

# Negation in Nonliteral Sentences

**Raluca Budiu (budiu@parc.com)**

Palo Alto Research Center,  
3333 Coyote Hill Rd., Palo Alto, CA 94304, USA

**John R. Anderson (ja@cmu.edu)**

Carnegie Mellon University,  
5000 Forbes Ave., Pittsburgh, PA 15213, USA

## Abstract

We investigate how people process negation in semantically distorted and metaphoric sentences. We present three experiments in which participants judged the truth of affirmative and negative sentences that were either literal or contained semantic illusions (Erickson & Mattson, 1981) or metaphors. In all experiments, negation increases processing times; although for semantic illusions, negation preserves the ordering of judgment times for literal and nonliteral sentences, for metaphors this ordering is reversed, with nonliteral negatives taking less time than literal negatives. This result presents evidence against the traditional Clark and Chase (1972) model of negation. We propose the negation-as distortion model and discuss how this model can explain the data.

**Keywords:** negation; Moses illusion; metaphor; sentence processing; computational modeling; INP; ACT-R

## Introduction

Negation markers (“no” and “not”) are a pervasive feature of natural language that has been only modestly addressed in the psycholinguistic literature. The classic attempt of a process model of negation belongs to Clark and Chase (1972), who assume that negation acts as a post-processing operator: the sentence as an affirmative is processed first, and then the negation is applied. Other studies (Kaup, 2001; MacDonald & Just, 1989; Lea & Mulligan, 2002) found that sometimes negation reduces the availability of the negated meaning. Giora, Balaban, Fein, and Alkabets (2004) propose the retention hypothesis for negation; according to this model a “negation marker does not suppress salient meanings activated initially but only modifies them.” In this paper we investigate yet another possibility – namely that negation acts as a semantic distortion, lowering the similarity of the overall (negative) sentence to other items in memory.

For a long time now, negation has been associated with similarity. In his 1965 paper, Wason has proposed the exceptionality hypothesis, which asserts that negative sentences are processed more easily if they state an exception to a rule than if they state “obvious”, implausible information. For instance, if it is known that Bill drives to work everyday, *Bill did not drive to work today* should be understood faster than *Bill did not walk to work today*. In Wason (1965)’s original study, participants were shown display cards with eight circles; seven circles were of one color (e.g., red) and another

was different (e.g., blue). After they saw the picture and described it verbally, participants had to fill in sentences of the type *Circle No.3 is ...* and *Circle No.3 is not ...*. In support of his exceptionality hypothesis, Wason found that, subjects needed less time to fill in exceptional negatives (e.g., *Circle No.3 is not red*) than the implausible negatives (e.g., *Circle No.3 is not blue*). De Villiers and Tager Flusberg (1975) performed a related experiment with young children, and also looked at “confusability”: how far away the discrepant item was from the similar items. They found that the more similar the discrepant items was to the other items, the easier was for the children to fill in the plausible negative (e.g., a “six horses and one cow” context was easier as opposed to a “seven cars and one baby’s bottle” context). Valle Arroyo (1982) also manipulated similarity and confusability and showed empirical support for the exceptionality hypothesis, without reaching a conclusive result about confusability.

We use a different paradigm to study how similarity affects negation: given a fact already existent in long term memory, we distort it (in varying degrees) and examine how this distortion affects the processing of negatives. In other words, we further explore the issue of confusability by manipulating language similarity. Previous experiments with semantic distortions in affirmative sentences have shown that people often tend to ignore them and extract the gist from the sentence. Erickson and Mattson (1981) were the first to document the so-called Moses illusion: when people were asked questions such as *How many animals of each kind did Moses take on the ark*, in about 80% of cases, they did not notice that the sentence incorrectly referred to *Moses* instead of *Noah*. Metaphors (and, more generally, nonliteral language) can also be thought of as another easy-to-process instance of semantic distortions. In fact, Budiu and Anderson (2004) have shown that both Moses illusion and metaphors can be explained through a unique sentence processing mechanism that involves similarity-based matching to concepts already existent in memory. In their INP (Interpretation-Based Processing) sentence-processing model, the input sentence is matched against other facts in memory and the proposition in memory that is most similar to the input is considered the current interpretation of the sentence. INP is able to find interpretations for nonliteral sentences (e.g., *The sky was filled with drops of molten silver*) by assuming a parallel

spreading activation model in which activation is spread from the words in the sentence to propositions in memory (e.g., *The sky is filled with stars*) that contain words semantically similar to them.

