Memory & Cognition
2004, 32 (2), 185-197

What plausibly affects plausibility? Concept coherence and distributional word coherence as factors influencing plausibility judgments

LOUISE CONNELL and MARK T. KEANE
University College Dublin, Dublin, Ireland

Our goal was to investigate the basis of human plausibility judgments. Previous research had suggested that plausibility is affected by two factors: concept coherence (the inferences made between parts of a discourse) and word coherence (the distributional properties of the words used). In two experiments, participants were asked to rate the plausibility of sentence pairs describing events. In the first, we manipulated concept coherence by using different inference types to link the sentences in a pair (e.g., causal or temporal). In the second, we manipulated word coherence by using latent semantic analysis, so two sentence pairs describing the same event had different distributional properties. The results showed that inference type affects plausibility; sentence pairs linked by causal inferences were rated highest, followed by attributal, temporal, and unrelated inferences. The distributional manipulations had no reliable effect on plausibility ratings. We conclude that the processes involved in rating plausibility are based on evaluating concept coherence, not word coherence.

Plausibility has the hallmarks of a phenomenon so pervasive and important that, like air, no one notices it. Time and again, many cognitive accounts appeal to the idea of plausibility without specifying its cognitive basis. Most of these accounts are grounded in operational definitions of plausibility, such as, for example, the plausibility ratings that people give to some set of stimuli. Neither do we gain much insight from dictionary definitions of the form that "something is plausible if it is apparently, seemingly, or even deceptively true."

Most theoretical statements on plausibility appeal to some idea of conceptual coherence—that some story, event, or thing is plausible if it is conceptually consistent with prior knowledge (see Collins & Michalski, 1989; Johnson-Laird, 1983). However, specifying what *consistent with prior knowledge* means in a computational model turns out to be nontrivial (see Costello & Keane, 2000, 2001). More recently, it has been suggested that plausibility may be based on the distributional properties of the words describing the scenario (Lapata, McDonald, & Keller, 1999). In this article, we examine both of these concept-level and word-level proposals in a novel paradigm for exploring plausibility judgments. But first, let us reflect on the centrality of plausibility before exploring its theoretical bases more fully.

This research was supported in part by University College Dublin, the Irish Research Council for Science, Engineering and Technology Embark Initiative, and the Irish Higher Education Authority's Multimedia Research Programme in collaboration with Media Lab Europe. The authors are grateful to Dermot Lynott, Martin Pickering, Matthew Traxler, and Lee Osterhout for valuable comments on earlier drafts of this article. We also thank the members of the UCD-TCD Cognitive Science Group for their feedback on the work. Correspondence concerning this article should be addressed to L. Connell or M. T. Keane, Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland (e-mail: louise.connell@ucd.ie or mark.keane@ucd.ie).

The Centrality of Plausibility

The centrality of plausibility is demonstrated by the diversity of cognitive acts in which it has been shown to play a role. In remembering—notably, when verbatim memory has faded—plausibility is used by people as an efficient strategy in place of more costly direct retrieval from long-term memory (Lemaire & Fayol, 1995; Reder, 1979, 1982; Reder & Ross, 1983; Reder, Wible, & Martin, 1986). In reading, people have been shown to use plausibility as a cognitive shortcut to speed parsing and resolve ambiguities (Pickering & Traxler, 1998; Speer & Clifton, 1998; Traxler & Pickering, 1996). In everyday thinking, plausible reasoning that uses prior knowledge in flexible ways appears to be commonplace (Collins & Michalski, 1989). In inductive inference, Smith, Shafir, and Osherson (1993) have shown that plausibility plays a role when familiar topics are used; that is, reasoning from the particular (e.g., *poodles can bite through wire*) to the general (e.g., *dogs can bite through wire*) depends on how plausible one finds the premise in question. In creative word combinations, Costello and Keane (2000, 2001) have argued that plausibility plays a fundamental role in constraining the interpretations produced for novel noun–noun compounds (e.g., *banana car*, *skunk squirrel*). Although these studies have demonstrated the pervasiveness of plausibility, most of them have not provided any theoretical account of its basis. Typically, they have defined plausibility operationally in terms of plausibility ratings, often using the term *plausible* inter-

changeably with other descriptions, such as *appropriate*, *sensible*, or *makes sense*. The result of this treatment of plausibility is that we have been left with few ideas as to the exact nature of the phenomenon.

Plausibility as Coherence of Concepts Based on Prior Knowledge

Although few papers have elaborated a theory of plausibility, there has been a shared view running through the literature that plausibility has something to do with the coherence of concepts as established by prior knowledge. This view holds that when people make plausibility judgments, they relate some stated target description to their prior experience, make the invited inferences in the statement, and somehow assess whether the result is a good match to what they have experienced in the past, either directly or vicariously. For example, if someone is asked to assess the plausibility of the statement *The bottle rolled off the shelf and smashed on the floor*, he or she might make the inference that the bottle rolling off the shelf *caused* it to smash on the floor. Then he or she might consider this elaborated description to be highly plausible because past experience has suggested that falling fragile things often end up breaking when they hit floors. In short, the description has a certain conceptual coherence. In contrast, if the target description was *The bottle rolled off the shelf and melted on the floor*, a causal inference could be made, but the statement seems less plausible because past experience has few cases of falling fragile objects melting on contact with floors (although a scenario can be constructed in which this could occur, such as if the room was made of metal and heated up like an oven). Stated simply, the description lacks a certain conceptual coherence.

This *concept coherence view* of plausibility has been examined by Black, Freeman, and Johnson-Laird (1986). Black et al. were concerned with showing that the plausibility/coherence of discourse was based not just on referential continuity (proposed by Kintsch & van Dijk, 1978), but also on suitable bridging inferences being found between parts of the story description. In their paradigm, they created variants of a story with reordered sentences that disrupted the causal sequence of its events, while holding referential continuity constant. People's ability to recall the resulting stories was shown to be sensitive to the amount of plausible relations present.

Although Black et al.'s (1986) study nicely illustrates the importance of concept coherence to perceived plausibility, it raises many other questions. First, could more subtle effects on plausibility judgments be shown by using descriptions that are not so radically disrupted? Reordering the sentences of a story quickly makes it nonsensical, and so its reduced plausibility is unsurprising. If sensible stories were used but their conceptual coherence somehow varied, would differences in plausibility judgments follow? Second, do all inferences have equal effects on plausibility judgments? Black et al. did not distinguish between different types of connecting relations—for example, do inferences based on attributal relations (part of, attribute of, or made of), as opposed to causal relations, differentially affect plausibility? If we can start to answer these sorts of questions, we should be able to build up a more detailed empirical picture of how concept coherence affects plausibility.

Plausibility as Coherence of Words Based on Distributional Knowledge

Apart from the widely accepted *concept coherence view*, there is another view that has emerged from the computational linguistics literature (Lapata et al., 1999), which might be glossed as the *word coherence view*. According to this view, plausibility judgments reflect the distributional patterns of word co-occurrences in the actual sentences of the described event or thing. That is, the particular relationships between words, as encoded in distributional knowledge, make certain sentences appear more plausible by virtue of the specific words used.

