A COGNITIVE THEORY AND MODEL OF PLAUSIBILITY

Louise Connell
B.Sc. (Hons), M.Sc.

The thesis is submitted to University College Dublin
for the degree of PhD in the Faculty of Science

December 2003

Department of Computer Science
Head of Department : Mr. Greg O'Hare
Supervisor of Research : Prof. Mark Keane

TABLE OF CONTENTS

SUMMARY

Whenever we must evaluate a theory, consider an excuse, or appraise a situation, we must judge how plausible things appear to us. We are aware when new information seems to "fit", and we notice when it seems "out of place". Plausibility judgement occupies a central position in human cognitive life, and contributes to many other tasks, from memory retrieval and encoding to sentence parsing and conceptual combination. However, this diversity of roles has meant that there is little agreement in the literature as to the exact processes and influences involved in plausibility judgement.

This thesis proposes the Knowledge-Fitting Theory of Plausibility, which views plausibility as being about fitting what we're told to what we know about the world. Plausibility judgement is described as spanning two stages; *comprehension* represents the presented scenario, and *assessment* examines this representation to gauge how well it fits prior knowledge. A series of experiments show that two factors influence the plausibility judgement process: the *word-coherence* of the description's distributional properties, and the *concept-coherence* of the scenario itself. The Knowledge-Fitting Theory is implemented computationally in the Plausibility Analysis Model (PAM). In simulations, PAM's performance closely models human responses in both plausibility ratings and judgement times.

The Knowledge-Fitting Theory gives plausibility a clarity of definition that has hitherto been absent from cognitive science, and impacts upon the wide variety of fields in which plausibility judgement plays a role.

ACKNOWLEDGEMENTS

Many people deserve my gratitude for a variety of reasons, and unlike the rest of this tome, I'll try to be concise…

First, I would like to thank my supervisor, Prof. Mark Keane, for all his generous time, help, encouragement, and insight over the last few years. This thesis couldn't have been written without his guidance and, importantly, vast reserves of patience.

Also, my thanks to Dermot Lynott for infinite quantities of support and patience, for being an uncomplaining sounding board, and for keeping me reminded of life outside these pages. Without his perceptive and constructive comments, this thesis would have been a much lesser work.

Thanks to the UCD-TCD Cognitive Science seminar group, who provided a good forum for testing whether any of my ideas about plausibility seemed plausible.

Finally, I would like to thank my family for their support over the years, and for their inexplicable continuing interest in what I actually do. And for happily accepting that I will never get a real job.

This thesis was funded by scholarships from the Department of Computer Science in University College Dublin, the Higher Education Authority under the Multimedia Research Programme in collaboration with MediaLab Europe, and the Irish Research Council for Science, Engineering and Technology under the Embark Initiative.

CHAPTER 1 – INTRODUCTION

"Plausible impossibilities should be preferred to implausible possibilities."
– Aristotle. (350 B.C.). *Poetics, 24*.

1.1 – Background and Objective of Research

Everyday, in many different scenarios, we judge plausibility. Whether we are evaluating the goodness of a theory, considering the alibi of a murder suspect in a crime novel, or weighing a child's claim that the cat left those muddy boot-prints on the floor, we are essentially judging how plausible each scenario seems to us. Intuitively, these judgements appear to involve quite rapid assessments based on aspects of the presented information and quick inferences made using our knowledge about the world.

Across the cognitive psychology literature, plausibility is often mentioned in the service of other phenomena rather than forming the prime focus of the research. The pervasiveness of plausibility is reflected in the many different cognitive contexts in which it has been studied (e.g., memory retrieval, arithmetic problem solving, probability judgement, discourse comprehension, conceptual combination). Indeed, this very pervasiveness seems to have made plausibility harder to explain, as the literature typically treats it merely as an operationalized variable (i.e., ratings of "goodness" or plausibility) or as an underspecified subcomponent of some other

phenomenon. In short, it seems that the very centrality of plausibility has made it invisible and ineluctable.

In fields other than cognitive psychology, plausibility has sometimes been aligned with probability, or, more specifically, with Bayesianism (Dempster, 1968; Friedman & Halpern, 1996; Halpern, 2001; Shafer, 1976). Bayesianism is the philosophical tenet that the mathematical theory of probability can be applied to the degree of plausibility of statements. This idea has several applications in artificial intelligence and expert systems (e.g., Norton, 1988; Spiegelhalter, Thomas & Best, 1996) in the form of Bayesian inference; a type of statistical inference in which probabilities are interpreted as degrees of belief. According to this account, the plausibility of a particular statement is interpreted as the degree of belief of rational agents in the truth of the statement, and this plausibility can be updated in light of new evidence using Bayesian probabilities. The principal advantage of Bayesianism is that it allows probabilities to be assigned to many kinds of statements, including statements about random events and personal beliefs. However, while Bayesian inference and other probabilistic approaches have been shown to approximate human performance in specific domains such as syllogistic reasoning (Chater & Oaksford, 1999; Johnson-Laird, 1983; see also Tenenbaum & Griffiths, 2001), this does not mean that people estimate probabilities when they judge the plausibility of a scenario. A scenario may be plausible without being probable; for example, a person may consider it plausible that there was once life on Mars without regarding it as probable. Conversely, a scenario may be probable without being plausible; for example, a person may regard a particular explanation as probable if all the alternatives have been eliminated, while still considering it to be an implausible

account. Indeed, an unpublished study carried out by the author showed an empirical distinction between people's plausibility judgements and probability judgements. While people rated scenarios with causal, attributal, temporal and unrelated events to have distinctly different levels of plausibility (see Experiment 1), this distinction was lost when people were asked to rate the probability of the same scenarios. Thus, plausibility is examined in this thesis quite separately from probability, and is characterised as a cognitive process in its own right.

