Running Head: COMPREHENDING ANAPHORIC METAPHORS

Comprehending Anaphoric Metaphors

Raluca Budiu

John R. Anderson

Carnegie Mellon University

Pittsburgh, PA 15213-3890

raluca@andrew.cmu.edu

ja+@cmu.edu

August 30, 2001

Manuscript #00MG-355 (01-174) accepted for publication by Memory & Cognition.

Please address correspondence to **Raluca Budiu**, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890. Electronic mail may be sent to raluca@andrew.cmu.edu or ja+@cmu.edu.

Abstract

In this study we investigate the comprehension of various kinds of anaphoric metaphors in context. We describe an experiment that manipulated the metaphoricity of simple noun + verb + ending sentences by using either a metaphoric noun or a metaphoric verb or both. Our results show that metaphoric nouns affect the sentence comprehension to a greater extent than metaphoric verbs. Thus, even though there were no sentence-reading time differences between metaphoric and literal targets, metaphoric nouns were read more slowly than literal nouns and they also affected the reading time of the following verb. Moreover, in trials involving metaphoric-noun targets, subjects read the endings of the targets faster and made more mistakes in answering post-trial questions than they did in literal-noun trials. We argue that these results suggest a comprehensioni deficit for anaphoric noun metaphors even when they are preceded by a context.

Comprehending Anaphoric Metaphors

Many studies suggested that people understand metaphors as easily as they understand literal sentences. For instance, in a 1978 experiment, Ortony, Schallert, Reynolds, and Antos (1978) showed subjects either a passage about a women's club meeting or about chickens on a farm and and followed each of them by a target sentence such as "The hens clucked noisily." When it came after the first passage, the sentence had a metaphoric interpretation; after the second passage it was literal. Participants in Ortony et al.'s experiment read this sentence as fast in both conditions. This result was interpreted as evidence that, when context is rich and supportive, people process metaphoric sentences as fast as literal sentences and contradicted Searle's (1979) theory of metaphor comprehension. Searle's theory asserts that, to understand a metaphoric utterance, people first need to compute its literal interpretation and, only if it does not make sense, do they proceed to search for a metaphoric interpretation. Further studies (Glucksberg, Glidea, & Bookin, 1982; Goldvarg & Glucksberg, 1998; Inhoff, Lima, & Carroll, 1984; Keysar, 1989; Shinjo & Myers, 1987) supported the assumption that similar processes are involved in the comprehension of both literal and metaphoric sentences and that metaphoric interpretation is not optional (i.e., people access it even when they do not need it for performing the task).

Janus and Bever (1985) replicated Ortony et al.'s (1978) findings for metaphors embedded within a rich context; however, beside measuring the sentence-reading times, they looked at the reading times for the metaphoric nouns. Even though, like Ortony et al. (1978), they found no significant difference between reading times for metaphoric and literal sentences, the reading times for metaphoric nouns were longer than those for literal nouns. This result threw some doubt over the idea that the same mechanism is involved in the comprehension of metaphoric and literal language.

A later study by Gibbs (1990) also provided some support to the Searle's model of metaphor comprehension. Gibbs showed subjects short passages followed by either a metaphoric or a literal sentence. For instance, one such passage was about a boxing match and ended either with a metaphoric sentence such as "The creampuff did not show up for the match" or with its literal equivalent "The boxer did not show up for the match". Gibbs did find a reading time disadvantage for metaphoric sentences with respect to literals, but attributed this result to the type of metaphors used — anaphoric in his study versus predicative in those studies that had provided evidence for similar literal- and metaphor-comprehension processes (Glucksberg et al., 1982; Inhoff et al., 1984; Keysar, 1989; Shinjo & Myers, 1987). Predicative metaphors are of the form "A is B" (e.g., "marriages are iceboxes", "time is money"). In contrast, in sentences that contain anaphoric metaphors the metaphoric term (vehicle) is used to refer anaphorically to some previously introduced concept (e.g., in the sentence "The creampuff didn't show up", "creampuff" refers metaphorically to the concept "boxer", introduced in the preceding discourse). However, even though the metaphors employed in their study were also anaphoric, Ortony et al. (1978) failed to find a difference between sentence-reading times for literal and metaphoric sentences.

We wanted to investigate more closely the reasons between the divergent results obtained by Ortony et al. (1978) and Gibbs (1990). We noted that one difference between Ortony et al. (1978) and Gibbs (1990) is that the former used as targets only sentences that made literal sense. In other words, whether preceded by a farm context or by a meeting context, the sentence "The hens clucked noisily" in Ortony et al.'s (1978) experiment is interpretable literally. In Gibbs' (1990) study, the targets seemed not to have a literal interpretation ("The creampuff didn't show up for the match" is not a valid literal sentence). Note that a metaphoric sentence with literal sense typically contains multiple words used metaphorically (e.g., when embedded in the meeting context, "The

hens clucked noisily" contains two metaphors "Women are hens" and "Talking is clucking"; compare that with "The creampuff didn't show up"). It is possible that in Ortony et al. (1978) metaphoric sentences with literal sense were not processed metaphorically at all, but rather understood in isolation from the preceding discourse. If that were the case, the initial delay for reading metaphoric nouns reported by Janus and Bever (1985) may be explained by subjects having not found antecedents for them; later, however, if the sentence were perceived as isolated from the preceding text, subjects may have skipped processes related to integration with preceding discourse and, thus, compensated for the initial slowdown.

