Word Learning in Context – Metaphors and Neologisms

Raluca Budiu*
Palo Alto Research Center, Palo Alto, CA 94304
John R. Anderson
Carnegie Mellon University, Pittsburgh,PA 15213

Abstract

Metaphors are a prolific source of new words in a language. In this chapter we study how people understand metaphors and how metaphors "freeze" and come to denote new meanings. We compare the understanding and learning of metaphors embedded in a text with the understanding and learning of equivalent nonsense words in the same context. We describe two experiments in which either metaphors or nonsense words are used anaphorically to refer to past concepts introduced in a preceding text. For anaphoric metaphors, subjects showed an initial bias to adopt a literal interpretation, which shifted as the experiment progressed. Subjects learned the meaning of the metaphors more rapidly and more accurately when compared with nonsense words. After repeated exposure to the words in appropriate contexts, metaphoric sentences were processed comparably with the sentences made only of literal words, whereas artificial-word sentences maintained a slight disadvantage. Results suggest that participants used context matching to understand and learn new words. We present a computational model that captures the essential trends in the data obtained from the two experiments.

Introduction

How do people acquire new words? Different estimates place the number of words that a child learns per year in the range of 2000 – 5000 (Nagy, Anderson, & Herman, 1987). There are differing opinions about whether this can or cannot be accounted for by children simply encountering words in context and inferring their meaning from the context. While agreeing that a child has about a 10% chance of learning a word each time it occurs in context, Landauer and Dumais (1997) conclude that there is no way a simple incremental word-by-word learning model could acquire

*The research was supported by grant SBR-94-21332 from the National Science Foundation and was conducted while this author was at Carnegie Mellon University, Pittsburgh, PA.

these many words, whereas Nagy, Herman, and Anderson (1985) conclude that this is just the rate of learning that would be necessary.

How do people manage to learn words from context? We aim to investigate this issue in what comes close to the actual situation people face in learning word meanings from context. As Nagy et al. (1985) point out, there are a number of unrealistic aspects to many studies of learning words from context. First, past studies have often explicitly instructed subjects that their task was to learn word meanings, while in natural situations the subjects' goal is presumably to understand text, and word learning occurs more in service of this goal. Second, the material subjects learn from is often unnaturally designed to facilitate the learning of the material. Third, many experiments have subjects learn a meaning for a known concept, whereas most real words we encounter are introduced because there is not another word for that meaning.

While these factors imply that word learning is often easier in the laboratory than in the real world, in other ways the laboratory situation can be more difficult. The words introduced in most laboratory experiments offer no hints to their meanings. On the other hand, as Nagy and Scott (1990) have shown, subjects often have expectations about word meanings based on the words themselves. One instance of such a situation is when an old existing word is given a new meaning as a metaphoric extension of its old meaning, such as the word *surf* in the phrase *surfing the web*.

We began our research interested in metaphoric extensions of old words. The first experiment in this paper is exclusively concerned with learning of this variety. However, we became convinced that the processes associated with this kind of word learning was not fundamentally different than the learning of completely new words, neologisms. The only difference is that it is easier to induce the features of new meanings in the context of a metaphor. Therefore, the second experiment in this paper compares learning of neologisms (new words) and metaphors.

One major goal of this research is to develop a computational model of how such learning takes place. We describe a computational model that has been developed in the ACT-R architecture (Anderson & Lebiere, 1998). This is an architecture that has been used to model a wide variety of cognitive phenomena including mathematical problem solving, list memory, strategy choice, analogical reasoning, and scientific reasoning. Our principal purpose for developing our model in this architecture is to establish that there is nothing different about word learning than these other varied cognitive functions.

Since our first experiment is exclusively concerned with metaphors as a source of new words and extends understanding of metaphoric processing, we begin with a brief review of previous research on metaphoric theory.

Previous Research on Metaphors

Although metaphors are most often associated with poetic language, they are truly pervasive in everyday language. Frequently encountered examples include *Time is money*, *neck of the woods*, *foot of the mountain*, *She is my light*, *Bob is a pig*, *war with inflation*. Metaphors explain a concept (called **topic** or **target**) by making direct or indirect reference to another one (called **vehicle** or **source**). In the aphorism *Time is money*, the vehicle is *money* and the topic is *time*. Sometimes the topic can be implicit — in *foot of the mountain* the topic is the region that lies at the base of the mountain. This kind of metaphor with implicit topic is called **anaphoric metaphor**.

Metaphor Comprehension

Perhaps the most famous and frequently refuted theory of metaphor comprehension is Searle's error-recovery theory (Searle, 1979). Searle claimed that, when confronted with a metaphor, people first try to understand the sentence literally, and, in case of failure, they look for a metaphorical interpretation. The decision whether the literal meaning is appropriate or not is based on context. Thus, the recognition of a metaphor consists of three steps: first, a literal interpretation of the sentence is built; second, this interpretation is matched against the context; third, if no consistent matching can be found, a metaphorical interpretation is considered. A corollary of this theory is that people take longer to understand metaphorical utterances than to understand literal utterances. But ulterior psychological evidence (Ortony, Schallert, Reynolds, & Antos, 1978; Inhoff, Lima, & Carroll, 1984; Shinjo & Myers, 1987; Keysar, 1989; Budiu & Anderson, 2002) shows that in many situations it does not take longer to understand metaphoric sentences.

Ortony et al. (1978) showed subjects two contexts for the same sentence: a literal inducing context and a metaphorical inducing context. For instance, the sentence *The hens clucked noisily* was preceded either by a passage about some chickens on a farm, or by a passage about a women's club meeting. The authors measured reading times for target sentences and obtained no difference between metaphorical and literal targets. However, when they made the prior context short (one sentence or less), they found a significant difference in reading times for the two kinds of targets. Inhoff et al. (1984) replicated this study and obtained similar results. They also collected eye movement data for the critical words (e.g. *hens* for the example above); there was no difference between the viewing times for these words in the literal and metaphoric contexts. These results were interpreted in favor of literal and metaphoric processes being similar, and against a stage model that predicts that metaphorical sentences take longer to comprehend.

Not only do subjects sometimes access the meaning of a metaphor as fast as the literal meaning, but the metaphoric meaning can interfere with the literal meaning. Glucksberg, Glidea, and Bookin (1982) had subjects judge the literal truth of sen-

tences of the form *A is B*. Subjects took longer to reject sentences that made sense metaphorically than nonsense sentences, albeit both being literally false (e.g. *Some jobs are jails*). Keysar (1989) extended Glucksberg et al.'s results in an ingenious experiment. He manipulated both the literal and metaphoric truth of sentences *A is B*, by varying the context that preceded them. Thus the target sentence could be both literally and metaphorically true or both literally and metaphorically false (congruent conditions), or literally true and metaphorically false or literally false and metaphorically true (incongruent conditions). He obtained shorter judgement times for the congruent conditions than for the incongruent ones, which argues for the inseparability of metaphoric and literal processing. Keysar (1989) repeated the experiment and measured comprehension times; he obtained fastest reading of sentences that were both literally and metaphorically true. The next fastest latencies were for sentences that were either literally or metaphorically true, and slowest reading times for sentences that were both metaphorically and literally false.

Gibbs (1990) performed an experiment where subjects had more rapid access to literal meanings. It showed that, in the case of anaphoric metaphors, when the topic of the metaphor is not explicitly mentioned in a sentence, but it is referred exclusively by the vehicle, the reading times for metaphoric utterances are longer than for equivalent literal utterances, even if the context is metaphor-supportive. The topics of the metaphors in Gibbs' experiment had been always introduced in the context preceding the sentence. For instance, in the sentence *The creampuff didn't show up*, the metaphor *creampuff* referred to a boxer introduced in an anterior passage. Worth mentioning is the fact that Gibbs' metaphoric sentences had no possible literal interpretation. Onishi and Murphy (1993) replicated Gibbs' findings; they also proved that those results were not an artifact of insufficient story knowledge or lack of referent salience. In another experiment, Onishi and Murphy modified the target sentences to have the form *A is B* (i.e. *The boxer is a creampuff*). For these metaphors, they obtained no reading time difference between metaphoric and equivalent literal sentences.

Based on Onishi and Murphy's study, one could infer that predicative *A is B* metaphors are as easy to comprehend as literals, whereas for anaphoric metaphors some extra time is needed. However, both Ortony et al. (1978), Budiu and Anderson (2002) are examples of anaphoric metaphor studies that did not find any difference between reading times for sentences involving anaphoric metaphors or literals. As we have discussed elsewhere (Budiu & Anderson, 2002, 2004), the mixed results that various studies have found for anaphoric and predicative metaphors can be understood in the light of a speed-accuracy trade off: for some anaphoric metaphors, the context of the sentence can be rich enough so that subjects are able to find an interpretation; however, for such sentences subjects pay a cost in the time to comprehend the sentence. For other sentences, the context may be too unsuportive and the participants may speed through the sentence without a good understanding of its meaning.

In conclusion, it seems that Searle was wrong in his assumption that literal interpretation must always precede metaphoric comprehension. In fact, the data are more consistent with a continuum of metaphoric and literal processing (see Budiu & Anderson, 2004 for a more elaborate discussion), in which the same kind of mechanisms are involved in both metaphor and literal comprehension. The various effects are produced by how similar the vehicle and the topic of the metaphor are and also by how much extra information is supplied by the sentence context.

We can assume that, as a metaphor gains familiarity, it comes to be processed as rapidly and perhaps more rapidly than the literal meaning. As a metaphor acquires increased familiarity, the word reaches the point where it just has another meaning. Indeed, there are a lot of words in natural languages that are such metaphorical extensions of old meanings, based on more or less obvious similarities between two concepts. It is often the case that an analogy gives rise to the same metaphor in several languages – like in the commonly met "*leg of a table*" or "*foot of the mountain*", or even in the less obvious cases of using verbs like the English *catch* or *grasp* to mean *understand* (Ullmann, 1966). Ullmann (1966) mentions other interesting examples of words of metaphoric origin: the word for *tongue*, the organ of speech, stands for *language* in many Indo-European and Finno-Ugrian languages; the word for the *pupil* of the eye in Latin (*pupilla*), Spanish (*niña*), Portuguese (*menina*) and other languages is the same with the words for *little girl* and comes from the "vague resemblance between a child and the minute figure reflected in the eye"; many words have antropomorphic origin (e.g. "*neck* of a bottle", "*eye* of a river", "*mouth* of a river" etc.), or, the other way around – words denoting parts of human body initially had non-antropomorphic meanings (e.g. *muscle* form Latin *musculus*, meaning *little rat*; Adam's *apple* etc.). Some of the new meanings of such words coexist with the older meanings, others replace them.

