PAM: A Cognitive Model of Plausibility

Louise Connell (louise.connell@ucd.ie)
Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland

Abstract

Plausibility has been implicated as playing a critical role in many cognitive phenomena from comprehension to problem solving. Yet, plausibility is usually treated as an operationalised variable (i.e., a plausibility rating) rather than being explained or studied in itself. This paper reports on a new model of plausibility that is aimed at modeling several direct studies of plausibility. This model, the Plausibility Analysis Model (PAM), used distributional knowledge about word co-occurrence (word-coherence) and commonsense knowledge of conceptual structure and relatedness (concept-coherence) to determine the degree of plausibility of some target description. A detailed simulation of several plausibility findings is reported, which shows a close correspondence between the model and human judgments.

Introduction

Plausibility is an ineluctable phenomenon of everyday life, whether it is used to assess the quality of a movie plot or to consider a child's excuse for a broken dish. It is perhaps this very ubiquity that has led to it being ignored in cognitive science. Typically, in the psychological literature, plausibility is merely operationalised (as ratings on a scale), rather than explained. This literature has shown plausibility to play a vital role in diverse phenomena; such as discourse comprehension (Speer & Clifton, 1998), conceptual combination (Costello & Keane, 2000), reasoning (Collins & Michalski, 1989; Smith, Shafir, & Osherson, 1993) and arithmetic problem solving (Lemaire & Fayol, 1995). In this way, the empirical literature leaves us with a sense that plausibility is important but without a good indication of what it is. Theoretically, the literature really only contains broad, statements suggesting that "something is plausible if it is conceptually supported by prior knowledge" (Collins & Michalski, 1989; Johnson-Laird, 1983). In short, plausibility is in need of a thorough computational and empirical treatment.

Recently, several proposals have emerged that might well provide a computational basis for plausibility. Costello & Keane (2000) have modeled plausibility in conceptual combination, illustrating what "conceptually supported by prior knowledge" might mean. Lapata, McDonald & Keller (1999) have suggested that plausibility might be modeled by the surface, distributional properties of words themselves, though some argue that this view overlooks conceptual structure (Zwaan, Magliano & Graesser, 1995; French & Labiouse, 2002). Finally, Halpern (2001) has a well-specified model of uncertainty assessment which he terms plausibility, but this work is not intended to be a cognitive model of human plausibility judgements.

These varied approaches provide pieces of the plausibility puzzle, informing our own cognitive model of plausibility (see also Connell and Keane, 2002, 2003a, 2003b). We argue that human plausibility is based upon both concept-coherence (i.e., the conceptual relatedness of the described situation) and word-coherence (i.e., the distributional information of the words used). In this paper, we review the evidence for this theory and describe its computational implementation the Plausibility Analysis Model (PAM).

Plausibility and Concept-Coherence

Notwithstanding the lack of specificity in definitions of plausibility, there is a shared view running through the literature that plausibility has something to do with the coherence of concepts as established by prior knowledge. For example, if we were asked to assess the plausibility of the scenario --*The bottle fell off the shelf and smashed* -- we might make the bridging inferences that the bottle falling *caused* it to smash *on the floor*. We may then judge this situation to be quite plausible because our prior experience suggests that fragile things often break when they fall on hard surfaces. In short, the smashing scenario has good concept-coherence. In contrast, if we were asked to judge the plausibility of the scenario --*The bottle fell off the shelf and melted*-- we may judge it to be less plausible because there is little in our prior experience to suggest that fragile things melt when they fall onto a surface. In short, the melting scenario lacks concept-coherence. Intuitively, these examples suggest that the way the concepts cohere in a scenario contributes to its perceived plausibility.