The objective of this thesis is to explain plausibility in and of itself. We propose a theory of plausibility, the Knowledge-Fitting Theory, that characterises the stages of the plausibility judgement process. This theory sees the plausibility of a scenario as being about the degree of fit between the scenario and prior knowledge, where this prior knowledge may draw on our conceptual knowledge of the world and our distributional knowledge of the words in the scenario description. In addition, we describe a computational implementation of the theory, the Plausibility Analysis Model (PAM), and show how the model parallels human plausibility responses. The following section offers a short introduction to the theory, and outlines the core ideas and operations.

1.2 – The Knowledge-Fitting Theory of Plausibility

When we wish to judge the plausibility of a scenario, be it a suspect's alibi or a politician's excuse, we must first properly understand the described scenario and then analyse it to see how it fits our knowledge of the world. In the Knowledge-Fitting Theory of Plausibility, there are two factors that influence how we do this:

the *word-coherence* of the description itself and the *concept-coherence* of the scenario's elements and events. This section presents a brief account of the Knowledge-Fitting Theory and offers a high-level introduction to its key components; a full and detailed description of the theory is given later in the thesis.

Regarding word-coherence, there are innumerable ways that any scenario can be described and the exact choice of words affects how easily we can understand the scenario. To be more specific, the description will be understood more easily if it sparks activations in our long-term memory of things related to the scenario. For example, if the scenario describes how a pack of hounds growled when they saw a fox, we will find it easier to understand if our knowledge of dogs' hunting behaviour is readily available. According to the Knowledge-Fitting Theory, the description activates our background knowledge about hunting dogs via our distributional knowledge of the kind of contexts in which we have previously encountered hounds, growling, and foxes. A particular description might activate either a lot or a little of our background knowledge, and the scenario will be understood with ease or difficulty, respectively. The way in which this activation mechanism works can be thought of in terms of a "*distributional spotlight"*. When we read a sentence, a spotlight falls on an area of distributional knowledge containing words that we usually find in similar contexts. The next sentence will spotlight another set of words in distributional knowledge, and so on. As this happens, the words in each distributional spotlight activate associated pieces of information in long-term memory (see Figure 1.1). In this way, the understanding of the scenario will be best facilitated if the distributional spotlights fall at a distance from each other and do not overlap, because this will ensure that the greatest amount of background knowledge

is made available. In other words, plausibility judgement is influenced by how much background knowledge the scenario description can activate in our long-term memory.

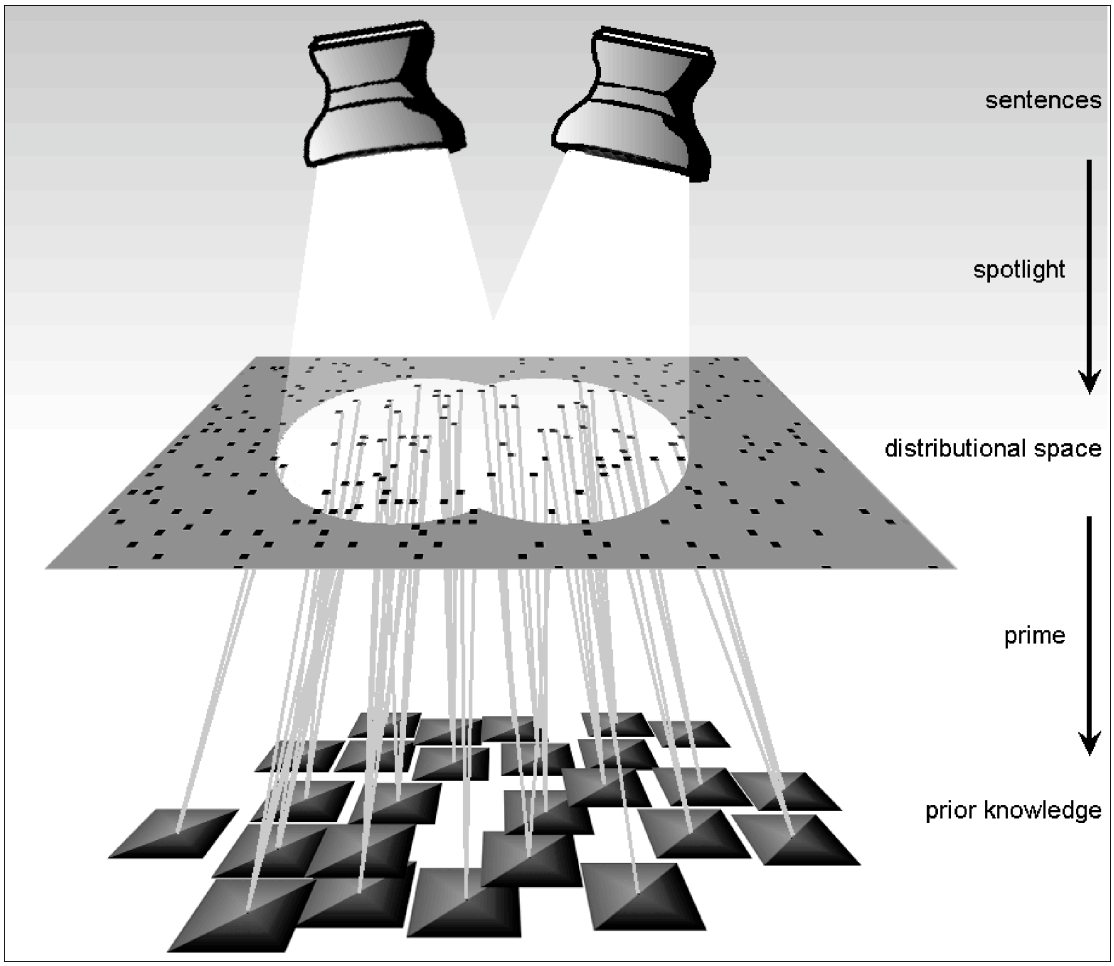


Figure 1.1 – the distributional spotlight: the scenario's sentences spotlights areas of distributional knowledge, which then activate associated pieces of information in long-term memory.