As an old word acquires a new meaning, it is reasonable to assume that subjects should no longer suffer a comprehension deficit. As the time to comprehend a sentence with an anaphoric metaphor approaches the time to comprehend sentence with a conventional word, the subject must be acquiring a new meaning for the metaphor. Thus, we can trace meaning acquisition without explicitly asking subjects to state the meaning of the word and, hence, we can get around the criticism of past experiments, which was that they were collecting explicit measures of word meaning and so made the learning of word meanings the subject's task. Thus, our first experiment involves repeated presentation of a word used as an anaphoric metaphor in various sentences and compares the comprehension of these sentences to that of similar literal sentences.

Experiment 1

The experiment compares literal and metaphoric sentence understanding under repeated exposure to the same words. We closely modeled our experiment after Gibbs

Table 1: Metaphors used in Experiment 1. Average familiarity (Fam) and goodness (Good) of each metaphor were rated on scale from 1 to 4 (1 = completely unfamiliar/very bad, 4 = very familiar/very good). Accuracy slopes (S) and intercepts (I) for each metaphor in Experiment 1.

Topic	Vehicle	Fam	Good	S	I
massive athlete	bear	2.30	2.60	10.32	44.92
cold room	freezer	3.30	3.20	1.47	70.42
hardworking farmer	ant	2.60	2.80	1.88	68.33
tough food	rubber	3.70	3.00	0.57	71.93
dependent husband	puppy	2.30	2.50	8.77	52.81
unfeeling official	iceberg	2.20	3.00	5.61	65.83
dishonest politician	fox	2.30	2.50	4.28	60.52
slow waiter	snail	3.10	3.00	5.79	63.57

(1990) study. Participants read short passages; then they read a target sentence and had to judge whether it was true or false based on the passage. The target sentence could contain an anaphoric metaphor or not. At the end of the experiment, we tested participants' preference for the metaphors that they had learned by having them choose between a sentence containing an "old", learned metaphor and the same sentence containing a new metaphor. We expect that subjects would be inclined towards using new metaphors. Our hypothesis is based on the assumption that, after a metaphoric word acquires a new meaning, it loses its figurative, metaphoric force. We further assumed that our subjects would prefer words with greater metaphoric force.

Method

Participants. Forty Carnegie Mellon University undergraduates who were native English speakers participated in the experiment for course credit.

Materials. We constructed 8 metaphors (see Table 1) and 64 short passages (8 for each metaphor) about the topics of these metaphors. Thus, we had 8 distinct passages about massive athletes, 8 about cold rooms etc. The average length of the stories was 129 words. For each story we created two true and two false possible target sentences to judge. Each of these two targets could be given in a literal form or in a metaphoric form. The metaphoric targets contained an anaphoric metaphor. For instance, for a story about a bulky athlete, a metaphoric target would be *The bear helped the housewife to crack the nuts* and the corresponding literal target would be *The wrestler helped the housewife to crack the nuts.* Table 2 contains an example of a story and the associated sentences.

For the last phase of the experiment, the metaphor preference testing, we created

Table 2: A story corresponding to the metaphor *cold room — freezer* and its associated sentences.

Jim went to a conference he was very excited about. The conference took place in a very large and cold room, and only the first rows were fully occupied. He sat somewhere in the middle rows. Jim was feeling bad — his hands and feet were frozen, and he doubted he could spend two hours in that awful room. But when the speaker started his presentation, it was so interesting that Jim wasn't able to feel anything else but enthusiasm for what he was listening to.
Metaphorical sentences:
Jim forgot very soon about being in a freezer. *[true]*
Only the last rows of the freezer were occupied. *[false]*
Literal sentences:
Jim forgot very soon about being in a cold room. *[true]*
Only the last rows of the room were occupied. *[false]*

8 new passages, each passage being about the topic of one metaphor. There were two possible ending sentences for each passage; they were both metaphoric and identical up to one word. The subjects had to choose between the two endings. One ending contained one of the metaphors in Table 1, and the other contained an alternative metaphor with the same topic. The alternative metaphors are described in Table 3.

Testing 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 Blasko & Connine, 1993; Tourangeau & Rips, 1991, for counterexamples).

Therefore, we have conducted a rating study in which 10 native English speakers, students or staff of Carnegie Mellon University, rated the goodness and the familiarity of the metaphors on a scale from 1 to 4 (1 — low familiarity/goodness; 4 – high familiarity/goodness). Participants rated the 16 metaphors (basic and alternative) described in the materials section plus other 173 metaphors we used for a different study. From these 173 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). These ensured that subjects use all the points on the scale.

Each participant saw the metaphors in a random order. Half of the subjects saw

Table 3: Alternative metaphors used in the metaphor preference task of experiment 1. Average familiarity (Fam) and goodness (Good) of each metaphor were rated on scale from 1 to 4 (1 = completely unfamiliar/very bad, 4 = very familiar/very good).

Topic	Vehicle	Fam	Good
massive athlete	mountain[a]		
cold room	igloo	2.60	2.90
hardworking farmer	beast of burden	2.20	2.30
tough food	cardboard	2.70	2.90
dependent husband	footman	1.30	2.80
unfeeling official	stone	2.30	2.80
dishonest politician	chameleon	2.00	2.30
slow waiter	sloth	2.40	2.60

[a]An error prevented us from collecting ratings for this metaphor.

Table 4: Average goodness and familiarity ratings for metaphors used in Experiments 1 and 2, compared with corresponding average ratings of good and nonsensical metaphors (scale: 1 — lowest; 4 — highest).

	Familiarity	Goodness
Experiment 1	2.72	2.82
Experiment 2	2.58	2.76
Good metaphors	2.75	2.72
Nonsense metaphors	1.40	1.73

the metaphors in the form of similes (*A is like B*) and the other half saw them in the *A is B* form. The ratings for the similes and the metaphors were well correlated ($r = .75$ for familiarity and $r = .74$ for goodness); henceforth we present aggregated data over the two conditions.

The averages of the goodness and familiarity ratings are given in Table 1 for the basic metaphors and in Table 3 for the alternative metaphors. Table 4 contains the averages of the metaphors used in Experiment 1 and the averages of the other metaphors that subjects rated. As one can see, the average familiarity and goodness for the metaphors we used were close to the same quantities for the good metaphors used in other research.

Procedure. The participants saw the materials on the screen of a MacIntosh computer. Participants used one of two keys (*K* for *true*, *D* for *false*). Each participant took about 45 minutes to complete the experiment.

A trial consisted of two phases:

Table 5: Design of Experiment 1. Met = Metaphoric; Lit = Literal.

	Topics 1-4	Topics 5-8
Block 1 (Pass 1 + 2)	4 Met Trues	4 Lit Trues
	4 Met Falses	4 Lit Falses
Block 2 (Pass 3+4)	4 Met Trues	4 Lit Trues
	4 Met Falses	4 Lit Falses
Block 3 (Pass 5+6)	4 Met Trues	4 Lit Trues
	4 Met Falses	4 Lit Falses
Block 4 (Pass 7+8)	4 Met Trues	4 Lit Trues
	4 Met Falses	4 Lit Falses

1. Participants read a short passage.

2. When finished, they pressed a key, and the passage on the screen was replaced by the target sentence (metaphorical or literal, true or false). Subjects had to decide whether the sentence was true or false and press one of the two keys to indicate their answer. They were instructed to be as fast as possible. After they gave their answer, they were given feedback (on a new screen) and pressed a key to continue with the next trial.

Table 5 illustrates the design of the experiment. For each subject four of the 8 topics from Table 1 were randomly selected to be referred metaphorically. The remaining four for that subject were referred to literally. That is, one subject might see metaphoric targets for all the stories concerning, for instance, massive athletes, and literal targets for all the stories concerning, for instance, cold rooms. Each target was selected to be true or false at random.

There were 8 passes through the material. In each pass the subject would see one story involving each topic for a total of 8 stories in all. The order of the stories in each pass was randomized for each subject, as was the assignment of stories to passes. Note that each story was used just once – there were 64 stories in all. These passes were organized into blocks of two passes. In each block, on average, one story for each topic would be followed by a true sentence and one story would be followed by a false sentence. The choice of which to follow with a true sentence was also random. Note that this methodology for material generation meant that each subject had a different random set of materials and metaphors and that in generalizing over subjects we are also generalizing over materials (Clark, 1973).

The independent variables we used for this experiment were (1) the 16-trial block (two passes aggregated), (2) the type of trial: metaphoric or literal, depending on the type of the target, and (3) the truth of the sentence to judge (target): true or false.

As a final test of whether the metaphors had completed their transitions to words with new meanings, at the end of the experiment participants had to perform the metaphor selection task. In this task, subjects read a passage, and then chose one out of two anaphoric metaphors to complete a sentence related to the passage. One metaphor came from our initial selection, the other was an alternative metaphor with the same target (as described in the materials section). In this task, there were 8 trials – each one corresponding to one of our 8 topics. For half of these trials neither of the two metaphors were encountered before (in our experiment) by the participant. We measured the proportion of times the subjects would use a metaphor, when they had seen that metaphor before and when they had not seen it.

Results

Because we wanted to study the learning of words over extended experience, we were constrained to look at relatively few topics in order to fit the learning experience into a reasonable period of time. This raises questions about whether there is something peculiar about our 8 topics that could be producing the results. Therefore, we report tests over both subjects (F_1 statistics) and items (F_2 statistics). Since the materials were randomly generated for each subject, Clark (1973) points out that these tests are rather redundant since in generalizing over subjects we are generalizing over materials and vice versa. However, we thought it was important to perform the item tests to provide assurance that our effects were not specific to particular items.

We eliminated from our analysis those sentences for whom the average accuracy (aggregated over both metaphoric and literal conditions) was less than 50%[1]. These sentences had also a low average accuracy for the literal condition as well as for the metaphoric condition, and that made us suspect that subjects were confused or misunderstood those passages or targets. For true target analysis, we eliminated 3 passages out of 64 (4.68%); and for false target analysis we eliminated 6 passages out of 64 (9.37%).