Connell and Keane (2002, 2003a) have provided empirical support for this intuition in studies of people's plausibility judgements of scenarios with differential concept coherence (i.e., scenarios that invite different bridging inferences). Many studies have shown that people simultaneously and independently monitor causal and temporal continuity when reading, making bridging inferences when necessary, to build up a coherent model of a described scenario (Zwaan et al., 1995). Connell and Keane found that causal inferences (*causal pairs*, like the smashing scenario) were judged more plausible than those that failed to invite obvious causal inferences (*unrelated pairs*, like the melting scenario), when other factors are being held

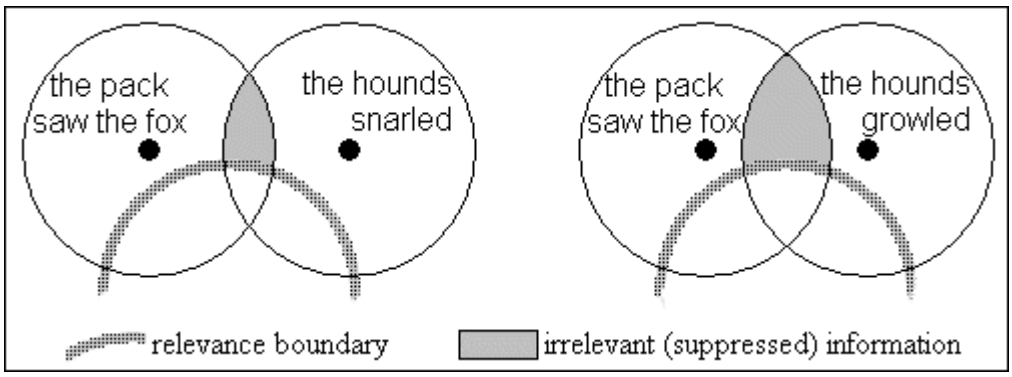

Figure 1: Illustration of distributional overlap

constant[1]. Furthermore, causal pairs were also found to be more plausible than sentence pairs that invited simple attributal inferences, which in turn were judged to be more plausible than inferences of temporal succession (see Table 1). In addition to inference type affecting how people rate the plausibility of situations, Connell and Keane (2003b) have also shown that inference type affects the time needed to make a plausibility judgement. People took significantly longer to make a binary (yes/no) decision of plausibility for causal sentence pairs than attributal sentence pairs. These studies provide specific concrete evidence that plausibility is influenced by the conceptual coherence of a situation, as shaped by the type of inferences involved.

Table 1: Example of inference types with mean plausibility scores for all materials in Experiment 1, Connell and Keane (2003a)

Inference Type	Sentence Pair	Mean Score
Causal	The breeze hit the candle. The candle flickered.	7.8
Attributal	The breeze hit the candle. The candle was pretty.	5.5
Temporal	The breeze hit the candle. The candle shone.	4.2
Unrelated	The breeze hit the candle. The candle drowned.	2.0

Note: All inference types were reliably different from one another.

Plausibility and Word-Coherence

Apart from the long-argued-for concept-coherence effect on plausibility, more recently some have argued for a word-coherence effect (Lapata et al., 2001). This view suggests plausibility judgements are sensitive to the distributional patterns of the specific words used to describe a situation. In other words, the distinctive relationships between words, as encoded in distributional knowledge, make certain sentences appear more plausible by virtue of the particular words used.

Distributional knowledge of a 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 summarised as a vector – or point in high-dimensional space – showing the frequency with which is it associated with other lexemes in the corpus. Similarly, a sentence may be represented as 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.

When a sentence is read, a neighbourhood of activation spreads out around its point in distributional space. The activated neighbourhood of a point is made up of words that are distributionally similar, such as those that the sentence in question may prime. If two sentences lie close to each other in distributional space, their neighbourhoods will have an overlap. For example, the sentence pairs:

(i) The pack saw the fox. The hounds snarled.
(ii) The pack saw the fox. The hounds growled.

have essentially the same meaning, but have different distributional overlaps. The differences in the distributional properties of *snarled* versus *growled* means that the sentences of pair (i) are further apart and thus have a smaller overlap of distributional information than the sentences of pair (ii) (see Figure 1). However, the entire distributional overlap does not contribute to the understanding of the sentences; only some of the overlapping information is relevant to the meaning of the sentence pair as a whole. For example, the overlap of pair (ii) may contain words like *leaped*, *bounded*, *beast*, *chasing*, *howling*, *lair*, etc., but many of these words (like *leaped*, *bounded*, *beast*) do not

[1] Factors controlled were word frequency (using counts from the British National Corpus), word-coherence (using scores from Latent Semantic Analysis, discussed below), and word appropriateness (of noun/verb and noun/adjective use).