With regard to concept-coherence, some scenarios are relatively straightforward and contain only simple descriptions and events. These scenarios can be understood simply by processing one clause after another, and they are equally easy to analyse when we wish to assess their plausibility. For example, if a

scenario describes a dog as being brown and white, and having floppy ears, and having a loud bark, we can understand and analyse it quite easily. On the other hand, some scenarios are quite complex and can only be properly understood by connecting the clauses that have causes and consequences, or happened in a specific order in time. Understanding these scenarios takes more effort because they involve drawing on our background knowledge of similar situations. For example, if a scenario describes a dog seeing a rabbit, and chasing after it, and whining when it disappears into a burrow, we must use our background knowledge of dogs to understand why it chased the rabbit and why it whined. This extra background knowledge also means that the scenario will also take more effort to analyse when we wish to assess its plausibility. In other words, plausibility judgement is influenced by what background knowledge we need to make the scenario fit what we know of the world.

In brief, the Knowledge-Fitting Theory of Plausibility proposes that the plausibility judgement process takes place across two stages: first, there is the comprehension stage where we understand the scenario, and second, there is the assessment stage where we examine its fit to prior knowledge. The distributional properties of the words in the description (word-coherence) can influence how easily we understand the scenario, by activating background knowledge from long-term memory. Our ease of understanding is also influenced by how much background knowledge we need to make the scenario fit what we know about the world (concept-coherence). The scenario's plausibility is then assessed based on the quality of this knowledge fit, by examining what knowledge was drawn in (concept-coherence).

1.3 – Structure of Thesis

The next chapter in this thesis is Chapter 2, which takes the form of a review of the theoretical and empirical literature that is relevant to the study of plausibility. This body of literature identifies two possible factors that may influence plausibility judgements: conceptual consistency with prior knowledge (*concept-coherence*) and the distributional properties of the words in the description (*word-coherence*). In the next chapters, we analyse the effects of these factors using two different experimental paradigms. Chapter 3 reports three experiments that measure *plausibility ratings*, which allow us to examine the end product of the plausibility judgement process. Chapter 4 reports two experiments that measure comprehension times and *plausibility judgement times*, which allow us to examine the time course of the judgement process. In Chapter 5, we return to the Knowledge-Fitting Theory of Plausibility, and describe in detail how it accounts for the effects found in the previous experimental chapters as well as those in the literature. We also describe the computational implementation of this theory, the Plausibility Analysis Model (PAM), and show how the model's performance parallels human plausibility responses. Finally, in Chapter 6, we give an overview of the work reported in this thesis and outline some possible directions for further research.

CHAPTER 2 – REVIEW

2.1 – Introduction

Plausibility has the hallmarks of a phenomenon so pervasive and important that, like air, no one notices it. The term "cognitively plausible" appears time and again across many fields of literature, yet the definition of this expression is fuzzy at best. Many cognitive accounts are content to appeal to the idea of plausibility without specifying its cognitive basis. Indeed, most of these accounts are grounded in transient operational definitions of plausibility; for example, the plausibility ratings people give to some set of stimuli. Even dictionary definitions of the term – "something is plausible if it is apparently, seemingly or even deceptively true" – fail to give us much insight.

This purpose of this chapter is to provide a review of the theoretical and empirical literature that is relevant to the study of plausibility. This body of literature spans three principal areas: treatments of plausibility, discourse comprehension and distributional analysis. First, we discuss how previous work has treated plausibility, and evaluate the extent to which these various studies have succeeded in specifying the nature of plausibility. Second, we review the discourse comprehension literature, and argue why a proper understanding of how people represent information is essential to any investigation of plausibility. Third, we describe the distributional analysis of linguistic information, and discuss how it has

been implicated in many cognitive phenomena and why it may be relevant to the analysis of plausibility. To conclude the chapter, we summarise the main points and delineate what the literature has to offer to the study of plausibility.

2.2 – Treatments of Plausibility

Although no specified account of plausibility has emerged from the literature, there is a shared view that plausibility has something to do with conceptual coherence and prior knowledge. This theoretical view holds that some scenario, event or discourse is plausible if it is conceptually consistent with what is known to have occurred in the past. Plausibility has been shown to play a central role in a wide variety of cognitive acts; some research adheres to this shared view but other work has suggested that plausibility may be based on very different word-level criteria.

2.2.1 – Plausibility in Memory

It has been argued that people use plausibility whenever possible in place of direct retrieval from memory (Lemaire & Fayol, 1995; Reder 1979; 1982; Reder, Wible & Martin, 1986). In one paradigm, Reder (1982) gave people a story followed by a test sentence, and asked either (A) if the sentence could be true given the story or (B) if the sentence appeared in the story. When tested after a delay, people were faster at task A (judging if the sentence could be true/plausible) than at task B (verifying if it had previously appeared). In addition, people appeared to be judging plausibility even when asked to perform a retrieval task, and that their choice

of strategy depended partly on task demands and partly on recency of memory. From these findings, Reder proposed that plausibility judgement was an automatic strategy that was more efficient than direct retrieval, at least once verbatim memory has faded.