Whereas most studies in the metaphor literature measured reading times for metaphoric sentences in context, their authors paid less attention to whether subjects had understood correctly the metaphoric targets. Although some studies do report accuracy figures, either the accuracy measures reflected passage comprehension rather than target comprehension (Gibbs, 1990; Janus & Bever, 1985; Onishi & Murphy, 1993) or the authors did not compare the accuracy in the two conditions of interest (metaphoric and literal) (Ortony et al., 1978).

In the face of this inconclusive array of results we decided that it would be useful to perform a study that could clear up some of the open issues. First, motivated by Janus and Bever's (1985) results, we wanted to separately assess the effect of metaphoricity on the reading of different parts of a sentence. Second, motivated by the divergence between Gibbs (1990) and Ortony et al. (1978), we wanted to investigate whether there was a difference between the processing of metaphoric sentences with a literal sense as well and metaphoric sentences without a literal interpretation. Third, concerned about the uncertain data about comprehension success, we wanted to determine whether there was a comprehension deficit for metaphoric sentences.

Experiment 1

In this experiment we showed subjects a short passage followed by any of the following target sentences: (a) metaphoric–literal sentences, in which both the subject and the verb are metaphoric (e.g., "The hens clucked noisily" in the context of a passage about some women in a meeting); (b) metaphoric–literal sentences, in which the subject only is metaphoric (e.g., "The hens talked noisily"); (c) literal–metaphoric sentences, in which the verb only is metaphoric (e.g. "The women clucked noisily"); and (d) literal–literal sentences, in which both the noun and the verb are literal (e.g. "The women talked noisily"). The metaphoric–metaphoric sentences correspond to those used by Ortony et al. (1978) and Janus and Bever (1985); the metaphoric–literal sentences correspond to those that Gibbs (1990) used in his study; and the literal–literal sentences correspond to literal controls in the previous studies. We collected reading time measures for different parts of the target sentence. In a pretest for Experiment 1, we also looked at reading latencies for the target sentences when they were presented in isolation, with no preceding passage, in a sentence completion task.

Method

Participants. Eighty-three undergraduates at Carnegie Mellon University participated in Experiment 1 as part of Psychology course requirement. They were all native English speakers.

Materials. We chose 28 metaphor pairs made of one noun metaphor and one verb metaphor (e.g., <women–hens, talk–cluck>). These pairs had to satisfy the constraint that the metaphoric-noun–metaphoric-verb combination ("hens cluck") makes literal sense. For each such pair we created a short passage that preceded the target sentence. The structure of all target sentences was noun + verb + ending part. The noun was articulated with a definite article and the verb was either in the present or in the past

tense. The ending part was chosen such as not to violate the literal sense constraint and was the same for all target sentences obtained from one metaphor pair (e.g., "noisily" for the pair <women – hens, talk – cluck>), but it varied from pair to pair. Usually, the ending part consisted of any of an adverb, an indefinite pronoun (such as "anybody" or "anything"), or prepositional noun phrase (e.g., "at school", "in the morning" etc.). There were four types of targets: (a) metaphoric–metaphoric targets, with metaphoric noun and verb; (b) metaphoric–literal targets, with metaphoric noun and literal verb; (c) literal–metaphoric targets, with literal verb and metaphoric noun; and (d) literal–literal targets, with literal noun and verb.

For each passage we created two probe sentences (one true and one false) that subjects had to judge as true or false. One probe sentence was selected randomly to be judged by each participant. Table 1 shows two examples of passages, targets and probe sentences.

Ratings of metaphors. It is reasonable to expect that the goodness and the familiarity of a metaphor may influence the speed at which it is processed. Nonetheless, previous studies (Gerrig & Healy, 1983; Tourangeau & Sternberg, 1981) showed that the goodness of a metaphor is not necessarily correlated with the ease of comprehension (but see Tourangeau & Rips, 1991; Blasko & Connine, 1993, for counterexamples). However, to make sure that the selected metaphors were not too good or too familiar, we conducted a rating study. In this study, 10 native English speakers, students or staff of Carnegie Mellon University, rated the goodness and the familiarity of 179 metaphors on a scale from 1 to 4 (1 — low familiarity/goodness; 4 — high familiarity/goodness). From those 179 metaphors, 16 were used in another study, and 56 corresponded to the 28 metaphor pairs from this experiment[1]. From the remaining 107 metaphors, 45 were fairly well known (and presumably good) metaphors used in every day language and 45 were nonsensical metaphors. A lot of the nonsensical metaphors were adapted from existent

literature (Gerrig & Healy, 1983; Ortony, Vondruska, Foss, & Jones, 1985). Some of the good metaphors were taken from Ortony et al. (1985) and from Inhoff et al. (1984). Using all these kinds of metaphors ensured that subjects use all the points on the scale.