The distributional structure of language can be gleaned from statistical analyses of how each word is distributed in relation to others in some corpora of texts. In these analyses, a given word's relationship to every other word is represented by a contextual distribution. The contextual distribution of a word is formed by moving through the corpus and counting the frequency with which it appears with other words in its surrounding context. Thus, every word may be summarized as a vector—or point in high-dimensional space—showing the frequency with which it is associated with other lexemes in the corpus. In a similar way, a whole sentence may be represented as a single point in distributional space by merging its word points; for example, the latent semantic analysis (LSA) model (Landauer & Dumais, 1997) uses the weighted sum of constituent word vectors to denote tracts of text. In this way, two sentences containing words that occur in similar linguistic contexts (i.e., that are distributionally similar) will be positioned closer together in this space than two sentences containing words that do not share as much distributional information. For example, LSA expresses this distributional similarity as a score from $[-1, +1]$, where high scores between words reflect a large amount of shared distributional information and the mean score for randomly chosen word pairs is .02 (Kintsch, 2001). This type of distributional information[1] has been implicated in many cognitive phenomena, including the prediction of priming effects (Landauer & Dumais, 1997; Lund, Burgess, & Atchley, 1995), context effects on typicality (Connell & Ramscar, 2001), children's acquisition of syntactic categories (Redington, Chater, & Finch, 1998), synonym selection (Landauer & Dumais, 1997), and metaphor interpretation (Kintsch, 2001).

Specifically, in the case of plausibility, Lapata et al. (1999) found that the judged plausibility of adjective–noun pairs is highly correlated with the distributional similarity of the two words. For example, *strong tea* is highly plausible, whereas *powerful tea* is not, and *tea* is more distributionally similar to *strong* than to *powerful*. However, as in the case of concept coherence, this work

raises more questions than it answers. First, do these effects generalize beyond very simple descriptions (like adjective–noun pairs)? Would they also hold for whole sentences? Second, is distributional information sufficient in itself to explain plausibility? Could it, for instance, remove the need for an explanation based on conceptual coherence? Third, if the word coherence account does not supplant the concept coherence one, do the two types of coherence combine to influence plausibility? It is these and other questions to which we now will turn in the remainder of the article.

Outline of Present Work

In this article, we provide the first systematic investigation of both the concept coherence and the word coherence views of plausibility. In our experimental paradigm, people are asked to rate the plausibility of two-sentence descriptions of various events (e.g., *The bottle fell off the shelf/The bottle smashed*). These descriptions are much more complex than adjective–noun pairs and, therefore, allow us to test the generality of word distribution effects. Their complexity also allows us to examine more subtle effects of conceptual coherence—that is, whether different classes of inferences impact plausibility differentially (note that we include both coreference and bridging inferences [Haviland & Clark, 1974] in our definition of inference). For example, the description *The bottle fell off the shelf/The bottle smashed* should invite a causal inference, whereas *The bottle fell off the shelf/ The bottle was pretty* should invite the inference that the *bottle* in the first sentence has the attribute *pretty*. Finally, since both word coherence and concept coherence can be varied in these descriptions, we can explore the effects of both factors on plausibility.

Two experiments are reported in which these questions were examined. In Experiment 1, we varied the concept coherence of event descriptions (via their inferences) while holding the word coherence of the descriptions constant. In Experiment 2, we selected a key concept coherence manipulation (attributal vs. causal) and crossed it with a manipulation of word coherence (strong vs. weak) to examine interactions between these two factors.

EXPERIMENT 1
Plausibility and Inference Types

In this experiment, the word coherence of the event descriptions (as measured by LSA scores) was held constant, and the types of inferences invited by the sentence pairs were manipulated (see Table 1). Four distinct categories of sentence pairs were used: causal, attributal, temporal, and unrelated. From a purely concept coherence viewpoint, we see plausibility as being about the fit between a given scenario and prior knowledge. This fit is influenced by many different factors, including the nature of the inferences made, the number of inferences made, the length of the inferential chains constructed, the amount of conjecture in the inference, the existence of possible alternative inferences, and so on (see Connell & Keane, 2003, for a computational model of such specific factors). For the explanatory purposes of this experiment, we can gloss these diverse factors as the strength of the inferential connection between the two sentences. Broadly speaking, if the scenario described in the two sentences is strongly connected, as determined by prior knowledge, the scenario will appear more plausible. In contrast, if the scenario described in the two sentences is weakly connected, the scenario will appear less plausible. This simple explanatory model allows us to make certain predictions for the four conditions in this experiment; specifically, we can predict a decreasing trend in perceived plausibility, with the following ordering: causal > attributal > temporal > unrelated (see Table 1 for sample materials). Let us consider how we arrived at these predictions.

First, we assumed that the greater the number of inferences that need to be constructed to connect the two sentences, the weaker the actual connection between them will be. In this respect, the causal and the attributal conditions are very strongly connected, in that a single inference connects both sentences. For example, in our causal scenario in Table 1, a single causal inference is found, using prior knowledge, to link the bottle's *falling* to its *breaking*. Also, in our attributal scenario, a simple referential inference is made to establish *prettiness* as a property of the falling bottle. On the other hand, the temporal and the unrelated conditions are less strongly connected, because they require many inferences to be made. For example, to establish the inference in our temporal scenario, one needs to construct the conditions under which the bottle might come to *sparkle* after it fell (e.g., its falling into a beam of sunlight, or a light passing across it on the floor). In our unrelated scenario, the two sentences have an even weaker connection, since they can be connected only by many interdependent inferences and the assumption of many additional conditions—for example,

Table 1
Sample of Sentence Pairs Used in Experiment 1

Sentence 1	Sentence 2 (Repeated Noun)	Sentence 2 (Alternate Noun)	Inference Type
The bottle fell off the shelf.	The bottle smashed.	The glass smashed.	causal
	The bottle was pretty.	The glass was pretty.	attributal
	The bottle sparkled.	The glass sparkled.	temporal
	The bottle melted.	The glass melted.	unrelated

one could assume that the room had a metal floor, that the floor had been heated to a very high temperature, and that when the falling bottle made contact with the floor, it *melted*. Indeed, the weakest connection of all—namely, when no connection can be found—is likely to occur in this unrelated condition when people fail to make the inferential leaps required. Such additional conditions do not need to be established in the causal condition, since the falling of fragile objects is known to lead typically to their smashing. Nor do these conditions need to be established in the attributal condition, where the direct assertion of a property of the bottle is noted as a simple putative fact. These proposals give us the following predicted ordering: (causal = attributal) > temporal > unrelated, with plausibility decreasing from left to right.