Plausibility judgement appears to hold an even greater advantage over straight retrieval when one knows a lot about a particular topic. The "fan effect" phenomenon shows that as people study more facts about a particular topic, their time to retrieve a particular fact about that topic increases (Anderson, 1974; Radvansky, Spieler & Zacks, 1993; Reder & Anderson, 1980). However, it has been shown that the fan effect is inverted for plausibility judgements; that is, the more facts one studies about a particular topic, the faster one can judge if a given fact is plausible / consistent (Reder & Anderson, 1980; Reder & Ross, 1983). For example, Reder and Anderson (1980) gave participants several facts to learn about a character called Marty doing his laundry, and later presented them with the test sentence "*Marty cleaned the lint trap*". Reder and Anderson found that the more facts that people had learned about Marty, the faster they could confirm that the test sentence was plausible / consistent, but the slower they were to confirm that the test sentence had been seen before. These findings were taken as further evidence that plausibility judgement was a more efficient strategy than direct retrieval.

In further research, the performance of younger and older participants was compared to see if cognitive performance for plausibility judgement degrades with age in a way that parallels degradation in memory retrieval (Reder, Wible & Martin, 1986). Using a similar story-sentence paradigm, they found that older participants had difficulty with the specific retrieval task, and that they tended to over-rely on

plausibility strategies at the expense of accuracy. In other words, age made little difference to performance so long as a plausibility judgement could provide the correct response. Reder et al. concluded that while careful retrieval is a much more costly process for older adults than for young adults, plausibility judgement is equally easy for both groups. Moreover, they suggested that plausibility judgement becomes an error-minimising fallback mechanism for older people to compensate for the increasing risk of forgetting due to age.

Plausibility judgement and retrieval have also been examined in studies of non-verbal information. Lemaire and Fayol (1995) examined performance for simple mathematical tasks, using mental arithmetic problems of the kind learned as "times-tables" (e.g., 7×3, 4×9). These problems had been categorised as either easy or difficult by an earlier group of adult raters. Lemaire and Fayol found that plausibility strategies were favoured over direct retrieval for more difficult problems, and concluded that this was because difficult problems took too long to retrieve from long-term memory. Indeed, young children were found to favour plausibility judgements over retrieval for easy as well as difficult problems, since for them, retrieval was equally difficult on all problems. Lemaire and Fayol argued that people used plausibility judgements to bypass normal retrieval processes when trying to retrieve something not easily available in long-term memory.

In addition to memory retrieval, plausibility may also have an effect on memory encoding (Loftus, 1979; Pohl, 1998; Thompson & Kliegl, 1991). Work by Pohl (1998) suggests that people encode plausible information more readily than implausible information. In a feedback paradigm, participants were asked to estimate the answer to numerical questions such as "*What year was the first reflector*

telescope built?" (e.g., *1600*) and after a delay were given a solution labelled as another person's estimate. When the given solution was plausible (e.g., *1671*), people tended to exhibit *hindsight bias* and "recollect" a better answer than they had actually given (e.g., *1650*). However, when the solution was implausible (e.g., *1171*), people could recall their own estimate fairly accurately. Pohl concluded that people use plausibility to decide whether information is worth encoding when the reliability of the source is in question; if information from another person seems plausible then it is encoded, but if it seems implausible it is rejected as incorrect. Plausibility effects on knowledge encoding were also found by Thompson and Kliegl (1991), who examined whether performance differs between older and younger subjects. They found that older participants have difficulty in encoding implausible information, and concluded that plausible information is in some way easier to encode.

Evaluation of the Memory Literature

None of these studies attempt to define plausibility, but instead consider plausibility judgement as a strategy that can sometimes allow us to avoid the costly and painstaking process of accessing memory. In memory retrieval tasks, we judge plausibility instead of determining whether presented information (e.g., a test sentence) exactly matches something in memory (e.g., a story read earlier). In memory encoding tasks, we first judge plausibility to evaluate whether the information (e.g., a date for an event) is worth remembering in the long term. However, these studies also found that error rates were higher in the plausibility judgement tasks – although plausibility judgement is quick and easy, it is not perfect.

In other words, plausibility judgement is a type of "cognitive shortcut" mechanism but its convenience comes with a price. Plausibility judgement seems to involve examining if the presented information *generally fits* with the current scenario. This can lead to mistakes, because information that "generally fits" may still not be correct. Memory research tends to operationalise plausibility as what people do in a variety of experimental contexts. Although plausibility judgement appears to be extremely useful, no real statement is made as to its nature.

2.2.2 – Plausibility in Text Comprehension

The idea that plausibility plays a key role in text comprehension has been around for some years. Black, Freeman and Johnson-Laird (1983) showed that implausible stories were harder to understand and remember than plausible stories. They created texts of varying levels of plausibility by manipulating the relatedness of the sentences while keeping referential coherence constant. For example, a story about an old man features the sentences "*He struggled with a youth who killed him. Then they stretched out indolently in the sun on the bank.*" While referential coherence is maintained (i.e., *they* in the second sentence refers to the *he* and *youth* of the first sentence), these sentences are difficult to plausibly relate unless one assumes that the story has taken a supernatural turn. Black et al. found that people's ease of comprehension and memory for the text was very sensitive to the number of plausible relations that could be made between the sentences in the story. From these findings, they concluded that a plausible text was one that allowed inferences to be made to establish relations between events, and that diminishing plausibility

meant a disruption of people's ability to construct a sensible representation of the text.