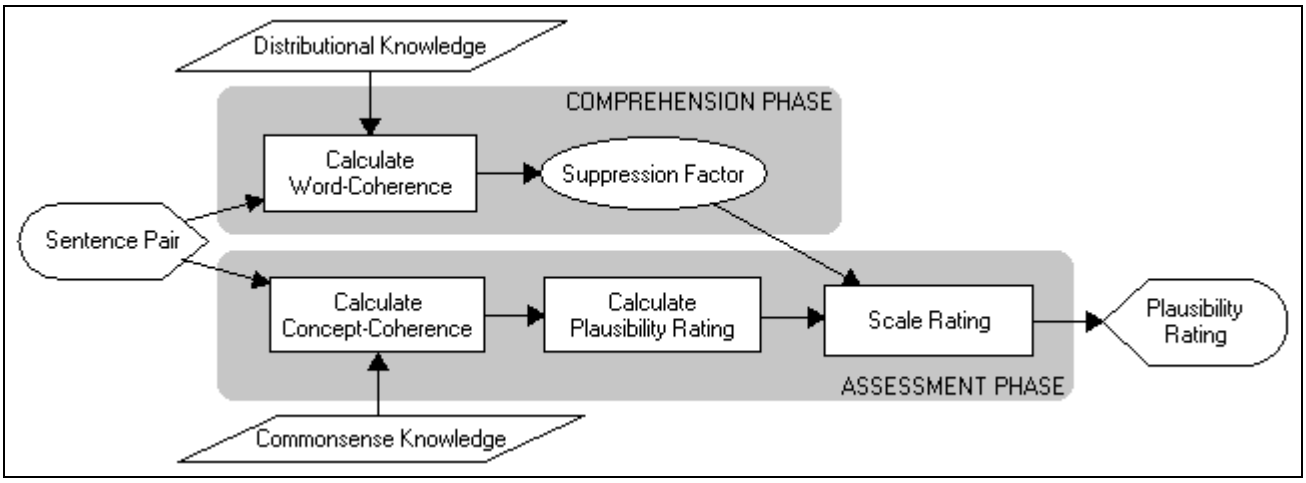

Figure 2: The Plausibility Analysis Model

play key semantic roles in scenarios about hounds hunting foxes. These words are irrelevant to the meaning of the sentence pair and must be suppressed – i.e., the information is superfluous to the task and so its activation must be dampened (Gernsbacher & Faust, 1991; Gernsbacher & Robertson, 1995). The words in the overlap (e.g., *chasing, howling, lair*) that are relevant to the meaning of the sentence pair will remain activated. In short, sentences with a small distributional overlap generally have less information to suppress than sentences with a large distributional overlap. This means that pair (i) has greater word-coherence than pair (ii) because it has a smaller overlap, and less of the activated distributional information has to be suppressed.

Connell and Keane (2003b) have shown that word-coherence measured in this way, has an effect on plausibility. They found that the greater the word-coherence of sentence-pairs, the faster people are to read them and to judge their plausibility. So, word-coherence has an effect on plausibility, albeit weaker than that of concept-coherence.

Plausibility Analysis Model

Given this recent evidence, the challenge for a cognitive model of plausibility is to capture the combined effects of concept- and word-coherence. In the remainder of this paper, we describe just such a model, the Plausibility Analysis Model (PAM). PAM takes sentence inputs and outputs a plausibility rating for the scenario described in the sentences. PAM judges plausibility using a combination of commonsense reasoning (for concept-coherence) and distributional analysis (for word-coherence). At present, PAM specifically deals with the sentences from Connell and Keane's studies though it can easily be extended, with further knowledge, to other inputs.