Coming back to negation, we propose that negation acts as a distortion. That is, (*NOT A*) is like a new concept that is highly similar to A, but not quite A. Then the comprehension of a negative sentence should resemble very much to that of a nonliteral sentence, in which a nonliteral item (e.g., *drops of molten silver* in *Drops of molten silver filled the night sky*) would have a lower semantic similarity to the “correct” interpretation (i.e., *The sky is filled with stars*) than the literal itself (e.g., *stars* or *bright celestial bodies*). However, for such a pure similarity-based model to work without confusing negatives and affirmatives all the time, we would need to augment it with a Clark and Chase (1972) type of tag model, in which a negation marker is also kept explicitly in the sentence representation.

Further, we present three experiments that were designed with this model in mind. (In the rest of the paper we refer to this model as negation-as-distortion model.) In particular, we are interested in exploring whether the negation as distortion assumption, combined with INP’s mechanism of sentence processing, is necessary, or a simple negation-tag model a la Clark and Chase (1972) is sufficient. We manipulate the level of distortion of sentences using nonliteral language (either semantic distortions or metaphors) and negation. If the Clark and Chase model is sufficient and negation does not act as a distortion, then we expect that the comprehension time orderings for literal and nonliteral affirmative sentences should be preserved in the negative. The negation-as-distortion model predicts that negation may actually interact with the level of nonliterality, with negated nonliterals having a possibly too low similarity to the concept being referred. The low levels of similarity may actually translate in poor comprehension of the negated nonliteral sentences and in faster response times<sup>1</sup>.

We study two kinds of nonliteral sentences: sentences that contain semantic distortions in Experiment 1 and metaphoric sentences in Experiments 2 and 3. The metaphoric sentences that we used contain predicative metaphors (e.g., *Jobs are jails*), in which the metaphor occurs as part of the sentence predicate. Predicative metaphors are fairly well studied and the general agreement is that they are the “easiest” of all metaphors.

## Experiment 1

Experiment 1 studied how negation affects literal sentences and sentences containing semantic distortions. Whereas Moses illusion is a very robust phenomenon and people have difficulty noticing distortions when the sentences contain “well-chosen” semantic distortions (Reder

---

<sup>1</sup>Budiu & Anderson, 2004, showed, in the context of metaphoric sentences, that if the similarity of the distorted sentence to the referring fact in memory is too low, then the process of sentence comprehension is truncated (no links have to be set to the background knowledge) and, thus, the response is fast, at the expense of an impoverished product of comprehension.

& Kusbit, 1991), not all distortions are equally good. Ayers et al. (1995) have shown that people fall less often for bad distortions (e.g., *How many animals of each kind did Adam take on the ark?*) than for good distortions. Ayers et al. conducted an informal rating of the “good” and “bad” distortions and established that the good distortion (*Moses*) shared more features with the undistorted term (*Noah*) than the bad distortion (*Adam*). Budiu and Anderson (2004) showed that a key factor in explaining Moses illusion is the degree of literalness of the distortion: the higher the similarity between the distortion and the undistorted term, the more likely the illusion is to work. (This assumption is confirmed by ratings conducted by van Oostendorp & de Mul, 1990; van Oostendorp & Kok, 1990.)

Since we were interested in how negation interacts with semantic similarity, we used two types of nonliteral sentences: high similarity sentences (i.e., sentences containing good distortions) and low similarity sentences (i.e., sentences containing bad distortions). The sentences were either affirmative or negative; participants had to judge the truth of these sentences.

**Participants.** Forty CMU students were given class credit for their participation. All were native speakers of English raised in United States or Canada.