Second, the strength of the connection between the sentences in the causal and attributal conditions is distinguished in an additional way—namely, by the informativeness of the single inferred relation. It is well known from work in other areas that causal relations (*cause* or *enable*) are much more informative than constitutive relations (*has* or *made of*; see, e.g., Keane, 1997; Keane, Ledgeway, & Duff, 1994; Mayer & Bromage, 1980). A causal inference tells us about the specific contingency between the event described in the first sentence and that in the second. By contrast, a constitutive inference simply tells us about a property of an object described in the first sentence. The strength of the attributal connection is, therefore, weaker by virtue of the informativeness of the inference. Hence, the sentence pairs in the causal condition should be viewed as more plausible than those in the attributal condition, giving the following predicted ordering: causal > attributal > temporal > unrelated, with decreasing plausibility.

So this simple explanatory model using the strength of the connection between the two sentences gives us the predicted ordering of decreasing plausibility across the four conditions. There obviously exists a more complex explanatory model of the effects that unpacks these ideas in terms of many different knowledge variables, but that is beyond the scope of this article (see Connell & Keane, 2003, for details).

Method

Materials and Design. Twelve basic sentence pairs were created and then modified to produce variants of the different materials. In each case, the second sentence was modified to produce causal, attributal, temporal, and unrelated pairs of sentences (see Appendix A), where the unrelated pairs provided a control in which no obvious inferences could be made. The causal pairs were designed to invite a causal inference by using a second sentence that was a reasonably direct causal consequence of the first sentence (e.g., *The bottle fell off the shelf/The bottle smashed*). The attributal pairs invited an attributal inference by using a second sentence that referred to an attribute of its subject in a way that was not causally related to the first sentence (e.g., *The bottle fell off the shelf/The bottle was pretty*). The temporal pairs invited a temporal inference by using a second sentence that could occur in the normal course of events, regardless of the occurrence of the first sentence (e.g., *The bottle fell off the shelf/The bottle sparkled*). The unrelated pairs used a second sentence that described an event that was unlikely to occur in the normal course of events and had no obvious causal link to the first sentence (e.g., *The bottle fell off the shelf/The bottle melted*).

In addition, the second sentence of each of these four pairs was modified to use either the same object as the first sentence (e.g., *bottle/bottle*) or something belonging to that object (e.g., *bottle/glass*, *cup/handle*). This manipulation was done to examine whether the repetition of terms would facilitate the participants' ability to construct inferences between the two sentences in the pair. Thus, the sentence pairs captured two within-subjects variables: *inference type* (causal, attributal, temporal, or unrelated) and *noun type* (repeated or alternate).

The word coherence of each sentence pair was controlled by comparing their distributional scores with LSA (Landauer & Dumais, 1997). All the LSA comparisons in these experiments were performed using "general reading up to 1st-year college" semantic space, with document-to-document comparison at maximum factors. This means that the LSA corpus used represents the cumulative lifetime readings of an American 1st-year university student and that the LSA scores were calculated as the distance between sentence points (which are calculated as the weighted sum of constituent word vectors). An analysis of variance (ANOVA) of the distributional scores of the sentence pairs revealed a significant difference between noun types [repeated $M = 0.72$, alternate $M = 0.29$; $F(1,88) = 217.780$, $MS_e = 0.020$, $p < .0001$], as was expected because repeated terms boost scores in LSA. However, there was no difference between inference types [causal $M = 0.52$, attributal $M = 0.48$, temporal $M = 0.51$, unrelated $M = 0.51$; $F(3,88) = 0.358$, $MS_e = 0.020$, $p > .7$; confirmed by pairwise comparisons using Bonferroni adjustments, all ps $> .9$] and no interaction between noun type and inference type [$F(3,88) = 0.029$, $MS_e = 0.020$, $p > .9$].

Word frequency was also controlled for by using British National Corpus (BNC) word frequency counts. The BNC's part-of-speech tags ensured that only word counts that corresponded syntactically with the sentence were used (e.g., *The branch fell* excluded the counts for *fell* in adjectival or nominal form). The accepted part-of-speech tags were *nn1* or *nn2* for nouns, *vvd* for causal, temporal, or unrelated verbs, and *aj0* for attributal adjectives. Ambiguous tags (e.g., *aj0–vvd*) were accepted and counted. An ANOVA of the frequency scores showed no reliable difference between the inference types [causal $M = 1{,}462.5$, attributal $M = 3{,}020.7$, temporal $M = 870.0$, unrelated $M = 1{,}025.9$; $F(3,44) = 2.007$, $MS_e = 5{,}778{,}177$, $p > .1$; with Bonferroni adjustments, all pairwise comparison ps $> .2$] or between noun types [repeated $M = 6{,}941.5$, alternate $M = 5{,}833.2$; $F(1,22) = 0.110$, $MS_e = 67{,}220{,}123$, $p > .7$].

The materials, 96 sentence pairs in all, were split into eight groups of 12 pairs apiece, selected to avoid repetition of nouns, verbs, or adjectives across the pairs. Each group contained 3 sentence pairs per inference type, counterbalanced between repeated nouns and alternate nouns. All 12 sentence pairs within each group were presented in a random order, resampled for each participant.

Participants. Forty native speakers of English were randomly assigned to the different groups in the experiment. All the participants were student volunteers at University College Dublin.

Procedure. The participants read instructions that explained the 0–10 plausibility scale (0 being *not at all plausible* and 10 being *highly plausible*) with an example of the sentence pairs—a causal pair that was not featured in the experiment [*The car rolled down the hill/The car skidded*]. They were asked to take their time over each decision and not to alter any answers already marked down. Each sentence pair was presented on a separate page with a marked space for the participants to note their 0–10 plausibility rating.

Results and Discussion

The results show that plausibility is affected by subtle changes in the conceptual coherence of simple event descriptions when different inferences are invited (see Fig-

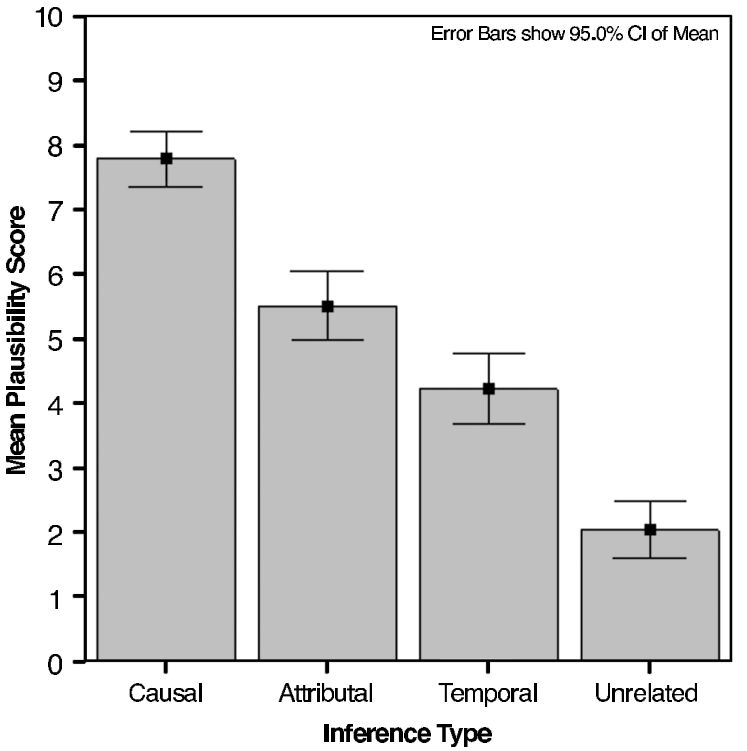

Figure 1. Mean plausibility ratings per inference type.

ure 1). Importantly, these plausibility differences are purely due to concept coherence and occur when word coherence is held constant (i.e., the distributional scores of sentence pairs were held constant across the different inference types). Table 2 gives the mean ratings for each condition: The causal pairs were rated the most plausible ($M = 7.8$), followed, as was predicted, by attributal ($M = 5.5$), temporal ($M = 4.2$), and unrelated ($M = 2.0$) pairs. The results support the traditional view that concept coherence is important in plausibility judgments, with the added finding that different inference types differentially affect plausibility.