PAM has two phases, as shown in Figure 2. The Comprehension phase models the word-coherence effect by using distributional analysis, and the Assessment phase models the concept-coherence effect by reasoning out a scenario and rating its plausibility.

Comprehension Phase

When a sentence is first read it is parsed and each word helps to activate a certain neighbourhood of distributional knowledge. This activated neighbourhood affects the ease with which any following sentence is read. Connell and Keane (2003b) have shown that even when word frequency and appropriateness are controlled for, people are slower to read and judge the plausibility of a sentence that has a large distributional overlap with its predecessor than a sentence that has little or no overlap. PAM models this effect by the use of a model of linguistic distributional knowledge, Latent Semantic Analysis (LSA: Landauer & Dumais, 1997)[2].

PAM uses LSA to calculate the 50 nearest neighbouring words for each sentence in the pair[3], and counts the number of common terms between the neighbourhoods (i.e., the sentence overlap). This number represents the amount of distributional information shared by the two sentences. PAM then uses LSA to calculate the 50 nearest neighbours of the sentence pair as a whole, and removes these terms from the sentence overlap. What is left is the information that must be suppressed, and is shown as the shaded area in Figure 1. This suppressed information is used by PAM as a downward-scaling variable in estimating the plausibility rating. In general, the larger the distributional overlap of two sentences, the greater the amount of suppressible information and the lower the plausibility rating will be.

However, distributional information on its own does not provide adequate knowledge to judge a sentence pair's plausibility. Regardless of their degree of distributional

[2] It is important to note here that we do not regard LSA as a model of meaning (c.f. Glenberg & Robertson, 2000), but rather as a model of a particular form of linguistic knowledge that reflects the distributional relationships between words.

[3] The LSA analyses were done in the 'General Reading up to 1st Year College' semantic space, with pseudodoc comparison at maximum factors. In order to exclude misspellings and other very low frequency words, and to maximize the sensitivity of PAM, any words with a corpus frequency of less than 10 were excluded.

overlap, the sentences must be conceptually analyzed to judge whether the events described are plausible or not. This is the task of the Assessment phase.

$$plausibility = 10 \times \left(1 - \frac{1 - \frac{1}{L+1}}{P - H + 1}\right)^2$$

Figure 3: PAM's formula for plausibility ratings

Assessment Phase

PAM analyses the sentence pair by breaking it down into propositional form and checking if its selection restrictions are confirmed by its knowledge-base. To start, the concept-coherence of the first sentence in the pair is examined. For example, the sentence [*The pack saw the fox*] is transformed into propositional form as `see(pack, fox)` and the selection restrictions for its arguments are checked. The first argument requires that something be an animal in order to *see* – a *pack* contains *dogs*, and *dog* is an animal, so that requirement is met. The second argument requires that something that must be a non-abstract entity in order to be seen – a *fox* is an animal, and animals are non-abstract entities, so that requirement is met. The way in which each requirement is met is listed, and if all selection restrictions are fulfilled PAM returns this list as a path of verification. If a path is found, it means that the first sentence has been conceptually verified, and so PAM can move onto examining the second sentence.

The sentence [*The hounds growled*] is the second sentence of the pair. Again, PAM breaks it down into propositional form as `growl(hounds)` and searches for different ways to verify its selection restrictions, noting the path taken each time. For example, `growl(hounds)` may be verified via several different paths, such as the hounds growling because hounds are generally aggressive, or because they are predators who have just encountered their prey (the *fox* of the first sentence), or because they are fighting amongst themselves, etc. It is likely that there are many paths in the knowledge base that could be followed in order to verify this sentence, and PAM will note them all.