Each participant saw the metaphors in a random order. Table 2 compares the averages of the goodness and familiarity ratings for the metaphors used in this experiment with those corresponding to good metaphors and to nonsensical metaphors.

Unfortunately our metaphoric nouns or verbs were not matched in terms of frequency and length to their literal counterparts. Based on Francis and Kucera's (1982) index, the average frequency was 11.21 for metaphoric nouns and 197.86 for literal nouns. Metaphoric verbs averaged 4.61 in frequency and literal verbs averaged 65.64. The difference in frequency between corresponding metaphoric and literal nouns was significant by a t test ($t(27) = -2.71$, $p < 0.05$), as was the frequency difference between metaphoric and literal verbs ($t(27) = -2.57$, $p < 0.05$). In terms of length (measured in characters), the metaphoric and literal items were more similar than in terms of frequency. The average noun length was 5.82 for metaphors and 5.25 for literals; the average verb length was 6.64 for metaphors and 6.32 for literals. There was no significant difference between literal and metaphoric nouns ($t(27) = 1.26$) or verbs ($t(27) = 0.73$). To compensate for the different lexical properties of metaphoric and literal items, we designed a pretest in which we collected reading times of the target sentences out of context. We report the pretest at the end of the Method subsection.

Ratings of probe sentences. A possible problem with judging probe sentences is that what an experimenter judges as a true or false sentence may be actually categorized differently by subjects. To avoid this problem, we asked the same 10 subjects who participated in the metaphor rating study to judge the truth of our probe sentences in a paper–and–pencil task. Each participant got a booklet containing 36 passages, including the 28 used in our main experiment. A story consisted of a passage ending with a target

sentence. The type of the target sentence was randomly selected to be any of metaphoric–metaphoric, metaphoric –literal, literal–metaphoric, or literal–literal. For each passage, subjects had to judge the truth of two probe sentences (one designed to be true, the other designed to be false) on a scale from 1 to 5 (1 — true, 5 — false). The order of the passages was randomized for each booklet. From the 28 passages used in the experiment, two had "true" probe sentences that were given a rating of 2.6 or higher and six had "false" probes that scored 2.8 or less. These probes were modified to be more clearly true or false.

Procedure. The participants saw the materials on the screen of a Macintosh computer. They used one of two keys (K for true, D for false) and the ENTER key to express their answers. Participants were instructed to answer as fast and as accurately as possible. They were also warned that they may encounter words used figuratively.

Each participant took about 20 minutes to complete the experiment. The experiment consisted of four practice trials and 28 experimental trials. A trial had three phases:

1. Participants read a short passage.

2. When finished, they pressed a key and the passage disappeared. Then the target sentence was presented in a "word by word" style: first, the subject (article plus noun) of the sentence appeared, then the participant pressed a key and the subject was replaced by the verb (in a simple tense, but sometimes containing a preposition — e.g., "give up"), and, finally, after another key press, the ending part (consisting possibly of one or more words) replaced the verb.

3. After they pressed another key, participants saw a probe sentence that they had to judge as true or false. They communicated their answer by pressing one of two keys and, after that, they received feedback (on a new screen) and continued to the next trial.

The selection of the target type (metaphoric–metaphoric, metaphoric – literal,

literal–metaphoric, or literal–literal) was random for each trial and subject. Also, the selection of the probe type (true or false) was done randomly for each trial and subject. The experiment consisted of 28 trials, ordered randomly. Further on, we use the term reading time of a sentence part as meaning the interval between the occurrence of the word(s) on the screen and the next key press.

The dependent variables used for this experiment were (a) the reading time for the subject of the target sentence (henceforth referred to as noun-reading time); (b) the reading time for the verb of the target sentence; (c) the reading time for the ending part of the target sentence; (d) the reading time for the target sentence (derived by summing noun, verb and ending reading times); (e) the accuracy for judging the probe sentences as true or false; and (f) the latency for making the truth judgments. The accuracy for judging the probe sentence can be regarded as a measure of correct metaphor comprehension, under the assumption that a correct answer depended on correct metaphor comprehension[2]. Judgment latencies may also reflect how easily the target was understood. The independent variables were: (a) the type of noun (metaphoric or literal) used in the target sentence; (b) the type of verb (metaphoric or literal) used in the target; and (c) truth of the probe sentence (true or false).

Out-of-context pretest.

Given the frequency difference and possible other differences between metaphoric and literal items, we designed a pretest experiment in which all targets were read out of context. The out-of-context reading-times were then used as covariates for the analysis of reading times in context.

