 + 1, RTs for both metaphor types are statistically nondistinct from those for the literal sentences.

In both experiments, we failed to find a reliable prototype effect within the group of literal sentences; that is, sentences with prototypical subjects were not processed faster than sentences with peripheral ones. This finding was constant across experiments and measurement positions (target, target2, target2 + 1). This is a remarkable finding considering the robustness of prototype effects in a variety of experimental tasks as reported in the literature. Recall that the lack of this effect cannot be due to item selection, as the same subjects and predicates taken from these experiments gave rise to a significant effect in a traditional category verification task (see Experiment 1, Pretest 4). Although further research is required to replicate this effect, we suggest that the category information that is mobilized in the context of a simple reading task differs from the information that comes to bear on the essentially metalinguistic tasks in which prototype effects have been demonstrated (category verification, rating tasks, member generation, etc.). In other words, the lack of a prototype effect in our self-paced reading experiments does not discredit the frequently attested prototype effects but suggests, rather, that the information on which these effects are based is less involved in online sentence processing. Eye-tracking data remain inconclusive in this respect (see Duffy & Rayner, 1990; Liversedge & Underwood, 1998).

Finally, the results reported here are also of some methodological importance. We pointed out earlier that previous research had generally failed to measure processing while participants are reading metaphorical (or, in general, figurative) sentences in real time, relying instead on rather crude measures like total sentence RT (but see Frisson & Pickering, 1999). Our finding of different response patterns at different sentence positions indicates that it is unwise to collapse data across sentence positions in this kind of research (which is the case when total sentence RTs are used). One may well lose important effects, which can turn up at only one theoretically relevant position, if these are drowned in the measurements for whole sen-

tences only. Especially if one wants to make statements on the literal first model, it is risky to neglect the precise time course of processing from the metaphorical term onward. This risk may be particularly high in the domain of familiar metaphors, where an initial processing delay on the metaphorical target itself may rapidly disappear if participants quickly manage to arrive at a metaphorical interpretation (which is not unlikely, given the interpretive success of apt unfamiliar metaphors toward the end of the clause in the present experiments). We emphasize that research on metaphor processing will benefit from the use of experimental techniques that can track the time course of such processing word by word.

ACKNOWLEDGMENTS

Steven Frisson is also affiliated with the University of Glasgow, Department of Psychology, Human Communication Research Centre.

This article is a revised and substantially elaborated version of an earlier preprint: Brisard, Frisson, and Sandra (1999). This study is supported by Grant G.0246.97 from the *Fonds voor Wetenschappelijk Onderzoek – Vlaanderen*.

We thank Kate Nation for helping us out with some of the analyses and allowing us to conduct a small supportive experiment at the University of York.

REFERENCES

- Abrams, K., & Bever, T. G. (1969). Syntactic structure modifies attention during speech perception and recognition. *Quarterly Journal of Experimental Psychology*, *21*, 280–290.
- Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 295–308.
- Brisard, F., Frisson, S., & Sandra, D. (1999). Processing unfamiliar metaphors during self-paced reading. In The Japanese Cognitive Science Society (Eds.), *Proceedings of the Second International Conference on Cognitive Science* (pp. 86–91). Tokyo: The Japanese Cognitive Science Society.
- Dascal, M. (1989). On the roles of context and literal meaning in understanding. *Cognitive Science*, *13*, 253–257.
- Duffy, S. A., & Rayner, K. (1990). Eye movements and anaphor resolution: Effects of antecedent typicality and distance. *Language and Speech*, *33*, 103–119.
- Frazier, L. (1999). *On sentence interpretation*. Dordrecht, The Netherlands: Kluwer.
- Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1366–1383.
- Gerrig, R. J. (1989). The time course of sense creation. *Memory & Cognition*, *17*, 194–207.
- Gerrig, R. J., & Healy, A. F. (1983). Dual processes in metaphor understanding: Comprehension and appreciation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 667–675.
- Gibbs, R. W. (1984). Literal meaning and psychological theory. *Cognitive Science*, *8*, 275–304.
- Gibbs, R. W. (1992). When is metaphor? The idea of understanding in theories of metaphor. *Poetics Today*, *13*, 575–606.
- Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. New York: Cambridge University Press.
- Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior*, *22*, 577–590.

- Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 8, 183–206.
- Glucksberg, S., Gildea, P., & Bookin, H. B. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, 21, 85–98.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). New York: Academic.
- Inhoff, A. W., Lima, S. D., & Carroll, P. J. (1984). Contextual effects on metaphor comprehension in reading. *Memory & Cognition*, 12, 558–567.
- Liversedge, S. P., & Underwood, G. (1998). Foveal processing load and landing effects in reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 201–222). Oxford, England: Elsevier.
- McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, 11, 1–37.
- Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior*, 17, 465–477.
- Searle, J. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 92–123). Cambridge, England: Cambridge University Press.