All ANOVAs by participants and by items were performed by treating participants (F_1) and sentences (F_2), respectively, as random factors. A two-way ANOVA by inference type and noun type showed a significant effect of inference type on plausibility scores [$F_1(3,117) = 84.57$, $MS_e = 7.5721$, $p < .0001$; $F_2(3,44) = 41.60$, $MS_e = 16.746$, $p < .0001$]. Planned pairwise comparisons revealed that all of the conditions were reliably different from one another, using Bonferroni adjustments ($ps < .05$). No effect of noun type was found [$F_1(1,19) = 0.83$, $MS_e = 4.827$, $p > .3$; $F_2(1,44) = 0.29$, $MS_e = 7.471$, $p > .5$]. Also, no significant interaction between inference type and noun type was found [$F_1(3,117) = 0.36$, $MS_e = 5.522$, $p > .7$; $F_2(3,44) = 0.73$, $MS_e = 7.471$, $p > .5$], showing that repeating a term between the first and the second sentences in a pair did not affect the participants' ability to construct inferences between them. As was predicted, the order of inference types (causal > attributal > temporal > unrelated) was found to be reliable, using Page's trend test by participants [$L(40) = 1{,}147$, $p < .0001$] and by items [$L(12) = 341.5$, $p < .0001$].

But how can we be confident that the different sentence pairs were indeed treated in accordance with the experimenter-defined categories?[2] For example, if a temporal pair was interpreted using a causal relation, the

Table 2
Mean Plausibility Scores for Each Condition in Experiment 1

	Inference Type				
Noun Type	Causal	Attributal	Temporal	Unrelated	*Overall Mean*
Repeated	7.85	5.62	4.15	1.67	4.82
Alternate	7.73	5.40	4.28	2.40	4.95
Overall mean	7.79	5.51	4.22	2.03	4.89

plausibility ratings for temporal pairs could have been artificially inflated. In order to test this possibility, we gave four independent raters descriptions of each inference type (detailed above in the Materials and Design section) and asked them to isolate the type of relation they understood to exist between the two sentences of each pair. A sentence pair was judged to have been appropriately classified (e.g., as causal or temporal) if there was 3/4 or 4/4 agreement with the original classification of the pair. Of the 96 original sentence pairs, 72 met the criterion. A reanalysis of the data for these 72 sentence pairs confirmed the original findings, with plausibility ratings being highest for causal pairs ($M = 7.8$), followed by attributal ($M = 5.4$), temporal ($M = 3.1$), and unrelated ($M = 2.2$) pairs. Again, as before, there was a significant effect of inference type [$F_1(3,117) = 124.98$, $MS_e = 5.538$, $p < .0001$; $F_2(3,32) = 21.02$, $MS_e = 4.503$, $p < .0001$], no effect of noun type [repeated $M = 4.6$, alternate $M = 4.7$; $F_1(1,39) = 0.25$, $MS_e = 3.649$, $p > .6$; $F_2(1,32) = 0.05$, $MS_e = 4.216$, $p > .8$], and no reliable interaction [$F_1(3,117) = 0.54$, $MS_e = 5.075$, $p > .6$; $F_2(3,32) = 0.37$, $MS_e = 4.216$, $p > .7$].

To summarize, these results show that plausibility judgments are sensitive to the type of conceptual coherence established in event descriptions when different inferences are made. The strength of the connection between the two sentences and, hence, its perceived plausibility are greatest in the causal pairs, where a single direct informative inference can be made. The connection strength and, hence, plausibility are lowest in the unrelated pairs, where extensive inferences and assumptions have to be made to form the connection (which, indeed, may fail to be formed at all). Ranged in between are the attributal and the temporal pairs, distinguished largely by the amount of inferencing and additional conditions that need to be assumed. In reality, the sources of these effects are much more complex than is given credit in this simple explanatory model. Connell and Keane (2003) have developed a computational simulation of these results in which the conditions are distinguished by many different knowledge variables. These variables interact to produce different calculated levels of plausibility that replicate the human judgments reported here.

On the basis of previous research, one might not be surprised to learn that the plausibility of the causal pairs was higher than that of the unrelated pairs, although we know of no previous study in which such a manipulation has been explicitly examined in a direct and controlled manner. The main novelty of the present experiment is the demonstration that there are four distinguishable empirical categories (causal, attributal, temporal, and unrelated) that can be ranged in terms of their impact on plausibility. This result has not been shown before. Furthermore, we can be confident that these effects are due specifically to concept coherence, and not to the possible effects of word coherence, to which we now will turn with the next experiment.

EXPERIMENT 2
Plausibility and Distributional Strength

In Experiment 1, we concentrated on the role of concept coherence in plausibility, controlling for the possible influence of word coherence. In the second experiment, we examined word coherence by crossing the variables of inference type (causal or attributal) and distributional strength (strong or weak). For example, although the two causal sentence pairs in Table 3 have essentially the same meaning, they differ markedly in their distributional scores. From a word coherence perspective, *growled* is much more likely to occur with the words in the first sentence than is *snarled*, even though, from a concept coherence perspective, both sentences invite the same causal inference. So these two test items vary distributional information while holding the inference type constant (see Table 3 for an example of attributal sentence pairs).

As in Experiment 1, we expected the causal descriptions to be rated as more plausible than the attributal pairs, because of their greater informativeness. Previous research on the role of distributional information would lead us to predict that the distributionally strong descriptions would be rated as more plausible than the distributionally weak descriptions (Lapata et al., 1999). However, it should be remembered that several studies outside the area of plausibility have shown that when target tasks involve additional inferences, the predictive power of LSA is reduced. For example, Lynott and Ramscar (2001) have found that while the interpretation of noun–noun compounds is aided by distributional information, it must be supported by more complex relation-building processes. French and Labiouse (2002) also have pointed out that distributional information is inadequate in interpreting analogies such as *John is a real wolf with the ladies*, which require mapping relational information about predator–prey interactions, rather than attributional information about long gray hair and

Table 3
Sample of Sentence Pairs Used in Experiment 2

Sentence 1	Sentence 2	Inference Type	Distributional Link Strength	LSA Score
The pack saw the fox.	The hounds growled.	causal	strong	.37
	The hounds snarled.	causal	weak	.20
	The hounds were fierce.	attributal	strong	.19
	The hounds were vicious.	attributal	weak	.12

sharp teeth. Similarly, Glenberg and Robertson (2000) found that distributional information did not distinguish between sensible novel situations (such as using a sweater filled with leaves as a pillow) and nonsensical novel situations (such as using a sweater filled with water as a pillow).