For the out-of context pretest, fifty-five native English speakers, undergraduates of Carnegie Mellon University[3], read the target sentences from Experiment 1 in the "word-by-word" manner described in the Procedure subsection and generated a causal completion for them. For instance, subjects read the different parts of the sentence "The

hens clucked noisily" and, then, on a separate screen, they had to generate a completion for that sentence. The completion always started with the word "because". An example of an acceptable completed sentence is "The hens clucked noisily because they liked the food".

We introduced the sentence completion task to make sure that subjects process the target sentences correctly and not press keys without actually comprehending them. Each participant fulfilled 56 trials: for each metaphor pair, two complementary targets were shown (e.g., either "The hens talked noisily" and "The women clucked noisily" or "The women talked noisily" and "The hens clucked noisily"). The first 28 trials contained no two targets pertaining to the same metaphor pairs, nor did the last 28 trials. Apart from this constraint, the order of presentation of targets was randomized for each subject.

We measured the reading times (i.e., the times between the onset of the word(s) on the screen and the subjects' key press) for noun, verb, ending, and overall sentence and used these measures as covariates in the analysis of noun-reading time, verb-reading time, ending-reading time and sentence-reading time, respectively, for Experiment 1. The reading times out of context are depicted in Table 3, together with the corresponding reading times from Experiment 1.

Results

The main results are given in Tables 3 (reading times) and 4 (accuracy measures and judgment times). In what follows we report statistics over both subjects (F_1 statistics) and items (F_2 statistics). We eliminated those trials that were more than 1.5 times the inter-quartile range from the upper and lower quartiles of the distribution, and we performed analyses of variance for each dependent measure. For noun-, verb-, ending-, and sentence-reading times, we also performed analyses of covariance (F_c statistics) over items with the corresponding out-of-context–reading times as covariate variables. Note

that the covariate differs for each dependent measure analyzed: for instance, for verb-reading times, F_c statistics is carried with out-of-context verb-reading time as a covariate. All the analyses take into consideration only the trials for which the answer to the final probe sentence was correct. Table 3 presents average reading times corrected for the corresponding out-of-context reading time. However, these corrected times are not substantially different from the uncorrected means.

Noun Reading Times. The ANOVA and the covariance analysis with noun as a factor[4] yielded a main effect of noun ($F_1(1,75) = 11.575$, $MSE_1 = 10331.229$, $p_1 < .001$, $F_2(1,25) = 8.896$, $MSE_2 = 7084.228$, $p_2 < 0.01$, $F_c(1,53) = 9.257$, $p_c < 0.005$), with the metaphoric nouns taking longer to be read than the literal nouns. The disadvantage of metaphoric nouns suggests that subjects paid a price in looking for a referent for the metaphoric noun.

Verb Reading Times. The ANOVA and the analysis of covariance with noun and verb as factors indicated a main effect of noun ($F_1(1,74) = 16.23$, $MSE_1 = 9411.635$, $p_1 < 0.001$, $F_2(1,26) = 6.992$, $MSE_2 = 7246.858$, $p_2 < .05$, $F_c(1,104) = 10.137$, $p_c < 0.005$): the verbs preceded by metaphoric nouns took longer to read than those preceded by literal nouns. Neither the main effect of verb or the interaction between verb and noun were significant. The noun effect on verb reading times possibly indicates a spillover of the processing load from the metaphoric noun to the following verb.

Ending Reading Times. The ANOVA with noun and verb as factors yielded again a significant main effect of noun over subjects, but this effect was not significant over items or in the covariance analysis ($F_1(1,72) = 8.273$, $MSE_1 = 42779.884$, $p_1 = 0.005$, $F_2(1,22) = 1.003$, $MSE_2 = 24158.412$, $F_c(1,99) = .411$). The verb effect and the interaction were nonsignificant. Surprisingly, the noun effect in the subject analysis was in the opposite direction of the previous effects: the ending parts of the sentences starting

with metaphoric nouns were read faster than the endings of sentences starting with literal nouns.

Sentence Reading Times. The sentence reading times were obtained by summing the noun reading times, the verb reading times and the ending part reading times. The ANOVA and the analysis of covariance with noun and verb as factors yielded no significant effects or interactions, although the metaphoric-noun sentences had a tendency to be slower.

Accuracy. The ANOVA for accuracy with noun, verb and truth as factors yielded a significant main effect of noun ($F_1(1,76) = 7.697$, $MSE_1 = 0.0393$, $p_1 < 0.01$, $F_2(1,24) = 6.319$, $MSE_2 = 0.01296$, $p_2 < 0.05$): subjects were more accurate for target sentences starting with literal rather than metaphoric nouns. Other interactions were significant only in the item analysis: noun and truth ($F_1(1,76) = 1.170$, $MSE_1 = 0.04234$, $F_2(1,24) = 5.526$, $MSE_2 = 0.0162$, $p_2 < 0.05$), verb and truth ($F_1(1,76) = 2.86$, $MSE_1 = 0.04648$, $F_2(1,24) = 4.812$, $MSE_2 = 0.007161$, $p_2 < 0.05$) and noun, verb and truth ($F_1(1,76) = 3.909$, $MSE_1 = 0.04616$, $F_2(1,24) = 5.195$, $MSE_2 = 0.01504$, $p_2 < 0.05$).