At a much lower level of detail, eye-tracking studies have shown that people make plausibility judgements on a word-by-word basis in order to aid sentence parsing (Pickering & Traxler, 1998; Speer & Clifton, 1999; Traxler & Pickering, 1996). For example, Speer and Clifton showed that highly plausible prepositional phrases (e.g., "The coach discussed a new technique *with the players*") were read more quickly than less plausible prepositional phrases (e.g., "The coach discussed a new technique *with the tailor*"), based on the plausibility of the match between the final noun and the rest of the sentence. Pickering and Traxler (1998) also showed that people used the plausibility of verb-object matches in order to resolve garden path sentences. In the sentence "*As the woman edited the magazine about fishing amused all the reporters*", people incorrectly read *the magazine* as the object of the verb *edited* until the syntactic choice point of *amused* is reached; at this point people re-analysed the sentence to resolve the garden path syntax. However, when the object *the magazine* was paired with a less plausible verb ("*As the woman sailed the magazine about fishing amused all the reporters*") people re-analysed the sentence at the point of reaching the verb *sailed*; they did not need to reach the syntactic choice point to resolve the garden path syntax.

There is some evidence to suggest that plausibility may be important to general comprehension skills rather than just to text comprehension. Gernsbacher (1985, cited in Gernsbacher, 1997) found that people could judge the plausibility of a sentence just as quickly when a picture substituted for one of its words as when the sentence contained only words or they read or heard text-based stories. Since it has

also been shown that people draw the same inferences about a sequence of events whether the events were presented as pictures or sentences (e.g., Baggett, 1979), Gernsbacher (1997) concluded that many of the mechanisms central to language comprehension may actually be central to any cognitive task involving information-processing.

Evaluation of the Text Comprehension Literature

The disparate approaches used to examine plausibility judgement in text comprehension have also led to disparate results. For example, the eye-tracking studies indicate that plausibility judgement is made on a word-by-word basis as one parses a sentence (e.g., some verb-object combinations are more plausible than others). In contrast, the work by Black et al. suggests that plausibility results from the ability to make inferences between events (e.g., that some event sequences are more plausible than others). No explanation of how to resolve these word-oriented and event-oriented accounts of plausibility has been offered by the literature. Furthermore, each account suffers somewhat from the limitations of its own paradigm. The eye-tracking studies examine particular syntactic clauses, and are thus focussed at too low a level to provide a general account of plausibility in reading. Black et al. examine relations between events without specifying whether the inference in question is causal, temporal, spatial etc., and is thus too vague to provide a meaningful account of plausibility in reading. In this respect, while Gernsbacher's cross-modal finding is interesting, we cannot begin to generalise how plausibility might be important to pictorial or verbal comprehension when its importance to text comprehension is so underspecified. All the comprehension

research leaves us with is the indeterminate impression that, however plausibility affects text comprehension, it probably also affects other comprehension modalities in the same way.

2.2.3 – Plausibility in Reasoning

Plausibility has sometimes been treated as part of a larger theory of reasoning and decision-making (Collins & Michalski, 1989; Smith, Shafir & Osherson, 1993; Thagard, 2000). In their Theory of Plausible Reasoning, Collins and Michalski (1989) aimed to construct a formal characterisation of the different patterns of plausible inference that people use in reasoning about the world, including deduction, induction and analogy. For example, given the information that England has daffodils, Collins and Michalski say that people reason about Brazil having daffodils by examining the similarity of the two countries' climates. The conclusion that Brazil is probably daffodil-free because of climatic dissimilarity is characterised by Collins and Michalski as a plausible deduction. By definition, this was not a theory of how people make plausibility judgements. Rather, it was a theory of how people reason about questions for which they do not have ready answers, based on parameters such as similarity, typicality and frequency.

Smith, Shafir and Osherson (1993) considered the relationship between plausibility and similarity to be quite different to that of Collins and Michalski. In inductive inference, Smith et al. showed that people use similarity and/or plausibility to decide if an argument is true. For example, if we are told that "*robins have sesamoid bones*", we infer whether the same is true of *ducks* based on the similarity

of *robins* and *ducks*. However, if told that "*robins can fly faster than 20 miles an hour*", we base our decision not only on the similarity of *robins* and *ducks*, but also on how plausible we find the idea of robins flying that fast. In other words, when reasoning about a topic for which we have some background knowledge, Smith et al. concluded that plausibility plays an essential role in weighing arguments before making a decision about likelihood.

A similar conclusion was reached by Thagard (e.g., 1989, 2000) in his Theory of Explanatory Coherence, where gauging the plausibility of an argument is regarded as a key part in all reasoning processes. In Thagard's account, reasoning is about evaluating how arguments "cohere" with each other. Knowledge is represented as a network of propositions, and coherence is considered to result from how relations among two or more propositions may "hold together" or "resist holding together". There are a number of detailed principles that are said to constrain and establish the pairwise relations between propositions, including explanation (a proposition that explains another proposition coheres with it), contradiction (a proposition is incoherent if it contradicts another) and simplicity (a proposition coheres if few co-propositions that are needed to help explain it). Of these principles, simplicity is singled out as being relevant to plausibility; in short, Thagard considers a simple argument to be a plausible argument.

Evaluation of the Reasoning Literature

These theories of human reasoning are quite diverse, and each of them has a different perspective on what plausibility involves. However, none of them succeed in specifying plausibility in its own right. Collins and Michalski's theory of

plausible reasoning seems to have a somewhat tautological view of plausibility: an inference that succeeds is plausible, and an inference must be plausible to succeed. Plausibility itself remains undefined, as the objective of their theory was not to characterise human plausibility judgement. In Smith et al.'s work, plausibility is regarded as part of certain types of probability judgement. This appears to be the case when reasoning about the properties of natural kinds (as in Smith et al's experiments) but it is imprudent to generalise this to more complex reasoning without more empirical work. Similarly, Thagard's theory can be criticised for its lack of empirical grounding, as it is motivated more towards creating artificially intelligent models of reasoning in specific domains. Plausibility is equated with simplicity of representation, but Thagard does not offer any explanation of how these detailed representations are constructed. Overall, theories of reasoning involving plausibility have tended to overreach themselves without having the weight of evidence behind them. It is not clear from the literature whether plausibility has some role to play in reasoning, or whether reasoning has some role to play in plausibility.