Method

Materials and Design. Fifteen basic sentence pairs were created and then modified to produce variants of the different materials. As in Experiment 1, several causal and attributal variants were produced, each of which maintained the basic meaning of the original sentence pair (see Appendix B). By using LSA, the highest and lowest scoring pairs were selected as the strong and the weak distributional pairs, respectively. The causal sentence pairs had a mean LSA score of .42 for strong pairs and .27 for weak pairs, whereas the attributal sentence pairs had a mean score of .35 for strong pairs and .23 for weak pairs. A two-way ANOVA showed a reliable difference between strong and weak scores of these materials [strong $M = .39$, weak $M = .25$; $F(1,56) = 13.543$, $MS_e = 0.020$, $p < .001$]. There was no reliable difference in the LSA scores for the different inference types of the pairs [causal $M = .34$, attributal $M = .29$; $F(1,56) = 2.067$, $MS_e = 0.020$, $p > .15$] and no reliable interaction between inference type and distributional strength [$F(1,56) = 0.060$, $MS_e = 0.020$, $p > .8$]. Thus, the sentence pairs captured the between-subjects variable of inference type (causal or attributal) and the within-subjects variable of distributional link strength (strong or weak).

Frequency was controlled as in Experiment 1. Causal pairs had a mean frequency of 1,327 for strong pairs and 2,517 for weak pairs, whereas attributal pairs had a mean frequency of 2,909 for strong pairs and 16,285 for weak pairs. There was no difference between the distributional strength pairs [$F(1,56) = 2.508$, $MS_e = 317{,}171{,}038$, $p > .1$], no difference between inference types [$F(1,56) = 2.786$, $MS_e = 317{,}171{,}038$, $p > .1$], and no significant interaction between inference type and distributional strength [$F(1,56) = 1.756$, $MS_e = 317{,}171{,}038$, $p > .15$].

In addition, a pretest examined whether the basic meaning of the second sentence was maintained in the strong and the weak variants. A group of 18 participants not used in any other experiment was asked to rate the appropriateness of the terms in the second sentence (e.g., the appropriateness of using *growled* in a sentence with *hounds*). On a scale from 1 (*not appropriate*) to 7 (*very appropriate*), this pretest showed little difference between the strong and the weak versions for noun/verb appropriateness in the causal pairs (strong $M = 5.8$, weak $M = 6.0$) or noun/adjective appropriateness in the attributal pairs (strong $M = 6.0$, weak $M = 5.8$). A two-way ANOVA of the appropriateness ratings confirmed that there was no effect of either distributional strength [$F_1(1,16) = 0.05$, $MS_e = 1.821$, $p > .8$; $F_2(1,29) = 0.01$, $MS_e = 2.505$, $p > .9$] or inference type [$F_1(1,16) = 0.01$, $MS_e = 8.738$, $p > .9$; $F_2(1,29) = 0.03$, $MS_e = 5.511$, $p > .8$]. The interaction of the factors also was not significant [$F_1(1,16) = 2.38$, $MS_e = 1.821$, $p > .1$; $F_2(1,29) = 1.74$, $MS_e = 2.505$, $p > .15$].

The materials, 60 sentence pairs in all, were split by inference type into 30 causal and 30 attributal sentence pairs (i.e., strong and weak versions of 15 basic sentence pairs). The matched strong and weak sentence pairs (e.g., *The pack saw the fox/The hounds growled* and *The pack saw the fox/The hounds snarled*, respectively) were presented together on each page. Two groups were formed per inference type: for each set of 2 matched sentence pairs, one group received a strong/weak order of presentation, whereas the other received a weak/strong order of presentation, and this was alternated for each of the 15 matched sets. All 15 matched sets within each group were presented in a random order, resampled for each participant.

Participants. Twenty-four native speakers of English were randomly assigned to a materials group. All the participants were volunteers at postgraduate level at University College Dublin. One participant was excluded from the causal inference group for failing to complete the experiment.

Procedure. The participants read instructions that explained the 0–10 plausibility scale (0 being *not plausible*, 5 being *moderately plausible*, and 10 being *very plausible*), with examples of the type of sentence pairs (using pairs not featured in the experiment, appropriate to the inference type of the group). Those in the causal group saw the strong pair *The chef poured the stew/The gravy dripped*, followed by the weak pair *The chef poured the stew/The gravy dribbled*. The attributal group saw the strong pair *The chef poured the stew/The gravy was delicious*, followed by the weak pair *The chef poured the stew/The gravy was tasty*. The participants were asked, if they found one sentence pair more plausible than the other, to make certain that the scores reflected this fact. One strong and one weak sentence pair (from the same matched set) were presented per page, each with the scale for the participants to circle their plausibility rating. The position of the pairs on the page relative to one another (i.e., strong above or below) was counterbalanced in the experiment.

Results and Discussion

The results replicated the causal–attributal effect found in Experiment 1, confirming the role of concept coherence, but showed no reliable effect of word coherence (see Table 4).

As before, a main effect of inference type was found, with the causal sentence pairs yielding higher plausibility scores than the attributal pairs in the two-factor mixed design ANOVA, though the by-participants analysis was outside significance [$F_1(1,21) = 1.17$, $MS_e = 35.978$, $p > .2$; $F_2(1,672) = 8.08$, $MS_e = 5.217$, $p < .005$]. We also performed a trend analysis of causal ratings against attributal ratings. Page's trend test confirmed that causal sentence pairs were reliably rated higher than attributal sentence pairs both by participants [$L(23) = 110.5$, $p < .005$] and by items [$L(15) = 71$, $p < .00001$].

In contrast, the main effect of distributional strength was not reliable across analyses [$F_1(1,21) = 7.26$, $MS_e = 1.516$, $p < .05$; $F_2(1,28) = 0.86$, $MS_e = 12.789$, $p > .3$]. In examining each inference type separately, planned comparisons showed that the distributional effect was insignificant for causal sentence pairs [$F_1(1,10) = 1.79$, $MS_e = 1.840$, $p > .2$; $F_2(1,14) = 0.65$, $MS_e = 15.664$, $p > .6$] and insignificant by items for attributal sentence pairs [$F_1(1,11) = 6.88$, $MS_e = 1.221$, $p < .05$; $F_2(1,14) = 0.84$, $MS_e = 9.944$, $p > .3$]. Indeed, the direction of the difference in the means was opposite to that predicted from previous work (i.e., weak was rated more plausible than strong; see Lapata et al., 1999). No reliable interaction

Table 4
Mean Plausibility Scores for Each Condition in Experiment 2

Distributional Link Strength	Inference Type		
	Causal	Attributal	*Overall Mean*
Strong	7.37	6.82	7.08
Weak	7.57	7.13	7.34
Overall mean	7.47	6.96	

between inference type and distributional strength was found [$F_1(1,21) = 0.32, MS_e = 1.516, p > .5$; $F_2(1,28) = 0.04, MS_e = 12.789, p > .8$].