The accuracies in this experiment were quite high, suggesting that subjects comprehended correctly the target sentences most of the time. However, another possibility is that subjects might have been able to judge the probe sentences even if they had not read the targets. We approach this issue in more depth in Experiment 2. In any case, the noun effect seems to indicate a comprehension deficit for metaphoric nouns.

We also computed sensitivity (d') and bias(β) measures from the hits for true sentences and the false alarms for foils. The d' values reflect the subjects' ability to discriminate between true and false probes. Bias values greater than 1 indicate subjects' tendency to guess that a probe is false. Table 4 shows the d' and β values for the four

conditions. An ANOVA analysis run on d's obtained from each subject resulted in a significant effect of noun ($F_1(1,63) = 13.816$, $MSE_1 = .545$, $p_1 < .0001$) and of verb ($F_1(1,63) = 4.413$, $MSE_1 = .670$, $p_1 < 0.05$), with metaphoric nouns or verbs yielding lower discriminability than their literal counterparts. The ANOVA on the bias measures β yielded a significant interaction between noun and verb ($F_1(1,63) = 5.481$, $MSE_1 = .941$, $p_1 < 0.05$), with the congruent (metaphoric–metaphoric and literal–literal) sentences having a higher bias than the other sentences.

Judgment Latencies. An ANOVA with noun, verb and truth as factors found a significant main effect of truth ($F_1(1,75) = 41.609$, $MSE_1 = 406102.99$, $p_1 < 0.001$, $F_2(1,27) = 9.734$, $MSE_2 = 349273.73$, $p_2 < 0.005$). The other main effects and interactions were not significant. Subjects were faster to judge true sentences than to judge false sentences.

Discussion

This experiment successfully replicated Ortony et al.'s (1978) findings that metaphoric sentences with literal meaning are read as fast as literal sentences. However, it failed to reproduce Gibbs (1990) result that sentences that have no literal meaning and contain anaphoric metaphors are harder than literal sentences. Like Janus and Bever (1985), we found that metaphoric nouns were read more slowly than literal nouns. Also, we found a comprehension deficit for metaphoric sentences, but only if they contained a metaphoric noun. The experiment resulted in two unexpected effects of the metaphoric nouns: first, they led to increased reading time of the subsequent verb; second, the ending parts were read faster for metaphoric-noun targets than they were in the case of literal-noun targets (at least as indicated by the subject analysis). These results, together with the poorer comprehension accuracy for metaphoric-noun targets, indicate that anaphoric sentences with metaphoric nouns do pose an extra load on comprehension.

The pattern of reading times in Experiment 1 matches the following model:

1. When the noun is read, subjects search for a discourse antecedent for it; in the case of metaphoric nouns, it is harder to find an antecedent, so the search process either is more complex or fails; thus, it takes longer. That is to say that sometimes the metaphor is not understood at this point.

2. When the verb is read, if the preceding noun was metaphorical and if no antecedent was found, subjects may still continue to search for a referent. Moreover, the verb (or part of its semantic features) may be used as a context hook for the sentence.

3. When the ending is processed, if no antecedent for the sentence was found in the discourse, subjects may insist in finding one using the discourse-related information (if any) in the ending. If they fail (and they can do so especially in the case of metaphoric-noun sentences, in which given information is low), they can skip the processes that integrate the current sentence with the preceding text. Thus, shorter ending times reflect lack of integration or poor integration with the preceding passage. Because the metaphoric-noun targets have shorter ending-reading times, we surmise that those sentences are poorly connected (if at all) with the discourse and, thus, less often understood in relation with the preceding passage (as indicated by the accuracy data).

One piece of evidence for the idea that ending-reading times reflect integration with the preceding passage comes from a post-hoc analysis of endings. Indeed, we noticed that our endings varied in whether they were given or new information. For instance, the ending "noisily" of the metaphoric sentence "The hens clucked noisily" after a passage about a women's meeting would not throw much light upon the understanding of the sentence in context (we call such an ending "unrelated" to the passage). However, after a passage about a dinner at a restaurant , the ending of a sentence such as "The snail crawled to the table" may help participants relate the metaphoric sentence to the restaurant scene, and thus comprehend that "snail" refers to the waiter (we call this type