2.2.4 – Plausibility in Theoretical and Computational Modelling

Plausibility has been used in theoretical and computational models across a wide variety of fields (Costello & Keane, 2000, 2001; Halpern, 2001; Lapata, McDonald & Keller, 1999). In many cases, plausibility is implemented as an operationalised metric. For example, Halpern (2001; Friedman & Halpern, 1996; see also Shafer, 1976) has created what he terms *plausibility measures*, but this is not intended to be a model of human plausibility judgement. Rather, the measures

constitute a mathematical metric of uncertainty for use in fuzzy logic and have little in common with what constitutes plausibility in the cognitive sense.

In their Constraint Theory of conceptual combination, Costello and Keane (2000; 2001) identified plausibility as a key constraint in generating interpretations for novel noun-noun compounds. For example, for the compound *shovel bird* the interpretation "*a bird which uses a shovel to dig for food*" is less acceptable because it is not particularly plausible. On the other hand, the interpretation "*a bird with a flat beak it uses to dig for food*" is more acceptable because it is more plausible. Thus, Constraint Theory operationally defines plausibility / acceptability as the degree to which the properties of an interpretation are consistent with prior knowledge. The C_3 model is the computational implementation of constraint theory, and computes plausibility as part of the process of generating interpretations. For the above example, the second interpretation is calculated as relatively plausible because there are several stored instances of birds having beaks of a particular shape, while the first interpretation is calculated as relatively implausible because there are no stored instances of a bird using a tool. In this way, C_3 models plausibility as the overlap of features between the current item and related instances in prior knowledge.

In the field of computational linguistics, Lapata, McDonald and Keller (1999) found that the distributional similarity of adjective-noun pairs could be used to model their plausibility. The distributional information of a particular word is calculated in corpus linguistics by counting the frequency of what words occur in its surrounding context. Across a large corpus, two words that tend to occur in the same contexts will be calculated as distributionally similar even though they may never

have occurred together. For example, the words *strong* and *tea* are more distributionally similar than the words *powerful* and *tea*. Lapata et al. found that this difference in distributional similarity could predict that people judge *strong tea* to be more plausible than *powerful tea*. They concluded that the strongly lexicalist nature of adjective-noun combinations was what allowed their plausibility to be operationalised as their distributional similarity.

Evaluation of the Theoretical and Computational Modelling Literature

Two notions of plausibility can be seen in the theoretical and computational modelling literature. First, Costello and Keane's work in conceptual combination suggests that plausibility judgement may be performed by comparing a particular concept to the contents of memory; the greater the match with prior knowledge, the more plausible the concept. However, while this theoretical notion seems relatively broad, the computational model has a much narrower focus. The C_3 model calculates the plausibility of concepts by counting the features that overlap with stored instances, and it is unclear how this paradigm can be extended to the plausibility of discourse scenarios that describe events. Second, Lapata et al. suggest an entirely different basis for plausibility judgement, namely the distributional properties of words themselves; the more distributionally similar the words, the more plausible their combination. However, this work is also rather difficult to generalise. Although the plausibility of adjective-noun pairs correlates with their distributional similarity, it is not certain whether a metric that works for word combinations will work in the same way for discourse scenarios that describe events. In summary, while this work gives some clues as to how conceptual or distributional properties

may affect specific types of plausibility judgement, their narrow focus prevents them from delivering a good account of plausibility. Further empirical work would be needed to test their generalisability.

2.2.5 – Summary of the Treatments of Plausibility

The ways in which plausibility has been treated across the literature can be roughly divided into two groups: those that adhere to the common view that plausibility stems from consistency with prior knowledge, and those that do not. The principal difference between these groups is whether they view people as judging the plausibility of particular events in the world, or the plausibility of linguistic descriptions of those events.

In the first group, we have the majority of the studies discussed here. From the "cognitive shortcut" mechanism to the various theories of reasoning and conceptual combination, much research considers plausibility judgement to be based on how well new information fits with our existing knowledge. Some of these types of plausibility judgement seem to be quite rapid while others do not. For example, the type of plausibility judgement used in place of direct retrieval (e.g., Reder, 1982) is a rapid process as it is faster than normal retrieval. On the other hand, the type of plausibility judgement involved in evaluating arguments (Smith et al., 1993) or connecting events in stories (Black et al. 1986) seems to be a slower process that has no particular time constraint. However, although they operate under different time constraints, these types of judgement all involve comparing the current conceptual representation of a scenario to prior knowledge. In other words, this research holds

that it is the global, concept-level properties (Hess, Foss & Carroll, 1995) of the scenario that are analysed during plausibility judgement, and that it is the events themselves rather than their description that matters.

In the second group, we have the studies that regard plausibility as stemming from word-level properties rather than from some sort of conceptual consistency. Again, some of these plausibility judgements are performed rapidly while others are not. For example, the plausibility judgements performed in sentence parsing (e.g., Pickering & Traxler, 1998) seem to be quite rapid processes. In contrast, rating the plausibility of adjective-noun pairs (Lapata et al., 1999) is a slower process that does not have any particular time constraint. However, irrespective of time constraints, such judgments are all rooted at the word level and appear to have the same underlying source. Lapata et al. showed that the distributional similarity of adjective-noun pairs can predict their plausibility, which raises the possibility that other word-level plausibility judgments may also be based on distributional information. For example, Pickering and Traxler found that an implausible verb-object match (e.g., "*sailed the magazine*") caused people to re-analyse sentences in a way that a plausible verb-object match (e.g., "*edited the magazine*") did not. Since *edited* and *magazine* are quite distributionally similar (i.e., they tend to occur in the same contexts), while *sailed* and *magazine* are not (i.e., they tend not to occur in the same contexts), it is possible that plausibility judgements made during parsing are also based (at least in part) on distributional analysis. In other words, this research suggests that it is the local word-level properties (Hess et al., 1995) of the scenario that are analysed during plausibility judgement, and that it is the description itself rather than its events that matters.