**Design and materials.** The experiment had a 3x2 design, crossing the level of distortion (no distortion, good distortion, bad distortion) with the verb form (affirmative or negative). We designed 72 sets of sentences, having six different versions, as in (1):

(1).1 Affirmative sentences:

- a. *The ancient temple of Parthenon was built in Athens by the Greek.* (no distortion)
- b. *The ancient temple of Parthenon was built in Athens by the Romans.* (good distortion)
- c. *The ancient temple of Parthenon was built in Athens by the Egyptians.* (bad distortion)

(1).2 Negative sentences:

- a. *The ancient temple of Parthenon was not built in Athens by the Greek.* (no distortion)
- b. *The ancient temple of Parthenon was not built in Athens by the Romans.* (good distortion)
- c. *The ancient temple of Parthenon was not built in Athens by the Egyptians.* (bad distortion)

Most of the materials (sixty four of the 72 sets of sentences) have been used before in Moses illusion experiments (Ayers, Reder & Anderson, 1995; Reder & Kusbit, 1991). Moreover, participants in these previous experiments fell more often for good than for bad distortions.

For each set of sentences, we also used a question variant (2) to establish whether the participants had correct knowledge about the fact stated by that sentence:

(2) *Who built the ancient temple of Parthenon in Athens?*

**Procedure.** The task was self paced and run on a PC. For each trial participants read one sentence corresponding to one of the variants in (1) and then had judge whether that sentence was true or false. Truth judgments were done by pressing one of the keys “D” or “K”; then the participants were shown a feedback message (right, wrong, too long to answer) and the next trial started. We recorded the participants’ responses and response times (RTs). At the end of the experiment, participants had to fill in a questionnaire that contained the question versions (2) of the sentences that they had seen during the experiment.

Each participant saw one sentence from each of the 72 sets of sentences (1), that was randomly selected. All participants was presented with 12 trials in each of the six conditions. The order in which the sentences were presented was randomized.

At the end of the experiment each participant had to fill in a pencil-and-paper questionnaire that contained versions (2) of the material sets. The order of questions in the questionnaire was randomized for each participant.

## Results and discussion

The results for Experiment 1 are shown in Table 1. The response times reported throughout the paper are for correct responses only. For each participant, we took into account only the trials for which, according to the final questionnaire, the participant knew the correct fact (e.g., that it was the Greek who built the Pantheon)<sup>2</sup>. We also eliminated 7 subjects whose average accuracies were below 60% and who may have given random answers.

As expected, the negation put an extra processing burden: the participants were significantly less accurate ( $F_1(1, 32) = 12.46$ ;  $MS = 0, 15$ ;  $p = 0.001$ ;  $F_2(1, 69) = 5.48$ ;  $MS = 0.27$ ;  $p < 0.05$ ) and slower ( $F_1(1, 32) = 52.482$ ;  $MS = 21201861$ ;  $p < 0.001$ ;  $F_2(1, 69) = 79.20$ ;  $MS = 56885916$ ;  $p < 0.001$ ) for negative sentences. Participants made more errors for distorted sentences (they “fell” for the illusion) than for undistorted ones. The effect of distortion was significant for accuracy:  $F_1(2, 64) = 68.53$ ;  $MS = 1.43$ ;  $p < 0.001$ ;  $F_2(2, 139) = 44.29$ ;  $MS = 3.13$ ;  $p < 0.001$ ), but participants did not take necessarily longer to make judgments of distorted sentences<sup>3</sup> (the distortion effect on RTs was significant in the item analysis only —  $F_1(2, 64) = 1.79$ ;  $p = 0.17$ ;  $F_2(2, 139) = 3.66$ ;  $MS = 2920360$ ;  $p < 0.05$ ). The interaction between negation and distortion was not significant for RTs or accuracies (except for the subject analysis for accuracies  $F_1(2, 64) = 52.48$ ;  $MS = 21201861$ ;  $p < 0.001$ ;  $F_2(2, 128) = 1.05$ ;  $p > 0.1$ ). We also ran planned comparisons between the two levels of distortion both in the negative and in the affirmative,

<sup>2</sup>As one reviewer pointed out, in the case of literal sentences the feedback may have provided a learning experience for the participants. Thus, it is possible that in the final questionnaire we actually overestimate the a priori knowledge of the subjects and hence the accuracies in Table 1 may represent underestimates in the literal case.

<sup>3</sup>This result is consistent with other studies (Reder & Kusbit, 1991) that have found no effect of distortion on latency.

Table 1: RTs and accuracies for Experiment 1.