of ending "related" to the passage). We hypothesized that, in the case of metaphoric sentences, the ending-reading times may be longer for related endings than for unrelated ones, because without a related ending participants may abandon any attempt to integrate the sentence with the preceding passage. To verify this assertion, we divided our passages into two groups: one in which the endings of the target sentences were related to the passage and the other in which they were not[5]. An ANOVA on ending reading times with noun, verb and relatedness as factors yielded a significant effect of relatedness ($F_1(1,81) = 4.855$, $MSE_1 = 33352.273$, $p_1 < 0.05$) and an interaction of noun by relatedness ($F_1(1,81) = 18.298$, $MSE_1 = 23905.481$, $p < 0.001$). In the item analysis, the effect of relatedness was not significant ($F_2(1,107) = 0.014$[6], $F_c(1,97) = 0.003$), but the interaction of noun by relatedness was reliable ($F_2(1,107) = 6.766$, $p_2 < 0.05$, $F_c(1,97) = 6.816$, $p_2 < 0.05$). Table 5 shows the averages of ending reading times for the related and unrelated conditions. Subjects are faster in reading the ending when the noun is metaphoric and the ending is not related to the meaning of the passage. This result gives support to our earlier hypothesis that subjects sometimes give up on trying to relate metaphoric sentences to the passage and read the endings quickly. Subjects are also almost as fast on the literal nouns in the related condition; it is possible that, for literal-noun sentences, a related-ending sentence is more easy to integrate with the preceding discourse, due to semantic redundancy.

We saw that our accuracy data is consistent with the hypothesis that comprehension of targets is poorer in the metaphoric noun condition: indeed, subjects are less accurate for metaphoric noun sentences than for other sentences. Whereas there are these differences, participants are still accurate (over 0.80) for metaphoric noun sentences. We hypothesize that much of their accuracy may reflect an ability to judge the truth of the probe sentence without reading the target sentences. We test this assumption in Experiment 2: in that experiment, participants read the same passages as in Experiment 1

and they had to answer the same probe sentences, but without seeing the target sentences.

Experiment 2

Method

Participants. Forty one undergraduates at Carnegie Mellon University participated for a partial requirement in a Psychology course. They were all native English speakers, and they also participated in the pretest for Experiment 1. None of the subjects participated in Experiment 1.

Materials. We used the 28 passages from Experiment 1, together with the associated probe sentences, which could be true or false.

Procedure. As in Experiment 1, the participants saw the materials on the screen of a MacIntosh computer. For each subject, we randomized the order in which the the pretest for Experiment 1 and Experiment 2 were completed. Subjects took about 10 minutes, on average, to complete this experiment.

A trial was very similar to a trial in Experiment 1, with the exception of phase two (target reading), which was removed. Participants in this experiments did not see any target sentences. After each trial, feedback was given. The feedback corresponded to the correct answers in Experiment 1 (i.e., an answer was considered correct if it would have been correct in Experiment 1).

Each participant completed 28 trials, presented in a random order. The truth of the probe sentence was also selected randomly. The instructions specified the possibility that, for some trials, passages did not contain all the information necessary to answer the final probe sentence. However, participants were encouraged to do their best to find the correct answer.

For this task, we used two dependent measures: the accuracy (defined with respect

to the passages of Experiment 1) of the answers and the judgment latency. The only independent variable was the truth of the probe sentence.

Results

Accuracy. A paired sample t test yielded a significant difference between the accuracy for true probe sentences and for false probe sentences ($t_1(40) = 2.124$, $p_1 < 0.05$) by subjects, but not by items ($t_2(27) = 1.488$). Subjects tended to be more accurate for true probe sentences (.78 correct) than for false probe sentences (.69 correct). As for Experiment 1, we computed d' and β values. In this experiment, the average discriminability d' was 1.43 and the bias β was .87, reflecting a tendency of subjects to answer "true".

Accuracy improvement in Experiment 1 versus Experiment 2. The main purpose of this experiment was to build a baseline towards which to judge subjects' answers in Experiment 1. Therefore, for each item we define a new measure of accuracy (which we call item improvement) reflecting the relative improvement in Experiment 1 with respect to the overall possible improvement. Thus, for a probe sentence q, let a_{1q} be the accuracy in Experiment 1 for all trials containing that probe sentence. Then let a_2 be the average accuracy in Experiment 2 for probe sentences with the same truth as q. Then the improvement I_q for probe sentence q is defined as $I_q = \frac{a_{1q} - a_2}{1 - a_2}$.

From the analyses that follow, we eliminated items that were too good (for which the average accuracy in Experiment 2 was over 0.90), because the possible improvement for such items was too small. For this reason, the score of 10 true probes and 10 false probes were not considered in the analyses[7]. ANOVA analyses on the improvement measure yielded an effect of noun on subjects, but not on items ($F_1(1, 76) = 5.973$, $MSE_1 = .547$, $p < 0.05$, $F_2(1, 27) = 1.964$, $MSE_2 = .253$). Table 4 presents the average improvement. Subjects show about 60 percent improvement on literal–literal sentences,

double of the 30 percent improvement on the metaphor–metaphor targets. This measured improvement is consistent with the estimate we get from the d' analysis. Subjects in the literal–literal condition increased their d' by 1.09 from the baseline whereas subjects in the metaphoric–metaphoric condition increased their d' by only 0.53.