The final part of the Assessment phase involves using this set of paths to calculate a plausibility rating. To do this, PAM uses three different variables taken from the set of paths (the exact formula used can be seen in Figure 3):

1. Total Number of Paths *P* (the number of different ways the sentence can be verified in the knowledge base)
2. Mean Path Length *L* (the average count of how many different requirements must be met per path)
3. Proportion of "Hypothetical" Paths *H* (proportion of all paths that can only be followed by meeting a requirement for something that is not explicitly mentioned – e.g. [*The bottle fell off the shelf. The bottle melted.*] is considered a plausible path if we allow that the bottle may have fallen into a hypothetical furnace)

The rating returned is between 0 (not plausible) and 10 (completely plausible), and is calculated according to the asymptotic function of Figure 3. In short, a high number of paths (*P*) means higher plausibility, because there are more possible ways that the sentence can be verified. A high mean path length (*L*) means lower plausibility, because elaborate requirements must be met to verify the sentence. Finally, a high proportion of hypothetical paths (*H*) means lower plausibility, because it is assuming the existence of entities that may not be there.

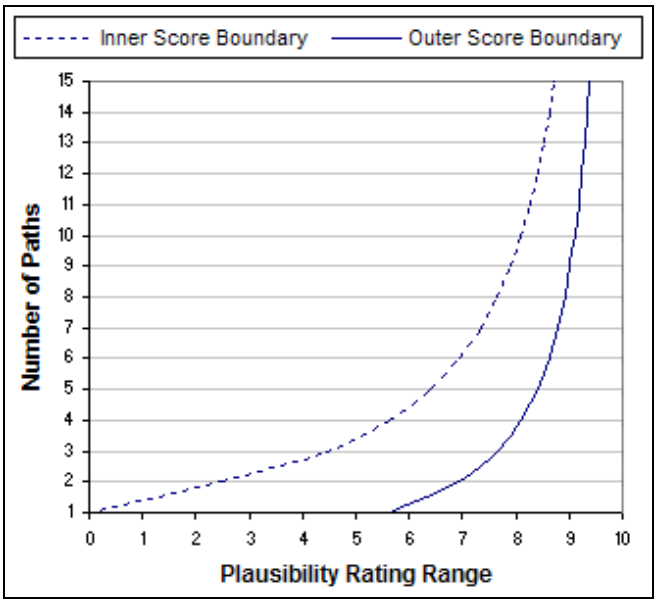

Figure 4: Graph of PAM's plausibility rating function

Figure 4 shows the boundaries of plausibility score that PAM generates for an increasing number of paths. The dotted line represents the inner (lower) score boundary, which is the worst-case situation where the mean path length approaches infinity and every path is hypothetical. The solid line represents the outer (upper) score boundary, where the mean path length is one and no path is hypothetical. For example, a set of four (non-hypothetical) paths with a mean length of three will have a rating of 7.2 out of 10, while a set of three paths (again with a mean length of three) will have a rating of 6.6 out of 10. If one of those three paths were a hypothetical path, then the score would drop to 6.3 out of 10.

When the path rating has been calculated, PAM then applies the scaling variable supplied by distributional knowledge in the Comprehension phase to represent the carry-over effect that the effort of suppression has on plausibility ratings. The scaling is of a lesser magnitude than that of the other variables in the model, but will still have a perceptible effect. In this way, PAM models the small difference in plausibility ratings found between versions of sentence pairs that vary in their distributional overlap but are conceptually identical.

Model Evaluation

PAM's performance in plausibility ratings was compared to human data. Using the sentence pair materials from Connell