Percentage Correct. Figure 1 presents the variation of the percentage of correct answers to true or false sentences over the duration of the experiment. The points in the two graphs are averages over one block — that is, over 16 consecutive trials. We performed an analysis of variance (ANOVA) on the percentage of correct sentence judgements. There were significant main effects of type of trial — metaphoric versus literal ($F_1(1,39) = 37.06$, $MSE_1 = 623.21$, $p_1 < 0.0001$, $F_2(1,7) = 9.83$, $MSE_2 = 316,599.27$, $p_2 < 0.05$), block ($F_1(3,117) = 5.22$, $MSE_1 = 390.21$, $p_1 < 0.005$, $F_2(3,21) = 44.51$, $MSE_2 = 245,697.82$, $p_2 < 0.0001$). There was an effect of truth — true versus false sentences — for the item analysis ($F_2(1,7) = 8.87$, $MSE_2 =$

[1] We actually eliminated those trials for which the accuracy in both Experiment 1 and Experiment 2 was less than 50%.

$100,218.46$, $p_2 < 0.05$), but it did not reach significance for the subject analysis ($F_1(1,39) = 2.25$, $MSE_1 = 2,011.14$). There was a significant interaction between trial type and block ($F_1(3,117) = 8.94$, $MSE_1 = 404.55$, $p_1 < 0.0001$, $F_2(3,21) = 5.79$, $MSE_2 = 335,852.48$, $p_2 < 0.005$). This is the most important interaction for the purposes of our interpretation. The other interactions were significant only for the subject ANOVA: between trial type and truth ($F_1(1,39) = 21.90$, $MSE_1 = 1,029.58$, $p_1 < 0.0001, F_2(1,7) = 0.004$, $MSE_2 = 354,267.95$), between truth and block ($F_1(3,117) = 13.32$, $MSE_1 = 490.39$, $p_1 < 0.0001$, $F_2(3,21) = 1.49$, $MSE_2 = 313,698.63$), between truth, trial type and block ($F_1(3,117) = 14.48$, $MSE_1 = 374.70$, $p_1 < 0.0001$, $F_2(3,21) = .93$, $MSE_2 = 186,929.28$).

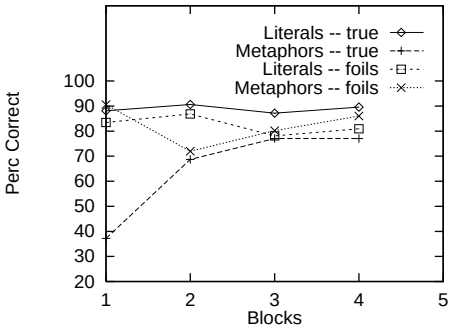


Figure 1: Percentage of correct answers in Experiment 1 as a function of blocks of practices.

The significant interaction between trial type and block indicates that subjects were learning the metaphors. Over the course of the experiment subjects started out performing poorly on metaphoric expressions but ended up with an accuracy close to that of literals. This supports the interpretation that subjects were learning new meaning for the metaphoric words. Table 1 gives the slope of the linear improvement for each of the 8 metaphors. They are all positive indicating that this effect was general to all the items – indeed, an effect significant by a sign test. We wanted to check whether the familiarity or the goodness of the metaphors affected their learning. For this purpose, we included only those trials when the target was metaphoric, and we looked at the slopes and the intercepts of the learning curves (which can be found in columns 5 and 6 of Table 1). We obtained that both the goodness and the familiarity were moderately correlated with the learning slope ($r = 0.61$ for goodness and $r = .69$ for familiarity). The correlation for goodness is non-significant, but the correlation for familiarity is marginally significant ($t(7) = 2.30, p < .10$) and indicates greater learning for the familiar metaphors. For the intercepts, there was a significant correlation with goodness ($r = .75$, $t(7) = 2.79$, $p < 0.05$) and a marginally significant correlation with familiarity ($r = .65$, $t(7) = 2.11$, $p < 0.1$), which indicates that good or familiar metaphors may have started with a comprehension advantage over the other metaphors.

At the beginning of the experiment, the accuracy for true metaphoric sentences is low, but it is high for foils in the same condition. Unlike for the overall data, this effect of truth is shown to be significant by an ANOVA performed only on the first block ($F_1(1,39) = 58.37$, $MSE_1 = 330.63$, $p_1 < 0.0001$, $F_2(1,7) = 72.75$, $MSE_2 = 59.64$, $p_2 < 0.0001$), which suggests that initially participants were often interpreting metaphoric sentences literally. By the end of the experiment, the percentage of correct answers for the metaphoric sentences became close to the same quantity for the literal condition. When performing ANOVA only for the last block of the experiment, none of the effects found in the overall data or in the data for the first block were significant[2].

Response Latencies for Judging Sentences. The ANOVA on judgement latencies showed main effects of trial type ($F_1(1,39) = 15.71$, $MSE_1 = 1,015,149.51$, $p_1 < 0.005$, $F_2(1,7) = 9.83$, $MSE_2 = 316,599.27$, $p_2 < 0.05$) and of block ($F_1(3,117) = 58.69$, $MSE_1 = 926,561.35$, $p_1 < 0.0001$, $F_2(3,21) = 44.51$, $MSE_2 = 245,697.82$, $p_2 < 0.0001$). The truth was not significant over subjects ($F_1(1,39) = 1.60$, $MSE_1 = 1,025,256.56$) but it was significant over items ($F_2(1,7) = 8.87$, $MSE_2 = 100,218.46$, $p_2 < 0.5$) . We also found a significant interaction between trial type and block ($F_1(3,117) = 7.89$, $MSE_1 = 800,726.10$, $p_1 < 0.0001$, $F_2(3,21) = 5.79$, $MSE_2 = 335,852.48$, $p_2 < 0.005$). Again, this is the most important interaction because it indicates learning of the metaphors. No other interaction proved to be significant. Figure 2 contains graphs for the variation of response latencies for the case of correct answers.

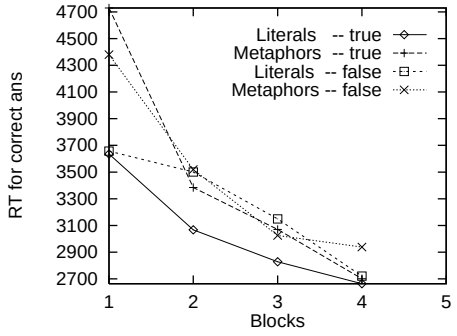


Figure 2: Response times for correct answers in Experiment 1 as a function of blocks of practice.

We ran separate ANOVAs for the data from the first and last blocks. In the first block, there was a strong effect of trial type ($F_1(1,39) = 28.78$, $MSE_1 =$

[2]The interaction between trial type and truth was significant, though ($F_1(1,,39) = 5.55$, $MSE_1 = 561.29$, $p_1 < 0.05$, $F_2(1,7) = 14.69, MSE_2 = 60.11$, $p_2 < 0.01$). Trues were still more accurate than falses for literals, but the reverse pattern held for metaphors.

$1,146,421.25$, $p_1 < 0.0001$, $F_2(1,7) = 16.04$, $MSE_2 = 549,758.20$, $p_2 < 0.01$), which disappeared by block four ($F_1(1,39) = 2.57$, $MSE_1 = 252,201.45$, $F_2(1,7) = .32$, $MSE_2 = 260,402.63$). Actually, no effects or interactions were found for the data from the last block, suggesting that by the end of the experiment the metaphoric sentences were comprehended as easily and as fast as literal sentences.

We also looked at the slopes and intercepts of the latency curves. There were marginally significant correlations of slopes with familiarity ($r = .65$, $t(7) = 2.09$, $p < 0.1$) and with goodness ($r = .64$, $t(7) = 2.09$, $p < 0.1$). We found no significant correlations for the latency intercepts.

Metaphor preference testing. Finally, we looked at subjects' preference for the metaphors used in Experiment 1 versus alternative, new metaphors. For the four metaphors that they had seen in the experiment, subjects showed a strong usage preference for the new metaphors, selecting the old metaphor only an average of 1.0 times, and the other 3.0 times. In the case of the four metaphors that they had not seen, subjects were much more evenly split, selecting the experiment (not seen) metaphor an average of 1.7 times and the other metaphor 2.3 times. This difference was significant whether tested over subjects ($t(39) = 2.58, p < 0.05$) or items ($t(7) = 4.61, p < 0.01$). Thus, we have further evidence that with practice subjects were no longer regarding the metaphoric words as metaphors, and so the words had lost some of their figurative force. We also tested whether the difference in familiarity or goodness between the two kinds of metaphors had any effects on the subjects' preference. The ANOVA did not find any effects or interactions.

Discussion

In the first trials, participants were slower and less accurate at judging metaphorical sentences than literal sentences. It is plausible that they first attempted a literal interpretation of the metaphorical sentences. However, during the course of the experiment, as participants were exposed to subsequent encounters of the same metaphors, their performance for metaphorical trials became comparable with the performance in the literal trials. A possible explanation is that, by the end of the experiment, they formed a new meaning for the vehicle of the metaphor and they retrieved this meaning, thus processing the metaphors as regular words that had entered their lexicon.

Budiu and Anderson (2003) presents a computational model for the data in the first block of this experiment[3]. That model accounts for the differences between metaphors and literals in the more general framework of the INP sentence processing theory (Budiu & Anderson, 2004), but is not concerned with how metaphors acquire

[3]Budiu and Anderson (2003) show how the model matches the metaphoric and literal data from the first block in Experiment 2; however, those data are closely correlated with the data from Experiment 1, and the same explanation can be used for both.

new meanings from context. After presenting Experiment 2, we describe another ACT-R (Anderson & Lebiere, 1998) computational model focused on capturing the building of a new meaning over time. That model explains most of the improvement over blocks as due to the gradual acquisition of elements of new word meanings for the metaphors. According to this model, there are only two effects of the use of metaphors. First, subjects started out with a literal bias, and one of the reasons for their high accuracy on metaphoric foils was a tendency to reject sentences that used words non-literally. Second, the metaphor itself guided subjects to the features that defined the new word and so facilitated the meaning acquisition. To test our computational model and check that these were the two features that distinguished metaphoric words, we ran a second experiment that also used neologisms – that is, completely new words with new meanings.

Experiment 2

Because it appeared that in Experiment 1 subjects were learning new word meanings, in Experiment 2 we introduced trials that contained artificial words. If subjects were learning new meanings for the metaphors, they would come to treat the metaphors and the artificial words equivalently. To assess the degree to which the participants really learned the words, we added a post test phase in which we asked them to define the words they encountered during the experiment.