The relatively high accuracy in Experiment 2 showed that subjects could answer our probe sentences correctly using plausibility strategies and supported the hypothesis that such strategies might have been used by subjects in Experiment 1. The improvement measure showed that subjects were only half as successful in comprehending metaphoric–metaphoric targets as literal–literal targets. Thus the comprehension deficit, corrected for plausibility strategies, was considerable for metaphors in Experiment 1.

Conclusions

We had entered this study believing that the difference between the Gibbs' (1990) and Ortony et al.'s (1978) results might have to do with whether both the noun and verb were metaphors. Gibbs' sentences were like our metaphoric-literal sentences and Ortony et al.'s were like our metaphoric-metaphoric sentences. We found little evidence for a difference between these two types of sentences. Instead, our participants seem to have allocated some sort of fixed processing time for the sentences. If they used too much time processing the beginning, they made up for it by speeding up reading of the end and paid a price in comprehension. Neither Gibbs (1990) nor Ortony et al. (1978) report a measure of comprehension differences for literals versus metaphors and so we cannot judge this issue for those studies. One can speculate that perhaps Gibbs' participants chose to spend more time to achieve higher levels of comprehension. Thus participants may make a speed–accuracy choice: for anaphoric metaphors, they either pay a price in comprehension time or comprehension success.

In conclusion, it seems that our results agree with Gibbs' (1990) and Onishi and

Murphy's (1993) claim that anaphoric metaphors are more difficult than literals and, indirectly, than predicative metaphors: however, in their studies that difficulty was reflected in reading times, whereas in ours, it is indicated by the accuracy measures. A methodological implication of this result is that sentence-reading time is not always a sufficient comprehension test in the case of metaphoric sentences.

We started this article with an evocation of Searle's (1979) theory of metaphor processing. As Janus and Bever's (1985) did, our study shows that for deciding pro or against it, sentence-reading–time granularity may be too coarse. Although at the time when it was articulated researchers seemed to believe that either this theory or the opposite must be true for metaphor comprehension, nowadays the picture becomes more nuanced. Thus, Blasko and Connine (1993) found that the familiarity of a metaphor plays a role in whether it is comprehended easily or not and for unfamiliar metaphors, goodness may facilitate understanding. Giora (1997) proposed the graded-salience hypothesis, according to which "salient meanings (e.g., conventional, frequent, familiar, enhanced by prior context) are processed first." Keysar (1994) showed that the context can influence whether the interpretation chosen for a sentence is literal or metaphoric. Our results indicate also that context plays a role in understanding metaphoric sentences: the amount of overlapping information between the discourse and sentence level modulates the comprehension of the target — the more given information, the better the integration with the preceding passage and the more accurate the final interpretation. Thus, for poor sentence context, metaphoric interpretation may be actually never computed. Indeed, we think that sentence context may be one of the keys for unifying the contradictory literature of metaphor-reading times: if the sentence context is facilitatory enough, then it will lead to a good understanding of the metaphor; moreover, if it precedes the metaphor in the sentence (as it does in the case of predicative metaphors) and if it is supportive enough, the context can make the metaphor understanding as smooth as literal understanding. In

this respect, we agree with Giora (1997) that the distinction between figurative and literal is an artificial one and should be replaced by "the continuum salient–nonsalient".

References

Blasko, D., & Connine, C. (1993). Effects of familiarity and aptness on metaphor processing. Journal of Experimental Psychology: Learning, Memory and Cognition, 19, 295-308.

Francis, W., & Kucera, H. (1982). Frequency analysis of english usage : lexicon and grammar. Boston: Houghton Mifflin.

Gerrig, R., & Healy, A. (1983). Dual processes in metaphor understanding: Comprehension and appreciation. Journal of Experimental Psychology: Memory and Cognition, 9, 667–675.

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

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. Cognitive Linguistics, 8, 183-206.

Glucksberg, S., Glidea, P., & Bookin, H. (1982). On understanding literal speech: Can people ignore metaphors? Journal of Verbal Learning and Verbal Behavior, 21, 85–98.

Goldvarg, Y., & Glucksberg, S. (1998). Conceptual combinations: The role of similarity. Metaphor and Symbol.

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

Janus, R., & Bever, T. (1985). Processing of metaphoric language: An investigation of the three-stage model of metaphor comprehension. Journal of Psycholinguistic Research, 14, 473-487.

Keysar, B. (1989). On the functional equivalence of literal and metaphorical interpretations in discourse. Journal of Memory and Language, 28, 375-385.

Keysar, B. (1994). Discourse context effects: Metaphorical and literal interpretations. Discourse Processes, 18, 247-269.

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

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

Ortony, A., Vondruska, R., Foss, M. A., & Jones, L. (1985). Salience, similes and the asymmetry of similarity. Journal of Memory and Language, 24, 569-594.

Searle, J. (1979). Metaphor. In A. Ortony (Ed.), Metaphor and thought. Cambridge University Press.