	RTs		Accuracies	
	Aff	Neg	Aff	Neg
No distortion	3879.61	4510.47	0.89	0.87
Good distortion	3991.08	4778.14	0.63	0.59
Bad distortion	4067.69	4613.17	0.69	0.60

but we did not find any significant differences either in RTs or accuracies. The lack of interaction for RTs is in agreement with the Clark& Chase model of negation: since negation is an operator applied to the affirmative proposition, there is no reason for sentences to behave differently in the negative than in the affirmative (except for a negation cost that presumably is the same for all sentences). However, the result is not necessarily contradictory with the negation-as-distortion model — it is possible that, although negation may have contributed to some degree to the level of distortion of the sentence, that level never got low enough to produce observable effects, due to the length of the experimental sentences (and, thus, to the amount of contextual information available). This kind of effect has been noticed before in gist versions of Moses-illusion (Reder & Kusbit, 1991; Ayers et al., 1995), where participants were able to correctly identify the fact that the sentence referred to, in spite of the sentences containing bad distortions. Thus, it is possible that this experiment did not succeed in differentiating between the simple Chase and Clark model and the negation-as-distortion model.

## Experiment 2

This experiment was very similar in design to Experiment 1, but used shorter metaphoric sentences of the form “Some As are Bs”. There were two types of nonliteral sentences: metaphoric (analogous to the good distortions in Experiment 1) and nonsensical (analogous to the bad distortions in Experiment 1).

**Participants.** Twenty two CMU students participated in the experiment for course credit. They were all native English speakers.

**Materials.** We used a 3x2 design in which the metaphoricity level (literal, good metaphor, nonsense metaphor) was crossed with negation (affirmative versus negative). There were 72 sets of “Some As are Bs” sentences, each having 6 different versions, as in (3):

(3).1. Affirmative sentences:

- a. *Some apartments are penthouses.* (literal)
- b. *Some apartments are palaces.* (metaphor)
- c. *Some apartments are angels.* (nonsense metaphor)

(3).2. Negative sentences:

- a. *No apartments are penthouses.* (literal)
- b. *No apartments are palaces.* (metaphor)

c. *No apartments are angels.* (nonsense metaphor)

To generate the nonsense metaphors, we scrambled “B” terms from other metaphors in our set (for instance, in the example (3), *angels* actually occurs in a different set as a metaphor — *Some babies are angels*).

**Procedure.** The procedure was identical with that in Experiment 1. We specifically instructed participants to judge the literal truth of the sentences. Instead of the final questionnaire, we had the subject rate the comprehensibility, metaphoric goodness, and familiarity of all affirmative nonliteral sentences (such as (3).1.b and (3).1.c). This rating was on a scale of 1 to 5 (1 lowest) and was administered using the same computer as for the experiment. The sentences were presented in random order. The ratings confirmed our choice of materials (the metaphors were significantly better than the nonsense on all dimensions) and we do not discuss them here.

## Results and discussion

The results are presented in Table 2. As in Experiment 1, negation significantly deteriorated both the RTs ( $F_1(1, 18) = 21.79$ ;  $MS = 3064353$ ;  $p < 0.001$ ;  $F_2(1, 69) = 23.58$ ;  $MS = 10904723$ ;  $p < 0.001$ ) and the accuracies ( $F_1(1, 18) = 13.68$ ;  $MS = 0.33$ ;  $p = 0.001$ ;  $F_2(1, 69) = 21.7783$ ;  $MS = 1.0468p < 0.001$ ). The effect of metaphoricity on the RTs was not significant ( $F_s < 1$ ); however, metaphoricity did affect the accuracies ( $F_1(2, 36) = 9.54$ ;  $MS = 0.15$ ;  $p < 0.001$ ;  $F_2(2, 139) = 4.97$ ;  $MS = 0.40$ ;  $p < 0.01$ ). For accuracies, there was a significant interaction of metaphoricity and negation ( $F_1(2, 36) = 2.64$ ;  $MS = 0.04$ ;  $p = 0.1$ ;  $F_2(2, 131) = 7.3226$ ;  $MS = 0.3454$ ;  $p < 0.001$ ): in the affirmative, participants made fewer errors for nonsense sentences than for metaphoric sentences, whereas in the negative, they were about the same. For RTs, there was a significant interaction between negation and metaphoricity ( $F_1(2, 36) = 4$ ;  $MS = 800715$ ;  $p < 0.05$ ;  $F_2(2, 118) = 4.94$ ;  $MS = 2512652$ ;  $p < 0.01$ ), with nonliteral sentences taking longer than literal sentences in the affirmative, but taking less in the negative. This latter effect contradicts the Clark & Chase model: if negation were just an operator applied to the positive proposition, the ordering of RTs in different conditions would be the same in the negative and affirmative. The reversal of RT ordering in the negative is consistent with the negation-as-distortion model. According to this model, negation lowers the similarity of the items to the corresponding concepts in memory; if the overall similarity of the sentence to any proposition in memory becomes too low (which may be the case for the negative metaphoric and nonsense sentences) the sentence is not linked to any known facts in memory and other processing is short-circuited, leading to shorter response times.