We used two kinds of foils: hard and easy. The reason was that, in Experiment 1, a possible strategy for participants was to ignore the metaphors and judge the sentences by checking whether the understandable part of the sentence matched the story. This strategy was practically infallible in the setting of the first experiment and saved the participant the effort of processing a word with an unknown meaning. This is perhaps why subjects never showed an accuracy effect for false metaphoric sentences but only for true ones. The hard foils in the second experiment were sentences for which this kind of matching strategy would fail.

Method

Participants. The participants were 40 Carnegie Mellon University undergraduates who were native English speakers and who participated in the experiment for credit. None of them was a participant in Experiment 1.

Materials. We used the eight metaphors pairs from the Experiment 1, to which we added one more metaphor (*poor musician – cricket*). We used the same old 64 passages plus eight more corresponding to the new pair. The average length of the passages was 132 words. This time, for each passage we constructed 6 possible target sentences: two literal, two metaphoric, and two containing an artificial word.

Table 6: A story corresponding to the pair *massive athlete – bear* and its associated sentences.

Jim was a philosophy junior. In one of his classes, he noticed a very massive young man who was always sleeping and never paid any attention to the discussions. The character came to the other seminars, too, but nothing seemed to raise his interest. One day, somebody told Jim the man was a very good linebacker that had been all-state in football. So the mystery was solved: he was accepted at the university for his athlete rather than for his philosopher qualities.
Metaphorical sentences:
The bear was sleeping in the philosophy class. *[true]*
The bear noticed a man sleeping in class. *[false]*
Literal sentences:
The athlete was sleeping in the philosophy class. *[true]*
The athlete noticed a man sleeping in class. *[false]*
Artificial-word sentences:
The carope was sleeping in the philosophy class. *[true]*
The carope noticed a man sleeping in class. *[false]*

As before, the metaphoric targets contained anaphoric metaphors. For the artificial-word targets, the metaphors were replaced by a nonce word. Some of the metaphoric and literal sentences were identical with sentences used in Experiment 1.

Half passages had hard false targets associated with them — that is, those targets could not be answered correctly without knowledge of all the words in the sentence. The false sentences from Experiment 1 were not hard – even if the participants did not know what *freezer* meant in that context, they could infer that the sentence *Only the last rows of the freezer were occupied* was false, since there was no object in the story whose *only last rows were occupied*. An example of a story with an associated hard false sentence is given in Table 6. If there is no knowledge about whom *bear* denotes in the sentence *The bear noticed a man sleeping in class*, the answer cannot be given by matching the rest of the sentence to the passage. Indeed, if *bear* refers to *Jim*, then the sentence is true, and if it refers to *the athlete*, then the sentence is false.

We selected three artificial words (*pinten, zolper, carope*) that were rated as the most plausible English words from a list of 14 artificial words. We assumed that no etymological or morphological knowledge could bias participants towards a special meaning of these words.

Procedure. We used the same experimental paradigm as for Experiment 1. Participants took about 45 minutes to finish the tasks. A trial consisted of the same

three phases as before. The only exception was that in this experiment the target could be literal, metaphoric, or artificial-word.

For each subject, one third of the targets were literal, one third were metaphoric and one third were artificial-word. We used the same design as illustrated in Table 5, except that now there were three columns for literal, metaphoric, and artificial-word with three topics in each.

For each subject we randomly assigned three topics for metaphors and three for artificial words. That is, a subject could see metaphoric targets for all the stories concerning, for instance, massive athletes, literal targets for all the stories concerning, for instance, cold rooms, and targets containing the artificial word *carope* for all stories concerning untalented musicians. The topics were associated with artificial words randomly — that is, once the topic *untalented musician* was selected for a subject, it was decided randomly whether *carope*, *pinten* or *zolper* would denote it for that subject.

Each target was selected to be true or false at random. Half of the false targets were written to be hard. That is, for a subject, about half of the targets were true, about a quarter were easy foils and another quarter were hard foils.

As before, the experiment was structured in 8 passes. Each pass had 9 trials containing 9 different passages pertaining to the 9 different metaphor topics. Hence, in total, subjects read 72 stories. As before, no subject saw the same passage or the same target twice, hence it was impossible for subjects to use a decision retrieval strategy.

At the end of the experiment, participants were asked to define the meanings of the metaphors and of the artificial words encountered during the experiment. Each participant had to define the word first out of context, and then with an example of its usage on the screen. The examples were sentences that had occurred in the experiment.

Results

As for Experiment 1, we report both F_1 and F_2. Again, we note that, since materials were randomly assigned to conditions for each subject, the item statistics are somewhat redundant. We eliminated from our analysis those sentences for which the average accuracy (aggregated over both metaphoric and literal conditions) was less than 50% in both Experiment 1 and Experiment 2[4]. For true target analysis, we eliminated 3 passages out of 72 (4.16%); and for false target analysis we eliminated 7 passages out of 72 (9.72%).

Percentage Correct. The percentage of correct answers is plotted in Figure 3. These results are very similar with the results obtained in Experiment 1 (see Figure 1). While this experiment involved new target and trial types (hard foils,

[4]See footnote 1. There was one passage that occurred only in Experiment 2 and was eliminated.

artificial-word trials), the correlation between the common conditions in Figure 1 and Figures 3 is $r = 0.80$. Participants started by being inaccurate for non-literal true sentences, but very soon their performance improved and became comparable with that for literal trials.

The ANOVA on percentage correct during sentence judgement showed significant effects for the trial type (literal, metaphor or artificial-word) — $F_1(2, 78) = 20.46$, $MSE_1 = 948.10$, $p_1 < 0.0001$, $F_2(2, 16) = 8.79$, $MSE_2 = 400.68$, $p_2 < 0.005$, and for truth ($F_1(2, 78) = 7.11$, $MSE_1 = 2,599.76$, $p_1 < 0.005$, $F_2(2, 16) = 4.67$, $MSE_2 = 687.36$, $p_2 < 0.05$) and significant interactions between truth and block ($F_1(6, 234) = 4.57$, $MSE_1 = 999.87$, $p_1 < 0.0005$, $F_2(6, 48) = 2.29$, $MSE_2 = 468.80$, $p_2 < 0.05$). The block effect was significant by the subject analysis ($F_1(3, 117) = 2.71$, $MSE_1 = 846.50$, $p_1 < 0.5$) but not by the item analysis ($F_2(3, 24) = .68$, $MSE_2 = 552.33$). The other interactions were also significant only for the subject analysis: between trial type and truth ($F_1(4, 156) = 5.56$, $MSE_1 = 1,286.66$, $p_1 < 0.0005$, $F_2(4, 32) = 2.46$, $MSE_2 = 542.06$), between trial type and block ($F_1(6, 234) = 2.98$, $MSE_1 = 779.23$, $p_1 < 0.01$, $F_2(6, 48) = 1.31$, $MSE_2 = 281.39$) and between trial type, truth and block ($F_1(12, 468) = 2.73$, $MSE_1 = 793.35$, $p_1 < 0.005$, $F_2(12, 96) = 1.67$, $MSE_2 = 244.98$, $p_2 < 0.1$).

ANOVA on the first block data found main effects of truth ($F_1(2, 78) = 26.90$, $MSE_1 = 1,287.86$, $p_1 < 0.0001$, $F_2(2, 16) = 6.60$, $MSE_2 = 514.77$, $p_2 < 0.01$) and of trial type ($F_1(2, 78) = 26.90$, $MSE_1 = 681.23$, $p_1 < 0.0001$, $F_2(2, 16) = 10.92$, $MSE_2 = 278.38$, $p_2 < 0.001$),and an interaction between trial type and truth ($F_1(4, 156) = 10.03$, $MSE_1 = 924.93$, $p_1 < 0.0001$, $F_2(4, 32) = 5.12$, $MSE_2 = 342.86$, $p_2 < 0.05$). For easy foils, participants were accurate from the beginning (see Figure 3), which suggests evidence for a context matching strategy in the first block. Subjects showed somewhat greater accuracy on artificial-word true sentences than on metaphoric true sentences, but lower accuracy on hard artificial-word foils than on hard metaphoric foils. This finding suggests a literal interpretation strategy for processing metaphoric sentences in which subjects were inclined to reject as false all metaphoric sentences.

On the other hand, in the last block the effect of trial type was preserved for the subject analysis ($F_1(2, 78) = 4.45$, $MSE_1 = 641.71, p_1 < 0.05$, $F_2(2, 16) = .97$, $MSE_2 = 212.31$); however there was no statistical significant difference between the metaphor and literal conditions ($F_1(1, 39) = .63$, $MSE_1 = 703.11$), which suggests that the artificial words were responsible for the trial type effect in the last block. Therefore, we can conclude that, while participants processed metaphors in the same way as literals at the end of the experiment, they still had some difficulty with the artificial words.

As for Experiment 1, we computed the learning slopes and intercepts of each of the items. Unlike in Experiment 1, we did not find a significant correlation between the learning slopes and the familiarity ($r = .44$) or goodness ($r = .11$) of metaphors. However, the correlation with familiarity was at least in the same direction as in

Experiment 1. The slopes for the metaphors were not significantly greater than 0 ($t(8) = .55$), but the slopes for the artificial words were ($t(8) = 2.92$). The intercepts for the metaphors were not correlated with either goodness or familiarity. The weaker effects in this experiment may reflect the fewer observations per item-by-type combination.

Response Time For Sentence Judgement. The judgement latencies for Experiment 2 are presented in Figure 4. The ANOVA on judgement latencies showed main effects of truth ($F_1(2,78) = 10.34$, $MSE_1 = 17,669,219.63$, $p_1 < 0.0001$, $F_2(2,16) = 4.25$, $MSE_2 = 1,133,824.38$, $p_2 < 0.05$) and of block ($F_1(3,117) = 66.44$, $MSE_1 = 2,036,181.94$, $p_1 < 0.0001$, $F_2(3,24) = 45.08$, $MSE_2 = 774,409.80$, $p_2 < 0.0001$). The effect of trial type was significant for subjects, but only marginally significant for items ($F_1(2,78) = 21.60$, $MSE_1 = 1,297,688.51$, $p_1 < 0.0001$, $F_2(2,16) = 3.26$, $MSE_2 = 2,512,082.02$, $p_2 < 0.1$). We found a significant interaction between trial type and block ($F_1(6,234) = 4.44$, $MSE_1 = 1,3914,415.52$, $p_1 < 0.0005$, $F_2(6,48) = 2.46$, $MSE_2 = 811,138.84$, $p_2 < 0.05$). Two other interactions were significant for subject analysis, but not for item analysis: between truth and block ($F_1(6,234) = 2.62$, $MSE_1 = 1,830,809.32$, $p_1 < 0.05$, $F_2(6,48) = 2.17$, $MSE_2 = 1,118,999.67$, $p_2 < 0.1$) and between trial type, truth and block ($F_1(12,468) = 1.83$, $MSE_2 = 1,351,573.47$, $p_1 < 0.05$, $F_2(12,96) = .71$, $MSE_2 = 873,536.81$). No other interactions were found significant. The latency effects are very similar to the ones in Experiment 1 — there is a $r = .88$ correlation between the conditions that overlap in Figure 2 and Figure 4.