It could be argued that the failure to find a reliable effect of distributional strength was due to a poor rendering of the difference between the strong and the weak distributional pairs. Some strong–weak materials were further apart than others in terms of their distributional strength (i.e., their LSA scores). However, recall that our pretests of the materials showed that the two conditions were reliably different in distributional strength. Moreover, regression analysis showed that the size of this difference between the strong–weak variants had little effect on the differences in plausibility ratings that were provided by the participants (adjusted $r^2 = -.003, p > .7$).

We also carried out a more stringent test of this issue. We grouped the sentence pairs according to how extreme their LSA differences were and reran the analyses. We divided the range of LSA score differences into thirds at the 33rd and 67th percentiles (i.e., forming three groups with increasingly extreme LSA difference between their strong and their weak forms) and examined the participants' scores for those sentence pairs only. Table 5 shows the mean LSA scores for these groups and their respective plausibility scores in each condition. The results confirmed our earlier findings that there was no significant difference between strong and weak sentence pairs, with distributional strength failing to achieve significance even in the group with the most extreme LSA score differences [least extreme group, $F_1(1,21) = 2.53, MS_e = 2.507, p > .1$; $F_2(1,11) = 1.17, MS_e = 5.398, p > .3$; mid-extreme group, $F_1(1,21) = 0.07, MS_e = 3.049, p > .7$; $F_2(1,6) = 0.01, MS_e = 29.372, p > .9$; most extreme group, $F_1(1,21) = 4.03, MS_e = 3.412, p > .05$; $F_2(1,7) = 0.88, MS_e = 15.546, p > .7$]. This analysis was just one of many we carried out. None of these analyses showed any obvious grouping of the materials set along the strong–weak dimension that generated robust differences in plausibility ratings.

Furthermore, the present experiment replicated a null effect for distributional strength that we had found in an earlier, unpublished study. This early experiment had an identical procedure but differed in that participants saw *either* the strong or the weak variant of each sentence pair, rather than both being printed together on the same page. The use of the strong/weak variant for each sentence pair was counterbalanced in the experiment, and the order of presentation of the materials was randomized for each participant. This experiment, which involved 24 student participants, showed no effect of distributional strength [strong $M = 6.6$, weak $M = 6.7$; $F_1(1,21) = 0.03, MS_e = 10.593, p > .8$; $F_2(1,30) = 0.04, MS_e = 7.865, p > .8$]. Upon dividing the materials into the same extreme groups outlined above, our reanalyses of this earlier experiment also confirmed that there was no significant effect of distributional strength to be found [least extreme group, strong $M = 7.0$, weak $M = 7.1$; $F_1(1,21) = 0.07, MS_e = 10.579, p > .7$; $F_2(1,12) = 0.13, MS_e = 4.859, p > .7$; mid-extreme group, strong $M = 6.7$, weak $M = 6.9$; $F_1(1,21) = 0.04, MS_e = 8.133, p > .8$; $F_2(1,7) = 0.23, MS_e = 9.413, p > .6$; most extreme group, strong $M = 6.0$, weak $M = 5.8$; $F_1(1,21) = 0.001, MS_e = 10.791, p > .9$; $F_2(1,7) = 0.26, MS_e = 13.51, p > .6$].

In conclusion, the present experiment replicated the concept coherence effect found in Experiment 1 but showed no reliable effect of word coherence. The results of further analyses verified that even the largest distributional strength manipulations do not reliably affect plausibility ratings. As such, we believe that it is safe to conclude that word coherence has no reliable effect on the judged plausibility of event descriptions. In short, the effects found by Lapata et al. (1999) for adjective–noun pairs do not generalize to more complex discourse.

GENERAL DISCUSSION

There are two novel findings in the work reported here. First, we have established not only that concept coherence plays a role in plausibility, but also that different types of inference have different effects on plausibility: Sentences with no obvious inferential link between them are rated as barely plausible, with temporal, attributal, and causal inferences ranged in increasing plausi-

Table 5
Mean Plausibility Scores for Each Condition in Experiment 2, Subclassified by the Extent of the Difference in Distributional Scores

LSA Score Difference	Distributional Link Strength	Inference Type		
		Causal	Attributal	*Overall Mean*
Least extreme	strong	7.31	7.19	7.25
($M = .09, N = 26$)	weak	7.84	7.01	7.44
Mid-extreme	strong	7.44	6.58	6.94
($M = .13, N = 16$)	weak	7.06	7.07	7.06
Most extreme	strong	7.38	6.56	7.00
($M = .22, N = 18$)	weak	7.49	7.38	7.44

Note—*M* represents the mean difference of strong minus weak latent semantic analysis (LSA) scores for the *N* sentence pairs in that category. Different *N*s per category result from tied LSA score differences.

bility. Second, we have shown that word coherence does not appear to play a role in the rating of plausibility. How can we relate these findings to those in the previous literature, and what constraints do they generate for future research into plausibility?

Some parallels may be drawn between word and concept coherence as described in this article and the local and global context models of discourse comprehension. Local context models propose that words act individually or in combination to affect subsequent words (e.g., by priming; see Duffy, Henderson, & Morris, 1989). These local context effects are alignable with word coherence and can be seen as the product of distributional knowledge; the distributional information activated in the vicinity of one word may include the subsequently presented word. Indeed, models of distributional knowledge have already been shown to predict priming effects (Landauer & Dumais, 1997; Lund et al., 1995). In contrast, global context models propose that processing is facilitated by ongoing discourse representations above the word level (Hess, Foss, & Carroll, 1995). These global context effects are alignable with concept coherence and can be seen as the product of conceptual knowledge; the conceptual model built around a description may prime related information. In support, it has been shown that priming effects do result from conceptual discourse models (Hess et al., 1995) and that these discourse models include knowledge drawn in by causal inferences, as well as information given in the description (Halldorson & Singer, 2002).

Regarding previous research from the concept coherence view of plausibility, the results presented here offer some interesting extensions of earlier findings. Black et al. (1986) have shown that varying the number of possible inferences that could be drawn between sentences in a story has an effect on plausibility judgments. We have demonstrated that it is not only the number, but also the *type* of inference that is important to plausibility. Causal inferences are found to be the most plausible because they provide the strongest concept coherence. Then, in decreasing order of concept coherence and, thus, plausibility are attributal, temporal, and unrelated (i.e., no relation at all) inferences. This finding emphasizes that plausibility is affected not just by the presence of inferences, but also by the type of inference in question. In other words, it is not just the presence of global context that is important to plausibility, but the actual knowledge drawn in as each of the inferences is made. Connell and Keane (2003) used this principle in their computational model of plausibility. Each particular inference type draws in different elements and levels of world knowledge, which allows the model to calculate concept coherence (and therefore, plausibility) by analyzing the resulting representation.