We also ran separate planned comparisons between the two nonliteral levels. We did not find any RT differences between the nonsense sentences and the metaphoric sentences; we did find a significant difference between the accuracies of affirmative metaphoric sentences and those of affirmative scrambled sentences ( $t_1(18) = 2.78$ ;  $p_1 = 0.01$ ;  $t_2(67) = 2.59$ ;  $p_2 = 0.01$ ), but

Table 2: Accuracies and RTs for Experiment 2.

	RTs		Accuracies	
	Aff	Neg	Aff	Neg
Literal	2338.32	2967.91	0.92	0.87
Metaphoric	2655.51	2705.99	0.82	0.72
Nonsense	2562.13	2865.77	0.90	0.73

this difference was not preserved in the negative. This result is consistent with the negation-as-distortion hypothesis: people were more likely to respond “true” to the false affirmative metaphor because the similarity of the metaphoric sentence to a proposition in memory was high enough that sometimes they could not notice any distortion. This was not true of the nonsense sentence. However, in the negative, the negation lowered the similarity of the metaphoric sentence and made it more like the nonsensical sentence.

## Experiment 3

The quantifiers introduced in Experiment 2 may have put some extra problem-solving burden on the participants; it is not clear whether such a burden is typical for normal linguistic processing. Therefore, we repeated Experiment 2, but with a different set of materials that contained no quantifiers.

**Participants.** Thirty CMU students participated for course credit. They were all native English speakers raised in the United States or Canada.

**Materials.** Instead of quantifier metaphors, in this experiment we used sentences such as “A is (not) a B”, where A was a proper name (two words). There were 72 such sentences, each with 6 variants (see (4)):

(4).1 Affirmative sentences:

- a. *Cruella Devil was a villain.* (literal)
- b. *Cruella DeVil was an iceberg.* (metaphor)
- c. *Cruella DeVil was an octopus.* (scrambled metaphor)

(4).2 Negative sentences:

- a. *Cruella Devil was not a villain.* (literal)
- b. *Cruella DeVil was not an iceberg.* (metaphor)
- c. *Cruella DeVil was not an octopus.* (scrambled metaphor)

The first two words in the sentence always formed a famous name.

**Procedure.** The procedure was the same as for Experiment 2.

## Results and Discussion

The results (accuracies and response times — RTs) are presented in Table 3. As in Experiments 1 and 2, participants were significantly slower and less accurate for negative sentences (RTs:  $F_1(1, 29) = 68.26$ ;  $MS = 11323357$ ;  $p < 0.001$ ;  $F_2(1, 63) = 91.22$ ;  $MS = 28419697$ ;  $p < .001$ ;

Table 3: Accuracies and RTs for Experiment 3.

	RTs		Accuracies	
	Aff	Neg	Aff	Neg
Literal	2093.91	2841.69	0.84	0.87
Metaphoric	2411.61	2785.91	0.95	0.88
Nonsense	2385.54	2768.35	0.98	0.91