For the first block, the latency results were consistent with the results obtained for percentage of correct answers. The ANOVA analysis showed a main effect of trial type ($F_1(2,78) = 11.70$, $MSE_1 = 2,451,692.91$, $p_1 < 0.0001$, $F_2(2,16) = 6.78$, $MSE_2 = 806,944.72$, $p_2 < 0.01$). Everywhere subjects were slowest on the artificial-word sentences. Except for the easy foils, they were fastest for the literal sentences. There was no significant effect of truth.

By the end of the experiment, the trial type effect was still present. The ANOVA performed only on the last block data showed a main effect of trial type for subjects ($F_1(2,78) = 3.23$, $MSE_1 = 460,922.43$, $p_1 < 0.05$;), which did not reach significance in the item analysis ($F_2(2,16) = .52$, $MSE_2 = 542,265.29$), and, similarly, a main effect of truth, significant only for subject analysis ($F_1(2,78) = 12.70$, $MSE_1 = 642,645.50$, $p_1 < 0.0001$, $F_2(2,16) = 2.32$, $MSE_2 = 853,285.52$), with hard foils still taking longer to understand than the other types of sentences. As for percentage correct, the trial effect was due to the artificial-word trials: the ANOVA performed only for literals and metaphors in the last block showed no significant trial effect ($F_1(1,39) = .73$, $MSE_1 = 438,655.34$).

As far as the latency slopes and intercepts, there were no significant correlations with either goodness or familiarity.

Table 7: Accuracy of definitions on a scale from 0 to 1. (Feat = Feature; Cat = Category.)

	No examples		Examples	
	Feat	Cat	Feat	Cat
Metaphors	.65	.50	.60	.57
Artificial Words	.20	.49	.34	.61

Accuracy of Word Learning. At the end of the experiment, participants had to define the words encountered in the experiment. Each definition was rated on a scale from 0 to 1. The definitions were rated on two dimensions: on how well they captured the salient feature of the meaning (e.g., for *bear*, that feature was *massive*) and on how well they captured the category (e.g. for *bear*, the category was *athlete*). If the definition contained a superordinate (e.g. for *bear*, *person*) or a subordinate (e.g. *wrestler*) of the category, we gave 0.5 credit out of the 1 maximum category score. The ratings were done twice: first, for words out of context, and second, for words occurring in a sentence. Table 7 contains average accuracy for definition ratings.

For this part of the experiment, the independent variables were (1) the unknown word type (metaphor or artificial word), (2) the presence of examples, and (3) the trait (category or feature).

The ANOVA showed main effects for word type — metaphor versus artificial word ($F_1(1,39) = 25.42$, $MSE_1 = 0.09$, $p_1 < 0.0001$, $F_2(1,8) = 49.86$, $MSE_2 = .01$, $p_2 < 0.001$), for presence of examples ($F_1(1,39) = 15.05$, $MSE_1 = 0.03$, $p_1 < 0.0005$, $F_2(1,8) = 16.14$, $MSE_2 = 0.01$, $p_2 < 0.005$). The trait effect (category versus feature) was not significant. There were also significant interactions between word type and trait ($F_1(1,39) = 23.98$, $MSE_1 = 0.11, p < 0.0001$, $F_2(1,8) = 7.75$, $MSE_2 = .07$, $p_2 < 0.05$) and between word type and the presence of examples ($F_1(1,39) = 7.75$, $MSE_1 = 0.03$, $p_1 < 0.01$, $F_2(1,8) = 7.45$, $MSE_2 = 0.01$, $p_2 < 0.05$). The three way interaction was significant by subject analysis, but only marginally significant by item analysis ($F_1(1,39) = 4.90$, $MSE_1 = .02$, $p_1 = 0.05$, $F_2(1,8) = 4.92$, $MSE_2 = 0.002$, $p_2 < 0.0575$).

In other words, subjects were more accurate at defining metaphors than artificial words, and more accurate when they were given usage examples for the words, with the artificial words benefiting more from the examples. Also, whereas the category accuracy was comparable for both metaphors and artificial words, the features were better captured for metaphors.

We also looked at the relationship between definition accuracy and familiarity and goodness for metaphors. Interestingly, the feature accuracy in the presence of examples was correlated with the goodness of the metaphor ($r = .79$, $t(8) = 3.38$, $p < 0.05$) and only marginally correlated with familiarity ($r = .67$, $t(8) = 2.36$, $p < 0.0506$). When there were no examples, the correlation between feature accuracy

and goodness was marginally significant ($r = .62$, $t(8) = 2.07$ $p < 0.1$). Thus, the better the metaphor, the greater the chance that subjects identify the correct feature for the definition. None of the other correlations were significant.

There are a couple of noteworthy results in these definitions. First, even in the presence of examples, the overall definition accuracy for artificial words was only about 47%. This result together with artificial word trials taking longer to comprehend than the other types of trials in the last block indicate that there is a difference between metaphor learning and artificial word learning. Further evidence for this hypothesis is supplied by the result that subjects were considerably worse at extracting the distinguishing feature for the artificial words than for metaphors. This result provides an example of how metaphors can facilitate the process of meaning creation: they bring attention to which features are critical to the acquisition of the meaning of the word.

Discussion

In the first trials, participants were slower and less accurate at judging metaphorical or artificial-word sentences than literal sentences. As in Experiment 1, initially, participants were biased to reject the metaphoric sentences as false, which suggests that they were first processing the metaphoric sentences literally. The initial differences in accuracy and latencies between the hard and easy foils suggests that, in an impasse due to meeting an unknown word, participants used a context matching strategy. This strategy might have also helped them to infer the meaning of the unknown word from the context.

At the end of the experiment, sentences containing words with artificial words maintained a comprehension difficulty with respect to literals and metaphors. This finding is not contradictory with the idea that the processing of literal and non-literal sentences becomes similar after repeated exposure. It may just be that metaphors indicate the salient features of the new meaning, and so the learning proceeds faster in the case of metaphors. Supporting this interpretation is the finding that the definitions for neologisms in Table 7 were worse than for metaphors in terms of the feature, but not in terms of category.

An ACT-R Computational Model of Word Learning

So far, we have sketched the general outline of our understanding of the processes occurring in the experiment. We believe that, with repeated exposure, subjects have increased probability of figuring out the components of the word meaning. Once they acquire the word meaning, there is no difference in the processing of sentences involving this word and of other sentences. If subjects do not have the meaning for the word, they fall back into special processes. For metaphors, subjects have a tendency to take a literal approach and reject the sentence. Otherwise, they tend

to judge whether the rest of the sentence matches the story.

There are certain trends in the data not addressed by this model. Subjects appear to get faster on all sentences, including literals. This means that there is some general speed up that we need to model with the rest of the data. It is also the case that in no condition are subjects responding perfectly. This fact probably reflects some mix of sloppiness in responding and failures of comprehension. Again, we need to add this to a full explanation of the data.

We have chosen the ACT-R architecture (Anderson & Lebiere, 1998) as a framework in which to develop a computational model that addresses all of the data. Our computational model uses a production system operating on the declarative knowledge base. The production system represents the rules for processing the sentences and for performing the judgement tasks. The declarative knowledge is used to represent the stories, the sentences to be judged, and the knowledge about the meanings of the words.

Figure 5 sketches out the flow of control when judging a sentence. It represents only an approximation of the process of sentence processing. (For instance, the predicate is not processed before the subject in normal sentence processing. However, this assumption is not vital for this model; it just makes the exposition simpler and saves a few extra paths in the diagram.) In Budiu and Anderson (2004) we present a detailed incremental model of sentence processing that accurately captures behavioral data from a number of domains. However, for the current purposes, we only use a rough approximation of the steps described in that model. In that model, as well as in the one in this chapter, we assume that, to judge the sentence (or comprehend it), subjects have to find a match within the preceding context.

Our model treats the sentence as consisting of a "subject" (which is potentially a metaphor or artificial word) and a "predicate", which stands for the rest of the sentence. The times shown in this flowchart can vary a little from sentence to sentence, depending on details of the ACT-R architecture. To explain this flowchart, let us first consider the situation where the model knows or has learned the meaning of the sentence subject. In this case, the model will successfully process the predicate and the subject, taking time $t_1 + t_2$. Then, it will try to retrieve a story sentence that corresponds to the probe, taking time t_3. If the retrieval fails, the model will respond *false* after taking an additional time f because of the failure (failed retrievals take longer in ACT-R). If a story sentence can be retrieved, the model will match the subject of the sentence against the subject of the story. First, the model will compare their general types, taking time t_4. If they agree (e.g. both words denote persons), and they always will agree for literal sentences, the model will continue with a more careful feature test, taking time t_5. If the more refined features match, too, the model responds *true*; otherwise, it responds *false*. Finally, in all cases, there is a time t_r to make the response. Thus, in the case when the meaning is known, the times are:

$t_1 + t_2 + t_3 + t_4 + t_5 + t_r$	for true sentences and hard foils[5]
$t_1 + t_2 + t_3 + f + t_r$	for easy foils

Now, let us consider the times for metaphors when the subject has not learned the new word meaning. The times for easy foils are the same as above. However, in the case of trues or hard foils, the predicate will enable a story sentence to be retrieved. The type of the story subject will not match the type of the sentence subject. When this happens, the model will search for a homonym, fail and either respond *false* or try to build a new meaning and respond *true*. The first branch occurs with probability $1 - p$ and the second branch occurs with probability p. Looking for a homonym takes time t_6; trying to build a metaphorical meaning takes time t_7. Thus, the response times are:

$t_1 + t_2 + t_3 + t_4 + t_6 + t_r$	for sentences treated literally
$t_1 + t_2 + t_3 + t_4 + t_6 + t_7 + t_r$	for sentences treated metaphorically

The final case concerns what happens when an artificial word is encountered and the subject has not yet learned a meaning for that word. In this, case the attempt to retrieve a subject meaning will fail, taking the extra time f (again, because retrieval failures take extra time in the model). The model will then try to retrieve the story sentence. In the case of easy foils, it will fail and respond *false*, taking time $t_3 + f$. In the other cases, the retrieval will succeed, and the model will have to guess if the sentence is true or false. We assume it will guess *true* 2/3 of the time, since this is the base rate for true sentences when the sentence predicate matches the story. This is an instance of probability matching (Lovett, 1998). It takes time t_7 to make this guess. Thus, the times are:

$t_1 + t_2 + f + t_3 + f + t_r$	for easy foils with artificial word
$t_1 + t_2 + f + t_3 + t_8 + t_r$	for trues and hard foils with artificial word

While there 10 time parameters in these equations, there are only 6 degrees of freedom, since $t_1 + t_2 + t_3 + t_r$ appears in all equations and t_4 is not independent of the other times. Thus, in creating a simulation model we estimate the parameters that controlled the intercept $t_1 + t_2 + t_3 + t_r$, t_5 (the time to match features), t_6 (the time to consider a homonym), t_7(the time to perform analogy), t_8 (the time to make a guess), and f (the time for a retrieval failure). Two parameters control the probability of correctly responding. One is p, which determines the probability of trying a metaphorical extension. The second is the q, the probability of producing the wrong response. This parameter models subjects not being perfectly accurate even in the case of literal trials or metaphoric easy foils, when all clues indicate what the correct response is. Such errors can occur because of miscomprehension, motor slips, or quick guesses.