In brief, the literature leaves us with two main perspectives on what plausibility judgement entails: sometimes plausibility is based on the consistency of events with existing knowledge, and sometimes it is based on the word-level distributional information in the linguistic description. Resolving these apparently dichotomous accounts of plausibility is not a trivial task. To begin with, we must examine the nature of both knowledge representation and distributional information, and attempt to find some common ground between them both. It is to this end that the following sections review the literature of discourse comprehension and distributional analysis.

2.3 – Discourse Comprehension

The field of discourse comprehension is concerned with how people understand information that is communicated through language. Much of this research focuses on how we mentally represent the information we are given, and how we infer connections and elaborations using our knowledge of the world. This section discusses such research, and examines why a proper understanding of how people represent information is essential to the analysis of plausibility.

The comprehension of a discourse involves constructing a mental representation of the described situation, aided by cues provided by the linguistic input and using inferences from prior knowledge (e.g., Gernsbacher, 1990; Johnson-Laird, 1983; Kintsch, 1998; Singer, Graesser & Trabasso, 1994; van Dijk & Kintsch, 1983; Zwaan, Kaup, Stanfield & Madden, 2001). Given a particular text to understand, there are several elements that comprise the comprehension process;

relevant information must be activated, inferences must be made, and all necessary information must be brought together into a coherent representation. Each of these elements contributes to the final form of the mental representation, and so plays a role in how the plausibility of the text is judged.

2.3.1 – The Role of Information Activation

For any given concept, we have a large amount of information available, but only some of it will be relevant to the current context in which we encounter it. For example, if we were reading a text about cooking chicken for dinner, then it would be more useful to have information available about chicken as a food than chicken as a fluffy favourite in a petting zoo. In a text, the words themselves (the local context) and the situation that they describe (the global context) both activate related information in background knowledge (e.g., Duffy, Henderson & Morris, 1989; Hess, Foss & Carroll, 1995; Meyer & Schvaneveldt, 1976; Swinney, 1979).

Studies of simple lexical priming show that individual words can be primed by related words in the preceding context (e.g., Foss, 1982; Meyer & Schvaneveldt, 1976; Swinney, 1979). For example, Foss (1982) showed that people are faster to process the word "fish" when it is presented following the phrase "gills and fins", compared to when it follows an unrelated phrase such as "spots and stripes". Words can also act in combination to prime items that they would be unable to prime individually. For example, Duffy et al. (1989) presented sentences such as "*the barber trimmed the mustache*" one word at a time, and measured the time it took people to name the last word. Even though neither *barber* nor *trimmed* are strongly

related to *mustache*, Duffy et al. found that the naming time for *mustache* was facilitated by the preceding context.

In addition to words being primed by related information, many studies have shown that different aspects of a concept can be highlighted by subtle changes in its global discourse context (e.g., Anderson & Ortony, 1975; Barclay, Bransford, Franks, McCarrell & Nitsch, 1974; Halff, Ortony & Anderson, 1976; Keane, 1985; McKoon & Ratcliff, 1988). For example, McKoon and Ratcliff (1988) found that after being presented with the sentence "*The little girl found a tomato to roll across the floor with her nose*", people were faster to confirm that tomatoes were "round" than that they were "red". Activated information can take the form of other words and concepts (e.g., the *redness* or *roundness* of a tomato), but it can also take the form of related knowledge that may be relevant to the situation. For example, reading the sentence "*The hiker shot the injured deer*" activates the related information that bullets kill animals, which in turn primes the outcome "*The deer died*" (Halldorson & Singer, 2002). This kind of global context priming has been shown to arise from the conceptual discourse representation (Hess et al., 1995) and from the knowledge incorporated in the representation when inferences are made between events (Halldorson & Singer, 2002).

Evaluation of the Role of Information Activation

Comprehending discourse involves integrating given information with our own knowledge of the world. The priming studies reported above illustrate that relevant information can be activated in long-term memory by the words themselves and also by the existing mental representation of the discourse. Both of these types

of priming are important to the comprehension process as they ease the problem of choosing what is relevant, in real time, from a vast repository of prior knowledge. This raises the possibility that the plausibility of a discourse may depend, in part, on whether the priming process succeeds in activating relevant information. As discussed in section 2.2.5, much research considers plausibility judgement to be based on how well new information fits with our existing knowledge. As it is not practicable to compare new information in real time to our *entire* repository of prior knowledge, it is possible that plausibility judgement is based on how well new information fits with our existing *activated* knowledge. In other words, an implausible scenario may be one that fails to activate relevant information in prior knowledge. Information activation may play a role in plausibility judgement by constraining the portion of background knowledge that is available for comparison with a given scenario.

2.3.2 – The Role of Inferencing

The scenario described in a text is not always complete, and may contain gaps or discontinuities. When we are reading a text and encounter a gap in information, we try to fill it with knowledge of what we know or assume about the world. For example, we may read about the events of pouring water on a fire and the fire going out. In order to fully comprehend this situation, we must find the connection between these two events – namely the causal inference that water *caused* the fire to extinguish – and then integrate the inference as part of the representation (see e.g., Halldorson & Singer, 2002; Keenan, Baillet & Brown, 1984; Kintsch & van Dijk, 1978; Singer & Halldorson, 1996). Researchers have identified

several different types of inference that use different amounts of prior knowledge to connect parts of a discourse (see Graesser, Millis & Zwaan, 1997, for a review).