accuracies:  $F_1(1, 29) = 5.52$ ;  $MS = 0.06$ ;  $p < 0.05$ ;  $F_2(1, 63) = 7.48$ ,  $MS = 0.22$ ;  $p < 0.01$ ). Metaphoricity did not impact RTs ( $F_1(2, 58) = 1.91$ ;  $p > 0.1$ ;  $F_2 < 1$ ), but did affect accuracies ( $F_1(2, 58) = 9.20$ ;  $MS = 0.13$ ;  $p < 0.001$ ;  $F_2(2, 127) = 8.83$ ;  $MS = 0.38$ ;  $p < 0.001$ ). As in Experiment 2, we found a negation by metaphoricity interaction, both for RTs ( $F_1(2, 58) = 7.3557$ ;  $MS = 681884$ ;  $p = 0.001$ ;  $F_2(2, 122) = 2.98$ ;  $MS = 977303$ ;  $p = 0.05$ ) and accuracies ( $F_1(2, 58) = 5.61$ ;  $MS = 0.05$ ;  $p < 0.01$ ;  $F_2(2, 127) = 4.47$ ;  $MS = 0.09$ ;  $p = 0.01$ ). Again, we found that in the negative participants were faster for nonliteral sentences than for literals. As for Experiment 2, the metaphoricity by negation interaction contradicts the Clark and Chase model and brings support to the negation-as distortion model. Planned comparisons between metaphoric and nonsense sentences revealed that, as in Experiment 2, the accuracy difference between affirmative metaphoric and nonsense sentences was significant ( $t_1(25) = 2$ ;  $p_1 = 0.05$ ;  $t_2(63) = 2.06$ ;  $p_2 < 0.05$ ); however, this difference disappeared in the negative. No significant difference between the two types of non-literals was found for the RTs.

## General discussion

We presented three experiments in which we compared literal sentences with nonliteral sentences involving semantic distortions — either metaphors or semantic illusions. Negation clearly made the sentence processing more difficult, as reflected in the longer RTs and higher number of errors. The accuracies were generally worse in the nonliteral conditions than in the literal conditions, except for Experiment 3. With respect to confusability, the level of distortion (low versus high; metaphoric versus nonsense) did not affect the RTs in the affirmative or in the negative. However, in Experiments 2 and 3 participants were significantly more accurate for the affirmative nonsense sentences than for the affirmative metaphoric sentences; this difference disappeared in the negative.

Whereas for the Experiment 1, the RT data were consistent with the Chase and Clark model (the ordering of RTs being preserved in the negative), in Experiments 2 and 3 the RT ordering was reversed in the negative: participants were faster to judge negative nonliteral sentences than negative literals, although the opposite trend was present in the affirmative. We interpret this result in favor of the negation-as-distortion hypothesis: both negation and nonliterality lower the overall similarity of the sentence to any proposition in memory. In negative nonliteral sentences this similarity may become too

low, resulting in a failure to link these sentences to any known fact in memory. The lack of an interpretation in the background knowledge leads to more shallow, restricted processing and, thus, to shorter response times for these sentences.

Next, let us see how the negation-as-distortion hypothesis on top of the INP processing model can account for the pattern of errors. But before that, we need to review some of INP’s basic mechanisms. INP is an incremental model of sentence comprehension, built in ACT-R (Anderson & Lebiere, 1998). It performs syntactic and semantic processing of an input sentence and has been shown to match behavioral data from different psycholinguistic domains. One of the outputs of INP is an interpretation for the sentence; the interpretation is the proposition in the background knowledge that is most similar to the input sentence. Since INP is incremental, it actually searches for an interpretation as it reads words one by one; if at any point, it is not able to find a proposition in memory that is similar enough to the words already read, it creates a so-called “bug”. The bug essentially records a difficulty in comprehending that sentence. The bug is used in a number of ulterior judgments about the sentence, the most two important for our purposes being the truth judgment and the literality judgment. If the end of the sentence is reached with no candidate interpretation, INP may actually skip integrating the sentence to the previous discourse or performing any extra processing on the sentence (e.g., literality checks)<sup>4</sup> and may in fact end up processing the sentence faster than if it had found an interpretation for it. We saw earlier how this feature may explain the reaction-time reversal for the negatives in Experiments 2 and 3.

Coming back to the negation-as-distortion model, let us see how this model generates truth judgments. Remember that, in the introduction, we established the need to keep a tag associated to negation in the representation of the sentence (as in Clark and Chase’s model). After the model has processed the input sentence and reached the end of the sentence, there are several possible outcomes:

1. No interpretation was found. Set *Response = False* and go to step 3.
2. A final interpretation was found. Check whether sentence was literal or not, by retrieving bugs formed during the processing of the sentence<sup>5</sup>:
  - 2.1. The sentence is metaphoric/distorted. Set *Response = False* and go to step 3.
  - 2.2. The sentence is literal. Set *Response = True* and go to step 3.
3. If the negation tag is present, flip *Response*.
4. The value of *Response* is the final answer.

<sup>4</sup>Just and Carpenter (1980) call this extra time at the end of the sentence “wrap-up time”.