There are two processes that control learning over time. One is the processing times speeding up as power functions of practice. Thus, if t_i is the initial time on block 1, then $t_i \cdot n^{-0.5}$ is the time on block n. This kind of learning is a basic part of the ACT-R architecture, and it produces the speed up in the literal condition. The 0.5 exponent is a constant in the ACT-R theory.

However, there is another learning process that is important in the metaphor and literal conditions, and this is the acquisition of word meanings. Learning new meanings happens either when an analogical extension is tried (see Figure 5) if the new word is metaphoric, or after the model is given feedback about an error in its answer. We assume that the subject learns the category of any new word the first time it occurs in a true sentence or in a hard foil. This is because the feedback in combination with the predicate that matches the story clearly identifies the category of the referent. On the other hand, we assume that subjects cannot learn with easy foils, because the predicate fails to identify any sentence in the story. We also assume that the feature of the metaphorical word is learned on the first appropriate trial, because the metaphor indicates which feature is critical. Since true sentences and hard foils occur on 3/4 of the trials and since the model has a $1/3 \times p + 2/3 \times (1-p) \approx 0.5$[6] chance of making an error for metaphoric trials until it learns the new meaning, there is approximately a .4 probability of learning these elements (category and feature for metaphors) in any trial. For artificial words, there is a $1/3 \times 1/3 + 2/3 \times 2/3 \approx .55$ chance of making an error, but we postulate a low probability, r, of learning the feature of an artificial word on such an error trial. There are many possible features of the referent in each story, and the artificial word gives no guidance as to how to learn the words. Learning the feature is critical because it is what enables the model to select the appropriate referent from the story. Essentially, if the feature has been learned the model can use the new word successfully for judging the sentence, and, if it hasn't been learned, the model is left to guess.

Thus, the probability parameters to be estimate are p (probability of trying a metaphorical extension), q (probability of making the unintended response), and r (probability of inducing the feature of an artificial word).

Model Predictions

None of the times or probability parameters are directly settable in ACT-R, but derive from underlying parameters controlling spreading activation and how alternative productions are evaluated. Nonetheless, we can state the approximate values the probabilities and time parameters take in the simulation. These are given in Table 8.

[6]For the hard foils (1/3 out of the 3/4), subjects will make an error if they try analogy and come up with the wrong one. For trues (2/3 out of the 3/4), subjects will make an error if they fail to make an analogy. We used an estimated value of $p = 0.43$ (see Table 8).

Table 8: Approximate parameter values for the simulation model.

Parameter	Expression	Value
Intercept	$t_1 + t_2 + t_3 + t_r$	3.25 s
Type Check	t_4	0.10 s
Feature match	t_5	0.31 s
Homonym search	t_6	2.04 s
Analogy	t_7	0.98 s
Nonword guess	t_8	1.25 s
Mismatch cost	f	0.75 s
Try analogy	p	0.43
Error	q	0.22
Induce feature	r	0.45

The reader is reminded that all the temporal parameters are not independent and therefore we have fewer degrees of freedom than parameters in the table. The reader is also reminded that the actual values can vary a bit depending on the details of the simulation. Therefore, the results we report are averaged over 200 simulation subjects.

Figure shows the predicted accuracy results. The patterns we see for true sentences in the data from Experiment 2 are reproduced by the ACT-R model: accuracy is high and stable for the easy foils; the difference among conditions almost disappears for trues, and the artificial-word condition lags behind the others for hard foils. The overall correlation between the plots in Figure and the ones in Figure 3 is $r = 0.81$. The fact that the correlation between Experiment 1 and Experiment 2 was 0.80 suggests we have accounted for most of the reliable trends in the data.

Figure shows model predictions for the times to judge the sentences. The overall correlation between the plots in Figure and the ones in Figure 4 is $r = 0.88$. The model produces very little difference between metaphors and literals for easy foils, while a large initial difference for trues and hard foils. The trues difference disappears over the trials. Except for metaphoric trues in the first block, the artificial-word trials tend to be the slowest.

The model makes predictions about the accuracy of the definitions. Table 9 shows the mean number of metaphors and artificial words with correct definitions. This reproduces the result of lower accuracy for features of artificial words. The overall level of accuracy is much higher than in Table 7 (except for the features of artificial words), but this might reflect differences in reporting and an overly stringent scoring criterion on our part.

Overall the model does a good job of accounting for the data and provides us with a computational view of how people may go about learning words and using them in context when these words are used referentially. This model uses the portion of a sentence that it does know, which in our experiment was the predicate, to identify

Table 9: Accuracy of definitions on a scale from 0 to 1 — model predictions.

	Feature	Category
Metaphors	.94	.97
Artificial words	.23	.84

the reference of the word in the past context. Then it uses the features of the reference to make hypotheses about the features of the object. It was able to learn from metaphors more rapidly because they better narrowed the relevant features. The model is quite successful, mastering the meaning of 95% of the metaphors in eight trials and mastering at least partially the meaning of 70% of the artificial words in the same time.

The model is "lazy" in its learning in that it only learns when it makes an error in the judgement task. This assumption reflects the view that word learning is driven by need. If the model had learned on every trial that it could, learning would have been faster. Still, this model paints a pretty optimistic picture of the ability of incremental word learning to succeed. Since subjects were not as accurate in their word definitions as the model (Table 7 versus Table 9), one might question whether the model accurately captures subjects' rate of word learning. However, we think the explicit definition task suffers from criterion difficulty both on our part as scorers and on the subjects' part as definition givers. In our view, the accuracy and latency in the judgement task in Figures 3 and 4 (and reproduced by the model in Figures and) are more to the point. By the end of the experiment, there are no differences between the accuracies and latencies for literals and metaphors, and the accuracy and latency differences with artificial words have been greatly reduced. Subjects were behaving as if they had mastered the new word meanings, at least in the case of metaphors.

On the other hand, one might argue that our learning situation was overly generous to the subjects in that 3/4 of trials (trues and hard foils) offered subjects an opportunity to infer the meaning of the new word. Perhaps more often when words occur in context in real life the context does not provide the same opportunities for inferring the meaning of the word. So this chapter really does not answer the question of whether it is possible for us to learn all the word meanings we do by incremental learning from context. Nonetheless, it does show that incremental word learning can be quite successful.

Figure 5 offered a picture of how words are processed when they occur in context. In the case of literals, the model just processed the sentences normally and in the case of artificial words they implemented a sophisticated guessing strategy. The most interesting case concerned the processing of metaphors. Not only did the model first try a literal interpretation, but over 50% of the time it did even go on to try a metaphoric interpretation. However, with practice the model acquired new metaphoric meanings for these words and was often retrieving them before the

literal meanings.

Conclusion

The results indicate that the context matching strategy can be successfully used for meaning creation (as theories of word learning in context assert) and comprehension. Convergence of latencies and percentages of correct answers in the literal and non-literal conditions supports the claim that participants are actually building new meanings for the artificial words and for the metaphor vehicles and use them to understand the sentences. We found no evidence for the claim that artificial-word learning is a process essentially different from metaphor learning. However, metaphorical meanings were easier to pick up: participants in our experiment gave more precise definitions for such words than for artificial words. It seems there are at least two cues people use to interpret utterances: literality and context.

It is interesting how quickly anaphoric metaphors changed their status and became understood as readily as literals. However, Experiment 1 showed that increased usage made them less appealing for further use. This result argues in favor of learned metaphors behaving like literals, if we assume that literals have less expressive power than metaphors and that people strive to achieve maximum expressivity.

Participants were able to do fairly well in comprehension without mastering completely the intended meaning of the new words. This result seems to support the view that word learning is incremental (Nagy et al., 1987) — the exact meaning is not acquired always from a single example. It is also in agreement with the claim of Landauer and Dumais (1997) that the meaning of a word is not precise, but rather determined by the variety of contexts in which it occurred.

References

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

Black, M. (1962). *Models and metaphors.* Ithaca, NY: Cornell University Press.

Black, M. (1979). More about metaphor. In A. Ortony (Ed.), *Metaphor and thought.* Cambridge University Press.

Blasko, D., & Connine, C. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of experimental psychology: learning, memory, and cognition*, *19*, 295-308.

Budiu, R., & Anderson, J. (2002). Comprehending anaphoric metaphors. *Memory & Cognition*, *30*, 158-165.

Budiu, R., & Anderson, J. (2003). Verification of sentences containing anaphoric metaphors. In *Proceedings of the 5th international conference on cognitive modeling.* Bamberg, Germany.

Budiu, R., & Anderson, J. (2004). Interpretation-based processing: A unified theory of semantic sentence comprehension. cognitive science. *Cognitive Science*, 1-44.

Cacciari, C., & Glucksberg, S. (1992). Understanding figurative language. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics.* San Diego, CA.

Carnine, D., Kameenui, E. J., & Coyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quaterly*, *19*, 188–204.

Clark, H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Dascal, M. (1987). Defending literal meaning. *Cognitive Science*, *11*, 259-281.