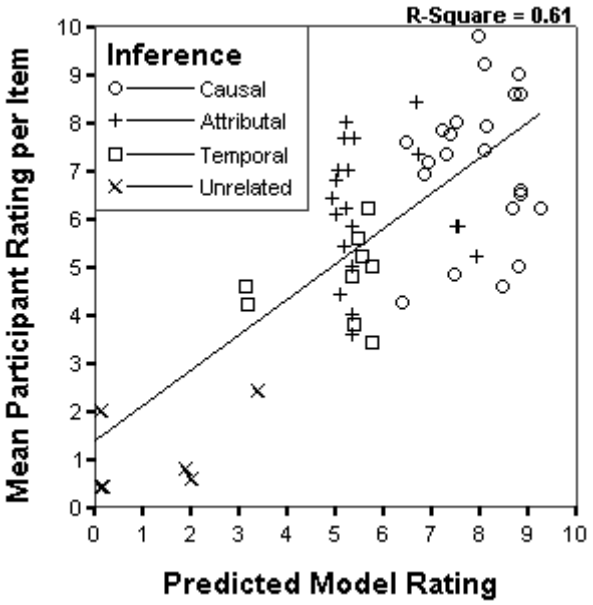

Figure 5: PAM's output against human plausibility ratings

and Keane (2003a), the simulation produced plausibility ratings that were then compared to the human judgments. The test items used were a different subset of Connell and Keane's materials than those used as PAM's training items.

It is important to note here that although the simulations were performed with materials from Connell and Keane's papers, PAM was designed to be generalisable to any other input simply by extending the commonsense knowledge base. We will address this issue further in the general discussion. Additionally, PAM's knowledge base was built in a "blind" fashion. That is, the knowledge was simply represented in local definitions of requirements, without checking possible path lengths that might emerge or without modifying the knowledge base to fit the data.

Simulation

Materials Connell and Keane's (2003a) materials were from two experiments, from which 60 sentence pairs were drawn as test items for the simulation. Of these, there were a number of different variants of each sentence pair. For example, some sentence pairs had variants manipulating concept-coherence (e.g., causal inference [*The bottle fell off the shelf. The bottle smashed.*] versus unrelated inference [*The bottle fell off the shelf. The bottle melted.*]) while others manipulated word-coherence (e.g., large distributional overlap [*The pack saw the fox. The hounds growled.*] versus small distributional overlap [*The pack saw the fox. The hounds snarled.*]).

Procedure The procedures in the two psychological experiments were slightly different. The first experiment, which manipulated just concept-coherence, presented each sentence pair on its own page in a booklet. Participants were then asked to judge the plausibility of the sentence pair and rate it on a 10-point scale (where 0 was implausible and 10 was very plausible). The second experiment, that manipulated both concept- and word-coherence, presented two sentence pairs per page in the booklet, where one pair was the variant with the large distributional overlap and the other pair was the variant with the small distributional overlap. Again, participants were asked to judge the plausibility of both sentence pairs and rate them on two separate 10-point scales. For the purposes of this simulation, the mean plausibility rating of each sentence pair was used. The procedure for PAM was to enter each natural language sentence pair and note the output from the Assessment phase, which took the form of a rating of plausibility (0-10).

Results & Discussion PAM returned plausibility ratings that were highly correlated with the human data from Connell and Keane (2003a), R=0.788, p<0.0001, N=60. A regression analysis confirmed that PAM's output could be used to predict human performance in plausibility ratings, R^2=0.621, p<0.0001. Figure 5 shows a scatterplot of the relationship between model output and participant means.

PAM performed well for all four concept-coherence variants (causal, attributal, temporal and unrelated). Table 2 shows the means per inference type for Connell and Keane's data against PAM's.

We also altered PAM's output to disregard the effect of word-coherence in the Assessment phase, and compared this to the human data. While we still found a significant correlation (R=0.779, p<0.0001), it was less than that found earlier and a regression analysis showed that PAM's performance had worsened by 1.4% without the word-coherence effect, R^2=0.607, p<0.0001. This confirms that word-coherence does indeed have a pertinent effect on PAM's plausibility ratings.

Table 2: Mean Plausibility ratings per inference type from PAM and Experiment 1, Connell and Keane (2003a)

Inference Type	Human Rating	PAM Rating
Causal	7.8	7.9
Attributal	5.5	5.7
Temporal	4.2	5.0
Unrelated	2.0	0.9

General Discussion

There are a number of novel achievements reported in this paper. The Plausibility Analysis Model (PAM) is the first computational model that specifically and accurately addresses human plausibility judgements. It does this by using a number of innovative techniques to capture the complex influences that empirical work has shown to bear upon plausibility, namely the use of both commonsense knowledge and distributional knowledge.