<sup>5</sup>If no bug can be retrieved the sentence is considered literal; otherwise metaphoric/distorted.

There are two chances of actually making an error: incorrectly judging the sentence as literal in step 2.2. and ignoring the negation tag in step 3. Judging a sentence as literal when in truth it is distorted can happen if the similarity of the distortion to the correct concept in memory is very high. It is the reason for which we often do not notice good metaphors and fall for Moses illusion sentences (see Budiu & Anderson, 2004 for an in depth explanation). In INP it is translated by the absence of a bug. The other source of error is incorrectly not retrieving (or not coding) a negation marker.

Therefore, according to this schema, most errors should occur in those cases in which processing should go through steps 3 (i.e., negative sentences) and 2.1 (i.e., the sentence should be metaphoric or distorted), but incorrectly skips them. Indeed, in Tables 1 and 2, the nonliteral negatives generate most errors. As expected, nonliteral sentences (whether negative or affirmative sentences) are usually worse in terms of accuracy than literal sentences<sup>6</sup> and negative sentences are more inaccurate than affirmative. Moreover, affirmative metaphoric sentences are more error-prone than affirmative nonsense sentences because, whereas affirmative nonsense sentences have low similarity to any proposition in memory and go through steps 1 and 3, the processing of affirmative metaphoric sentences goes through steps 2 and 3, and is likely to skip step 2.1 (if the similarity of the metaphor to its referent is high enough and, thus, the metaphor is good enough to be unnoticed).

To summarize, we have presented three experiments that show that the classic Chase and Clark (1974) model is not sufficient to account for the processing of negative sentences. The experiments are consistent with a negation-as-distortion model, based on the following principles:

- A negated item (NOT A) is represented as if it were a new concept highly similar with A, but not quite the same. Thus negation of A can be viewed as a distortion of A.
- A negation tag is kept for each negative sentence.
- Negation lowers the similarity of a sentence to propositions in background knowledge<sup>7</sup>. If this similarity is too low, the input sentence may not be connected to the background knowledge, some of the comprehension processes may be shortcircuited, and the sentence may be processed faster.

## References

Anderson, J.R. and Lebiere, C. (1998). *The atomic components of thought*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

<sup>6</sup>In Experiment 3, however, we see more errors for literal affirmatives than for any other sentences.

<sup>7</sup>Of course, this assumes that the sentence being processed contains NOT A and all the information in the background knowledge is about A. If the background knowledge contained facts about NOT A then the similarity to those sentences would not be lowered by NOT A (but would be lowered if A was present instead of NOT A in the input).

- Ayers, M., Reder, L., & Anderson, J. (1996). Accepting false information now and believing it later: Partial matching and false information in the Moses illusion. Unpublished manuscript.
- Budiu, R. and Anderson, J.R. (2004). Interpretation-based processing: A unified theory of semantic sentence comprehension. *Cognitive Science*, 28, pp. 1-44.
- Clark, H.H., & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472-517.
- Erickson, T., & Mattson, M. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20, 540-552.
- Giora, R., Balaban, N., Fein, O., & Alkabetz, I. (2004). Negation as positivity in disguise. In: H. L. Colston, & A. Katz (Eds.), *Figurative language comprehension: Social and cultural influences*. Hillsdale, NJ: Erlbaum.
- Just, M., & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.
- Lea, B. R., & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 303-317.
- MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 633-642.
- Mayo, R., Schul, Y., & Burnstein, E. (in press). "I am not guilty" versus "I am innocent": The associative structure activated in processing negations. *Journal of Experimental Social Psychology*.
- Oaksford, M.R., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 835-854.
- Reder, L., & Kusbit, G. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory and Language*, 30, 385-406.
- Valle Arroyo, F. (1982). Negatives in context. *Journal of Verbal Learning & Verbal Behavior*, 21, 118-126.
- van Oostendorp, H., & de Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, 74, 35-46.
- van Oostendorp, H., & Kok, I. (1990). Failing to notice errors in sentences. *Language and cognitive processes*, 5, 105-113.
- de Villiers, J., & Tager Flusberg, H.B. (1975). Some facts one simply cannot deny. *Journal of Child Language*, 2, 279-286.
- Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning & Verbal Behavior*, 4, 7-11.
- Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, 11, 92-107.