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

Tourangeau, R., & Rips, L. (1991). Interpreting and evaluating metaphors. Journal of Memory and Language, 30, 452-472.

Tourangeau, R., & Sternberg, R. (1981). Aptness in metaphor. Cognitive Psychology, 13, 27-55.

Author Note

This research was supported by grant 9975220 from the National Science Foundation. We thank Glenn Gunzelmann and Mihai Budiu for commenting on earlier drafts of this manuscript.

Footnotes

[1]Initially we considered more than 28 metaphor pairs; based on this rating study we rejected some of our initial choices.

[2]We test this assumption in Experiment 2.

[3]Forty one of the pretest subjects participated also in Experiment 2.

[4]The verb or the truth of the final probe sentence could not influence the reading time for the noun, as the noun preceded the verb and the probe sentence.

[5]The two authors rated the materials independently and initially agreed in 27 out of 28 cases. The remaining case was determined after discussion. There were 14 items in each of the two groups.

[6]The item analysis was not a repeated analysis.

[7]The stories with high true probe score were not the same with the stories with high false probe score. Therefore, these items could not be eliminated completely, but rather their scores in the true respectively false condition were treated as missing values for that condition.

Table 1

Sample Passages, Targets and Probes from Experiment 1.

EXAMPLE 1:

Mary was taking a cooking course. She wanted to show her husband what a good cook she had become, so she decided to prepare the turkey recipe they were teaching at school. She followed the recipe down to the last detail. However, when her husband tasted the turkey, he found it too tough to cut or eat. He threw Mary a scornful look and said: "You'd better spend money on something more useful than cooking school."

TARGET SENTENCES:

The dagger cut deeply.(METAPHORIC–METAPHORIC)

The dagger offended deeply. (METAPHORIC–LITERAL)

The insult cut deeply. (LITERAL–METAPHORIC)

The insult offended deeply. (LITERAL–LITERAL)

PROBE SENTENCES:

The insult hurt Mary. (TRUE)

The insult made Mary ambitious. (FALSE)

EXAMPLE 2:

When Cinderella arrived at the ball, she was wearing a wonderful white dress. Her fine long neck was adorned by an exquisite diamond necklace. Many handsome men wanted to invite her to dance. She was a little bit anxious, because she had never danced before.

TARGET SENTENCES:

The swan floated beautifully. (METAPHORIC–METAPHORIC)

The swan waltzed beautifully. (METAPHORIC–LITERAL)

The girl floated beautifully. (LITERAL–METAPHORIC)

The girl waltzed beautifully. (LITERAL–LITERAL)

PROBE SENTENCES:

Cinderella danced with grace. (TRUE)

Cinderella talked with grace. (FALSE)

Table 2

Average Goodness and Familiarity Ratings for Metaphors Used in Experiment 1, Compared with Corresponding Average Ratings of Good and Nonsensical Metaphors

	Familiarity	Goodness
Experiment 1	2.18	2.54
Noun Metaphors	2.16	2.65
Verb Metaphors	2.20	2.45
Good Metaphors	2.75	2.72
Nonsense Metaphors	1.40	1.73

Note. Scale: 1 = lowest; 4 = highest.

Table 3

Average Reading Times in Context (Experiment 1) and Out of Context (Pretest to Experiment 1)

	In Context				Out of Context			
	Met–Met	Met–Lit	Lit–Met	Lit–Lit	Met–Met	Met–Lit	Lit–Met	Lit–Lit
Noun RT	672		634		698		707	
Verb RT	567	562	538	529	645	638	672	631
Ending RT	787	769	805	783	855	981	1100	849
Sentence RT	2031	2036	1975	1978	2266	2348	2513	2191

Note. All reading times in context (Experiment 1) are corrected for the corresponding out-of-context reading time. All reading times are given in milliseconds. Met = metaphoric, Lit = literal, RT = reading time.

Table 4

Average Accuracy, d', β, Accuracy Improvement, and Judgment Time in Experiment 1

	Met–Met	Met–Lit	Lit–Met	Lit–Lit
Accuracy				
True Probes	0.84	0.88	0.90	0.86
False Probes	0.79	0.82	0.82	0.91
d'	1.96	2.32	2.45	2.52
β	1.35	1.15	1.04	1.40
Accuracy Improvement				
True Probes	0.28	0.452	0.557	0.378
False Probes	0.332	0.428	0.43	0.71
Judgment time				
True Probes	2237	2179	2096	2368
False Probes	2526	2531	2570	2588

Note. The accuracy improvement in Experiment 1 was computed based on the normative data from Experiment 2. Judgment times are given in milliseconds. Met = metaphoric, Lit = literal, RT = reading time.

Table 5

Average Ending Reading Times (ms) for Endings Related or not to the Passage in Experiment 1

Context	Met Noun	Lit Noun
Unrelated	748	826
Related	811	761

Note. The numbers were corrected for ending-reading times out of context. Met = metaphoric, Lit = literal.