PAM uses a commonsense knowledge base to assess concept-coherence. This assessment is based upon an analysis of the requirements that must be met for a proposition to be true. Many of these requirements are based upon what is intuitively regarded as common sense. For

example, for an entity *X* to *melt*, one of the requirements is that *X* is currently *solid*. For *X* to be *solid*, there is a further requirement that *X* is *non-abstract*, and so on. In general, this precludes the use of figurative language in the sentence pairs that PAM takes as input, but it would be possible to build up such a requirements set for future versions.

In addition to concept-coherence, PAM also assesses word-coherence by using linguistic distributional knowledge. It does this through the use of Latent Semantic Analysis (LSA). However, rather than the conventional use of LSA scores that represent the distance between points in a high-dimensional space (c.f. Kintsch, 2001; Landauer & Dumais, 1997), we have taken the alternative approach of neighbourhood activation. By treating words and sentences as activating only a certain area of distributional knowledge, we believe our implementation of a high-dimensional distributional space to have greater cognitive plausibility.

There is an interaction between commonsense knowledge and distributional knowledge as shown in the empirical work of Connell and Keane (2002; 2003a; 2003b). For a considered plausibility rating, PAM models the interaction as sequential: the conceptual soundness of the situation is fully explored and afterwards a lingering effect of distributional knowledge is applied. While the simulations were run with all available human data, it is our intention to use PAM to create more sentence pairs and examine how its output predicts additional human plausibility ratings. It is also our intention to extend PAM to deal with other discourse inputs, which will require only that the commonsense knowledge base be extended accordingly. The distributional knowledge accessed in the Comprehension phase need not be altered, as LSA already deals with the full English language.

PAM is the computational implementation of the plausibility theory put forward by Connell and Keane (2003a; 2003b), and as such is the first model specifically of human plausibility judgements. Although still in development, the simulations reported here demonstrate the importance and accuracy of PAM's modeling techniques. When people judge the plausibility of a scenario, they are influenced both by the concept-coherence of the situation in hand and by the word-coherence of the description they have read or listened to. Any future models of human plausibility judgements must therefore take account of both these factors, and implement conceptual and distributional knowledge, and the interactions between them.

Acknowledgments

This work was funded in part by grant from the Irish Research Council for Science, Engineering and Technology.

References

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

Connell, L., & Keane, M. T. (2002). The roots of plausibility: The role of coherence and distributional knowledge in plausibility judgements. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (p. 998). Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M. T. (2003a). What Plausibly Affects Plausibility? Concept-Coherence & Distributional Word-Coherence As Factors Influencing Plausibility Judgements. *Manuscript in submission.*

Connell, L., & Keane, M. T. (2003b). The effect of distributional information on plausibility decision times. *Manuscript in preparation.*

Costello, F., & Keane, M.T. (2000). Efficient Creativity: Constraints on conceptual combination. *Cognitive Science, 24,* 299-349.

French, R., & Labiouse, C. (2002). Four problems with extracting human semantics from large text corpora. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 316-321). Hillsdale, NJ; Erlbaum.

Gernsbacher, M. A., & Faust, M. (1991). The role of suppression in sentence comprehension. In G.B. Simpson (Ed.), *Understanding word and sentence.* Amsterdam: North Holland.

Gernsbacher, M. A., & Robertson, R. R. W. (1995). Reading skill and suppression revisited. *Psychological Science, 6,* 165-169.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379-401.

Halpern, J. Y. (2001). Plausibility Measures: A General Approach for Representing Uncertainty. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence,* (pp. 1474-1483). San Mateo, CA: Morgan Kaufmann.

Johnson-Laird, P.N. (1983). *Mental Models.* Cambridge: Cambridge University Press.

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

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. *Proceedings of the 9th 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 rule effect in simple arithmetic. *Memory and Cognition, 23,* 34-48.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67-96.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26*(5), 965-978.

Zwaan, R.A., Magliano, J.P., & Graesser, A.C. (1995). Dimensions of situation-model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 386-397.