Fischer, U. (1994). Learning words from context and dictionaries: An experimental comparison. *Applied Psycholinguistics*, *15*, 551–574.

Gentner, D., & France, I. (1988). The verb mutability effect: studies of the combinatorial semantics of nouns and verbs. In S. Small, G. Cottrell, & M. Tanenhaus (Eds.), *Lexical ambiguity resolutions: perspectives from psycholinguistics, neuropsychology and artificial intelligence* (p. 343-382). Morgan Kauffman.

Gerrig, R. (1989). The time course of sense creation. *Memory and Cognition*, *17*, 197–207.

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.

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.

Herman, P., Anderson, R., Pearson, P., & Nagy, W. E. (1987). Incidental acquisition of word meaning from expositions with varied text features. *Reading Research Quaterly*, *22*, 263–284.

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

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *105*, 221–240.

Lovett, M. (1998). Choice. In J. Anderson & C. Lebiere (Eds.), *The atomic components of tought.* Lawrence Erlbaum Associates.

McKeown, M. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quaterly*, *20*, 482–496.

Mondria, J., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*, 249–267.

Nagy, W., Anderson, R., & Herman, P. (1987). Learning word meanings from context during normal reading. *24*, 237–270.

Nagy, W., & Genter, D. (1990). Semantic constraints on lexical categories. *Language and Cognition Processes*, *5*, 169–201.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quaterly*, *20*, 233–253.

Nagy, W., & Scott, J. A. (1990). Word schemas: Expectations about the form and meaning of new words. *Cognition and Instructions*, *7*, 105-127.

O'Brien, E., & Myers, J. (1985). When comprehension difficulty improves memory for text. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 12-21.

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.

Shefelbine, J. (1990). Student factors related to variability in learning word meanings in context. *Journal of Reading Behavior*, *22*, 71.

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

Sternberg, R., & Powell, J. (1983). Comprehending verbal comprehension. *American Psychologist*, *38*, 878–893.

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.

Townsend, J. (1974). Issues and models concerning the processing of a finite number of inputs. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (p. 133-186). Hilsdale, NJ: Lawrence Erlbaum Associates.

Ullmann, S. (1966). *Language and style.* New York: Barnes and Noble, Inc.

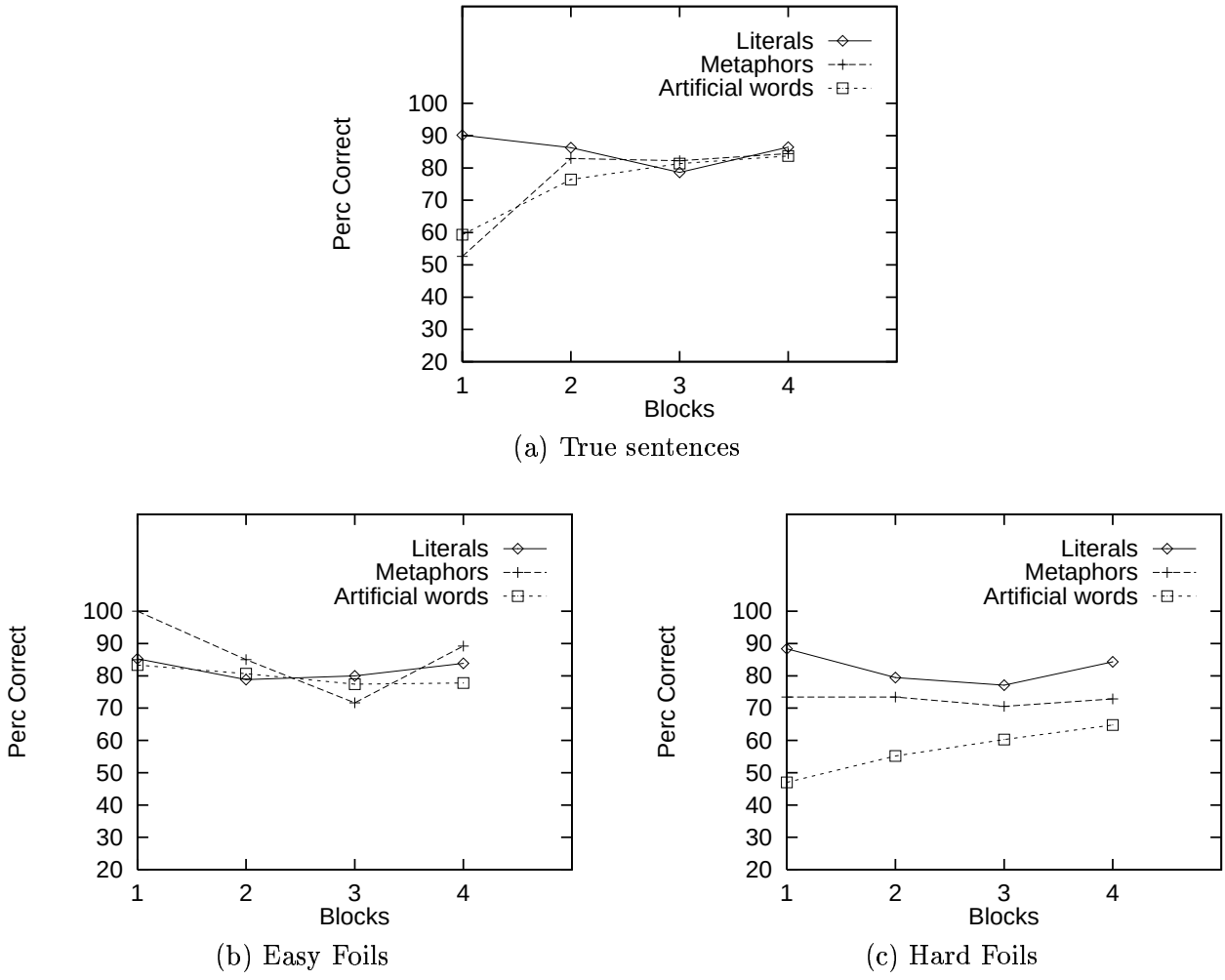
van Daalen–Kapteijns, M. M., & Elshout-Mohr, M. (1981). The acquisition of word meanings as a cognitive learning process. *Journal of Verbal Learning and Verbal Behavior*, *20*, 286–399.

Webster's 9th new collegiate dictionary. (1991). Springfield, MA: Merriam-Webster.

Werner, H., & Kaplan, E. (1950a). The acquisition of word meanings: a developmental study. *Monographs of the Society for Research in Child Development*, *15*(51).

Werner, H., & Kaplan, E. (1950b). Development of word meaning through verbal context: an experimental study. *Journal of Psychology*, *29*, 251–257.

Xiaolong, L. (1988). Effects of contextual cues on inferring and remembering meanings of words. *Applied Linguistics*, *9*, 402–413.

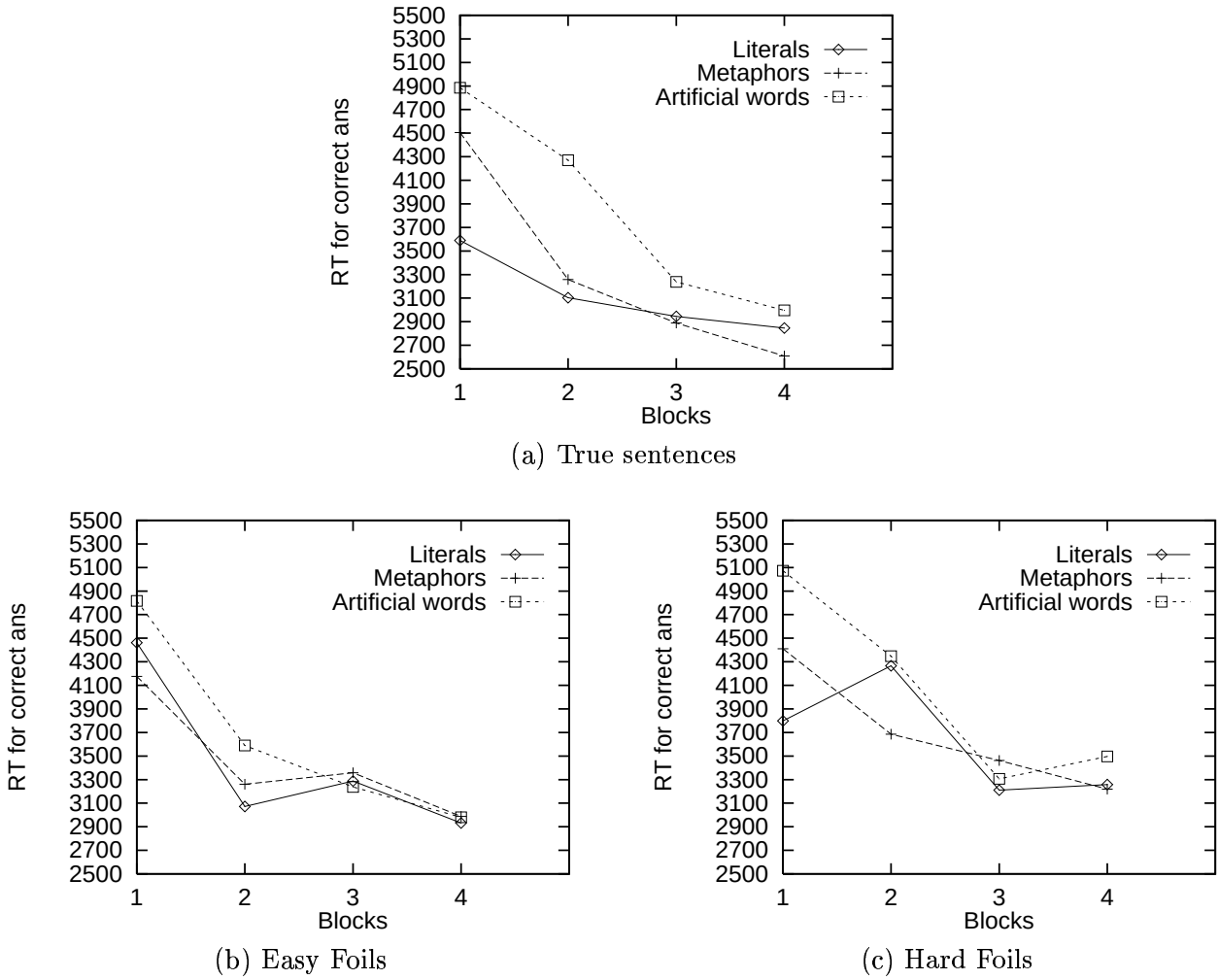


(a) True sentences

(b) Easy Foils

(c) Hard Foils

Figure 3: a) Percentage of correct answers to true sentences in Experiment 2 as a function of blocks of practice. b) Percentage of correct answers to easy false sentences in Experiment 2 as a function of blocks of practice. c) Percentage of correct answers to hard false sentences in Experiment 2 as a function of blocks of practice.