With regard to the word coherence view, the present results suggest that Lapata et al.'s (1999) findings are limited to adjective–noun combinations. Lapata et al. gave their participants simple adjective–noun pairs, which by their nature provide local, but not global, contexts. It has been proposed by Hess et al. (1995) that a local context is useful only when people are given no global context, which would suggest that Lapata et al.'s participants based their plausibility ratings on the only thing they had—the distributional information of the local context. This represents a simple situation in which word coherence is the sole basis of the plausibility rating. In contrast, our materials consisted of two sentences that required an inference to connect them, providing a global context on which to base a plausibility assessment. In this case, concept coherence is the basis of the plausibility rating, and no reliable effect of word coherence is found, because word coherence is rendered somewhat irrelevant by concept coherence.

The finding that word coherence plays no role in the rating of the plausibility of events does not obviate an earlier, supportive role. In their conceptual combination work, Lynott and Ramscar (2001) suggested that distributional information can aid the interpretation of noun–noun compounds by constraining meaning to a property- or relation-based interpretation, while having no effect on the processes that identify the details of the interpretation itself. Similarly, distributional information could aid sentence comprehension, while having no effect on the processes that establish inferences (see Burgess, Livesay, & Lund, 1998). Burgess et al. argued that although distributional information cannot in any way be used to make inferences, it can provide some of the necessary constraints by offering thematic cues to situation goals and word semantics. This would suggest that word coherence may contribute to the early stages of plausibility judgment but that the effect is not discernible once concept coherence comes into play.

We have shown that when people rate the plausibility of events, they are not influenced by manipulations of word coherence. Once they have created a concept-level model of the discourse, they are no longer affected by the distributional properties of the particular words used. Plausibility ratings are based on the processes of evaluating concept coherence, and it is the nature of the inferences themselves that determines how plausible people find the discourse to be.

REFERENCES

BLACK, A., FREEMAN, P., & JOHNSON-LAIRD, P. N. (1986). Plausibility and the comprehension of text. *British Journal of Psychology*, **77**, 51-60.

BURGESS, C., LIVESAY, K., & LUND, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, **25**, 211-257.

COLLINS, A., & MICHALSKI, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science*, **13**, 1-49.

CONNELL, L., & KEANE, M. T. (2003). PAM: A cognitive model of plausibility. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (CD-ROM). Mahwah, NJ: Erlbaum.

CONNELL, L., & RAMSCAR, M. (2001). Using distributional measures to model typicality in categorization. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 226-231). Mahwah, NJ: Erlbaum.

COSTELLO, F., & KEANE, M. T. (2000). Efficient creativity: Constraints on conceptual combination. *Cognitive Science*, **24**, 299-349.

COSTELLO, F., & KEANE, M. T. (2001). Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 255-271.

DUFFY, S. A., HENDERSON, J. M., & MORRIS, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 791-801.

FRENCH, R. M., & LABIOUSE, C. L. (2002). Four problems with extracting human semantics from large text corpora. In L. R. Gleitman & A. K. Joshi, *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 316-321). Hillsdale, NJ: Erlbaum.

GLENBERG, A. M., & ROBERTSON, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, **43**, 379-401.

HALLDORSON, M., & SINGER, M. (2002). Inference processes: Integrating relevant knowledge and text information. *Discourse Processes*, **34**, 145-161.

HAVILAND, S. E., & CLARK, H. H. (1974). What's new: Acquiring new information as a process in comprehension. *Journal of Verbal Learning & Verbal Behavior*, **13**, 512-521.

HESS, D. J., FOSS, D. J., & CARROLL, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, **124**, 62-82.

JOHNSON-LAIRD, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.

KEANE, M. T. (1997). What makes an analogy difficult? The effects of order and causal structure in analogical mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 946-967.

KEANE, M. T., LEDGEWAY, T., & DUFF, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, **18**, 287-334.

KINTSCH, W. (2001). Predication. *Cognitive Science*, **25**, 173-202.

KINTSCH, W., & VAN DIJK, T. A. (1978). Towards a model of text comprehension. *Psychological Review*, **85**, 363-394.

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*, **104**, 211-240.

LAPATA, M., MCDONALD, S., & KELLER, F. (1999). Determinants of adjective–noun plausibility. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 30-36). San Mateo, CA: Morgan Kaufmann.

LEMAIRE, P., & FAYOL, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd–even effect on product verification. *Memory & Cognition*, **23**, 34-48.

LUND, K., BURGESS, C., & ATCHLEY, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 660-665). Hillsdale, NJ: Erlbaum.

LYNOTT, D., & RAMSCAR, M. J. A. (2001). Can we model conceptual combination using distributional information? In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 1-10). Kildare, Ireland: NUI Maynooth.

MAYER, R. E., & BROMAGE, B. K. (1980). Different recall protocols for technical texts due to advance organizers. *Journal of Educational Psychology*, **72**, 206-255.

PICKERING, M. J., & TRAXLER, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 940-961.

REDER, L. M. (1979). The role of elaborations in memory for prose. *Cognitive Psychology*, **11**, 221-234.

REDER, L. M. (1982). Plausibility judgments vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, **89**, 250-280.

REDER, L. M., & ROSS, B. H. (1983). Integrated knowledge in different tasks: The role of retrieval strategy on fan effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 55-72.

REDER, L. M., WIBLE, C., & MARTIN, J. (1986). Differential memory changes with age: Exact retrieval versus plausible inference. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 72-81.

REDINGTON, M., CHATER, N., & FINCH, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, **22**, 425-469.

SMITH, E. E., SHAFIR, E., & OSHERSON, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, **49**, 67-96.

SPEER, S. R., & CLIFTON, C., JR. (1998). Plausibility and argument structure in sentence comprehension. *Memory & Cognition*, **26**, 965-978.

TRAXLER, M. J., & PICKERING, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory & Language*, **35**, 454-475.

NOTES

1. It is important to note here that we do not regard LSA, or any high-dimensional distributional model, as a full model of meaning (cf. Glenberg & Robertson, 2000) but, rather, as a model of a particular form of linguistic knowledge that reflects the distributional relationships between words.

2. We thank Martin Pickering for noting this possibility.

APPENDIX A

The materials used in Experiment 1 are in the following format:

Sentence 1.
Causal: Sentence 2 repeated noun (LSA score); alternate noun (LSA score)
Attributal: Sentence 2 repeated noun (LSA score); alternate noun (LSA score)
Temporal: Sentence 2 repeated noun (LSA score); alternate noun (LSA score)
Unrelated: Sentence 2 repeated noun (LSA score); alternate noun (LSA score)

The boy fumbled with his knife.
Causal: The boy bled (.93); The thumb bled (.13)
Attributal: The boy was ugly (.92); The thumb was ugly (.17)
Temporal: The boy twitched (.94); The thumb twitched (.17)
Unrelated: The boy appeared (.92); The thumb appeared (.10)

The cat pounced on the bird.
Causal: The cat gripped (.74); The claws gripped (.43)
Attributal: The cat was scary (.74); The claws were scary (.40)
Temporal: The cat relaxed (.74); The claws relaxed (.45)
Unrelated: The cat hovered (.75); The claws hovered (.51)

APPENDIX A (Continued)

The woman swiped at his face.
Causal: The woman missed (.78); The hand missed (.15)
Attributal: The woman was petite (.79); The hand was petite (.19)
Temporal: The woman waved (.80); The hand waved (.19)
Unrelated: The woman vanished (.79); The hand vanished (.17)

The bottle fell off the shelf.
Causal: The bottle smashed (.60); The glass smashed (.22)
Attributal: The bottle was pretty (.47); The glass was pretty (.18)
Temporal: The bottle sparkled (.58); The glass sparkled (.20)
Unrelated: The bottle melted (.50); The glass melted (.23)

The dress snagged on a nail.
Causal: The dress ripped (.79); The silk ripped (.26)
Attributal: The dress was costly (.73); The silk was costly (.26)
Temporal: The dress glittered (.79); The silk glittered (.29)
Unrelated: The dress shrank (.81); The silk shrank (.28)

The knife caught on the fork.
Causal: The knife snapped (.73); The blade snapped (.47)
Attributal: The knife was sharp (.67); The blade was sharp (.44)
Temporal: The knife gleamed (.73); The blade gleamed (.40)
Unrelated: The knife bubbled (.74); The blade bubbled (.40)

The girl shook the box.
Causal: The box rattled (.81); The lid rattled (.47)
Attributal: The box was wooden (.78); The lid was wooden (.38)
Temporal: The box gleamed (.81); The lid gleamed (.44)
Unrelated: The box floated (.81); The lid floated (.40)

The girl hit the mirror.
Causal: The mirror cracked (.67); The glass cracked (.25)
Attributal: The mirror was huge (.61); The glass was huge (.23)
Temporal: The mirror shone (.65); The glass shone (.24)
Unrelated: The mirror bubbled (.66); The glass bubbled (.24)

The waitress dropped the cup.
Causal: The cup smashed (.80); The handle smashed (.21)
Attributal: The cup was delicate (.70); The handle was delicate (.19)
Temporal: The cup glistened (.78); The handle glistened (.21)
Unrelated: The cup floated (.77); The handle floated (.19)

The lightning struck the tree.
Causal: The tree fell (.95); The branch fell (.57)
Attributal: The tree was huge (.93); The branch was huge (.46)
Temporal: The tree grew (.91); The branch grew (.45)
Unrelated: The tree melted (.92); The branch melted (.50)

The breeze hit the candle.
Causal: The candle flickered (.43); The flame flickered (.26)
Attributal: The candle was pretty (.35); The flame was pretty (.25)
Temporal: The candle shone (.43); The flame shone (.29)
Unrelated: The candle drowned (.44); The flame drowned (.27)

The lever closed the cage.
Causal: The cage rattled (.81); The bars rattled (.47)
Attributal: The cage was rusty (.78); The bars were rusty (.38)
Temporal: The cage tilted (.81); The bars tilted (.44)
Unrelated: The cage crumbled (.81); The bars crumbled (.40)

APPENDIX B

The materials used in Experiment 2 are in the following format:

Sentence 1.
Causal: Sentence 2. (*strong* LSA score)
Sentence 2. (*weak* LSA score)
Attributal: Sentence 2. (*strong* LSA score)
Sentence 2. (*weak* LSA score)

The opposition scored a penalty.
Causal: The goalie wept. (*strong* .29)
The goalie cried. (*weak* .04)
Attributal: The goalie was sluggish. (*strong* .29)
The goalie was slow. (*weak* .13)

The cat pounced on the bird.
Causal: The claws tore. (*strong* .46)
The claws cut. (*weak* .04)
Attributal: The claws were sharp. (*strong* .36)
The claws were pointy. (*weak* .27)

The woman swiped at his face.
Causal: The hand slapped. (*strong* .18)
The hand hit. (*weak* .11)
Attributal: The hand was petite. (*strong* .19)
The hand was little. (*weak* .13)

The pack saw the fox.
Causal: The hounds growled. (*strong* .37)
The hounds snarled. (*weak* .20)
Attributal: The hounds were fierce. (*strong* .19)
The hounds were vicious. (*weak* .12)

The flowers wilted in the vase.
Causal: The petals dropped. (*strong* .63)
The petals fell. (*weak* .53)
Attributal: The petals were velvety. (*strong* .70)
The petals were soft. (*weak* .53)

The dress snagged on a nail.
Causal: The satin ripped. (*strong* .29)
The satin tore. (*weak* .17)
Attributal: The satin was priceless. (*strong* .26)
The satin was valuable. (*weak* .13)

The knife caught on the fork.
Causal: The blade bent. (*strong* .44)
The blade curved. (*weak* .31)
Attributal: The blade was broad. (*strong* .37)
The blade was wide. (*weak* .28)

The sail caught the wind.
Causal: The canvas flapped. (*strong* .36)
The canvas fluttered. (*weak* .44)
Attributal: The canvas was strong. (*strong* .28)
The canvas was durable. (*weak* .18)

The girl shook the box.
Causal: The lid hopped. (*strong* .49)
The lid jumped. (*weak* .32)
Attributal: The lid was flimsy. (*strong* .43)
The lid was weak. (*weak* .22)

APPENDIX B (Continued)

The girl hit the mirror.
Causal: The reflection quivered. (*strong* .52)
The reflection shook. (*weak* .41)
Attributal: The reflection was indistinct. (*strong* .50)
The reflection was faint. (*weak* .40)

The wolf raced toward the flock.
Causal: The sheep ran. (*strong* .52)
The sheep fled. (*weak* .45)
Attributal: The sheep were uneasy. (*strong* .45)
The sheep were nervous. (*weak* .34)

The lightning struck the tree.
Causal: The branch scorched. (*strong* .53)
The branch burned. (*weak* .44)
Attributal: The branch was huge. (*strong* .46)
The branch was big. (*weak* .32)

The breeze hit the candle.
Causal: The flame flared. (*strong* .26)
The flame grew. (*weak* .17)
Attributal: The flame was hot. (*strong* .18)
The flame was warm. (*weak* .08)

The lever shut the cage.
Causal: The bars rang. (*strong* .34)
The bars resonated. (*weak* .25)
Attributal: The bars were rigid. (*strong* .17)
The bars were solid. (*weak* .04)

The wave crashed against the ship.
Causal: The vessel keeled. (*strong* .54)
The vessel tilted. (*weak* .39)
Attributal: The vessel was antique. (*strong* .48)
The vessel was ancient. (*weak* .24)

(Manuscript received August 22, 2002;
revision accepted for publication September 4, 2003.)