(a) True sentences

(b) Easy Foils

(c) Hard Foils

Figure 4: a) Response times for correct answers to true sentences in Experiment 2. b) Response times for correct answers to false easy sentences in Experiment 2. c) Response times for correct answers to false hard sentences in Experiment 2.

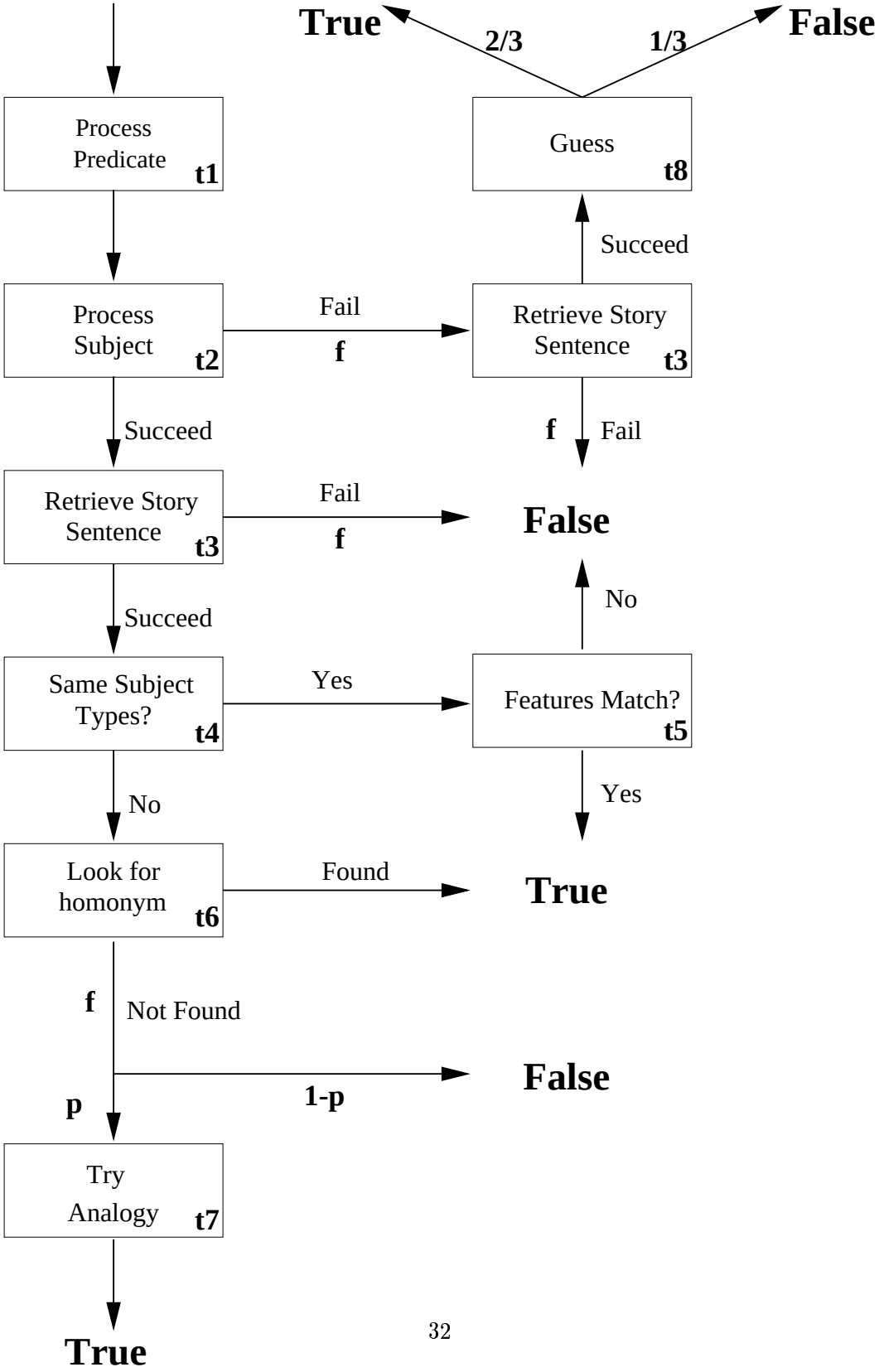
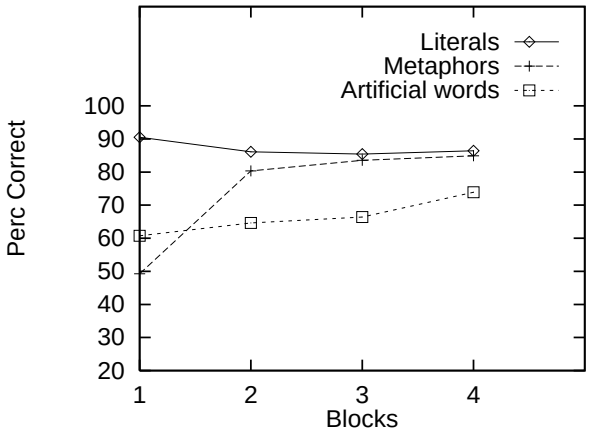
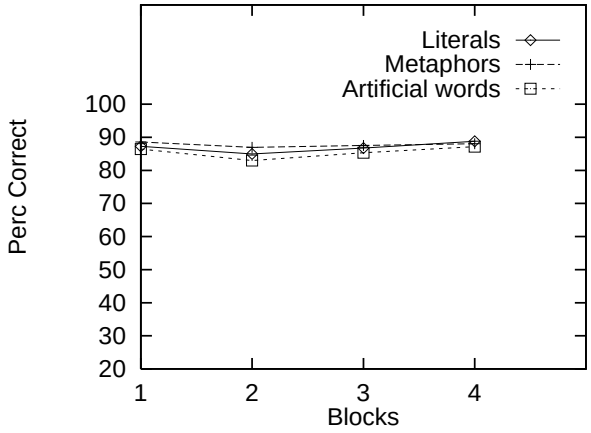


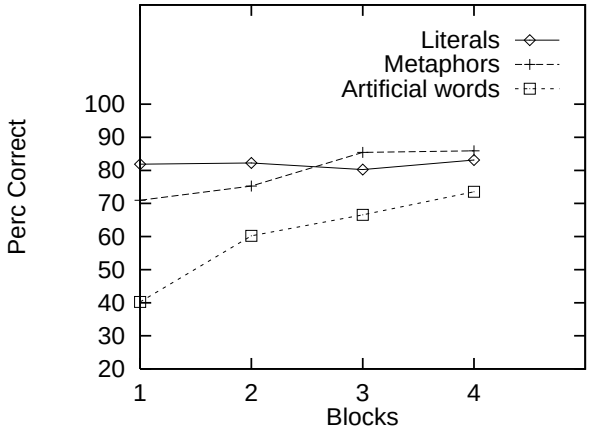
Figure 5: Flow of control for sentence judgement.



(a) True sentences

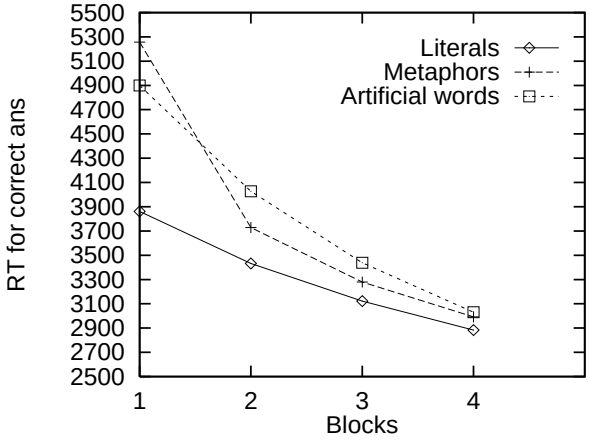


(b) Easy Foils

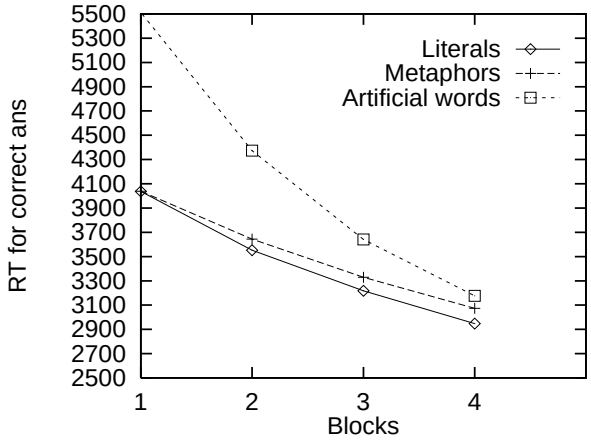


(c) Hard Foils

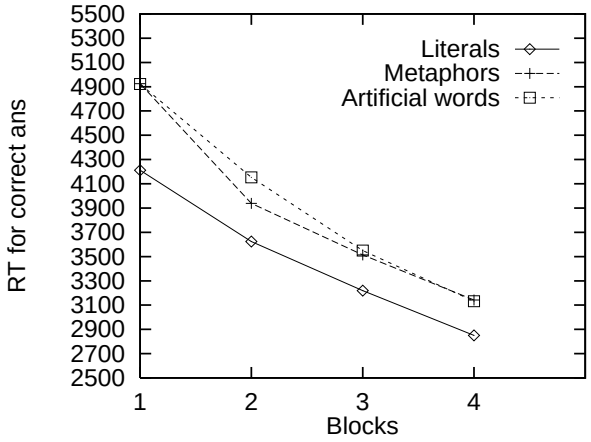
Figure 6: a) Percentage of correct answers to true sentences as a function of blocks of practice — model prediction. b) Percentage of correct answers to easy false sentences as a function of blocks of practice – model prediction. c) Percentage of correct answers to hard false sentences as a function of blocks of practice — model prediction.



(a) True sentences



(b) Easy Foils



(c) Hard Foils

Figure 7: a) Response times for correct answers to true sentences as a function of blocks of practice – model prediction. b) Response times for correct answers to false easy sentences as a function of blocks of practice – model prediction. c)Response times for correct answers to false hard sentences as a function of blocks of practice – model prediction.