McKoon and Ratcliff (1992) outline a minimalist hypothesis of inferencing, which holds that people only make the inferences that are necessary to make sense of a discourse, as well as those that are easily accessible when a sentence is presented. Some inferences are made to connect adjacent clauses through argument overlap (Kintsch & van Dijk, 1978; McKoon & Ratcliff, 1992). For example, the sentences "*Mary went to the restaurant. She ordered a salad.*" are connected by inferring the referential match between "Mary" in the first sentence and "she" in the second sentence. These inferences are concerned with maintaining local coherence, and can usually be made without requiring extra background knowledge. It has also been shown that people often infer causal antecedents when they read a text (McKoon & Ratcliff, 1986, 1992). For instance, people may infer from the above sentences that Mary was hungry. McKoon and Ratcliff argue that these inferences are made because they require little effort, as the background knowledge that they require (e.g., that people eat because they are hungry) is already in working memory or is highly active in long-term memory.

However, Singer et al.'s (1994) constructionist hypothesis of inferencing challenges McKoon and Ratcliff's assertions, and states that people make as many inferences as are necessary to explain why actions, events and states occur. It has been shown that when people read a text, they infer information about characters' motives and goals (Graesser, Singer & Trabasso, 1994; Singer et al. 1994). For example, from the sentences "*Mary ate the salad. She left a large tip.*" people may infer that Mary was pleased with the food and service in the restaurant. These

inferences are concerned with establishing and maintaining global coherence across the discourse, and usually require background knowledge to make connections between parts of the text. Singer et al. argue that people make these inferences even if they take more effort, and that the background knowledge they require will be retrieved from long-term memory as necessary.

Evaluation of the Role of Inferencing

Comprehending a text involves inferring connections between parts of the discourse, using our own knowledge of the world. As discussed in section 2.2.2, plausibility depends on the ability to make inferences, as an implausible text is one in which people cannot infer connections between events (Black et al., 1983). However, the discourse studies discussed above identify different *types* of inferential connection. Some inferences can be made rapidly with little effort and do not require much extra information from prior knowledge (McKoon & Ratcliff, 1992). In contrast, other inferences take more time and effort to make and require relevant information from prior knowledge (Singer et al. 1994). This raises the possibility that plausibility may be influenced not only by the presence of an inferential connection (Black et al., 1983), but also by how much effort and prior knowledge the inference required. In other words, an inference that draws in a lot of background knowledge will require more effort to make than an inference that draws in little background knowledge, and this extra effort may make the scenario seem less plausible. As earlier discussed, much research considers plausibility to be based on how well new information fits with our existing knowledge. The role of inferencing

in plausibility may be to integrate prior knowledge with the discourse in a way that allows us to judge the quality of this fit.

2.3.3 – The Role of Represented Information

When we read a text, we construct a mental representation of the described situation. However, there is a vast amount of information that could possibly be represented. For example, the sentences "*The cat sat on the mat. It fell asleep.*" may be represented by a verbatim memory of the sentence, by a situation model of cat falling asleep because it was tired, by a mental image of a sleeping cat, and so on. Much discourse research has focussed on ascertaining exactly what information people represent when they read a description of a scenario (e.g., Gernsbacher, 1990; Johnson-Laird, 1983; Kintsch, 1998; Singer, Graesser & Trabasso, 1994; van Dijk & Kintsch, 1983; Zwaan, Kaup, Stanfield & Madden, 2001).

Early discourse research assumed that referential and some inferential information is represented in a *textbase*, a notion proposed by Kintsch and van Dijk (1978) which has been adopted by many subsequent researchers (e.g., McKoon & Ratcliff, 1992; Myers & O'Brien, 1998). In essence, a textbase is a network of propositions. Comprehenders transform the natural language input into propositional form, and connect these propositions when they share an argument. For example, the above sentences could be represented propositionally as:

```
sit(cat, mat)

sleep(cat)
```

because the *it* in the second sentence refers to the *cat* in the first sentence. When comprehenders cannot connect a new proposition to the existing representation in working memory, they make a backward or *bridging inference* by retrieving knowledge from long-term memory that shares arguments with the new proposition and those currently in working memory. Also, comprehenders may sometimes make *elaborative inferences* that use propositions from long-term memory that share arguments with a new proposition, but do not connect it to the rest of the representation. Kintsch and van Dijk found that the textbase representation successfully predicted how well people could recall the "gist" of stories they had read up to three months previously.

Other researchers quickly proposed that language comprehension entails a lot more than the construction of a textbase, and that a *situation model* of the discourse was needed to deal with causal, temporal, and spatial aspects (Anderson, 1983; Johnson-Laird, 1983; Gernsbacher, 1990, 1997; Kintsch, 1998; Singer et al., 1994; van Dijk & Kintsch, 1983; Zwaan, Magliano & Graesser, 1995). There is little consensus between researchers as to how exactly a situation model is structured. For example, Kintsch's (1998; Singer & Kintsch, 2001) Construction-Integration theory proposes a connectionist network of propositions, while Gernsbacher's (1990, 1997) Structure Building Framework rejects propositions as a possible representation and instead proposes layered structures of (undefined) "memory nodes". However, situation models are agreed to include many subtle aspects of discourse representation that a textbase cannot. For example, take these two sentence pairs:

1a) "The actress walked across the stage. A moment later she collapsed."

1b) "The actress walked across the stage. An hour later she collapsed."

Zwaan (1996) found that the different temporal expressions differentially affected processing; people are faster to confirm that the word *walked* appeared in the text for pair 1a than for pair 1b. In other words, the representation of the first event (*walked*) is more available to a comprehender if it is temporally close to the current state of affairs (*collapsed*). This temporal sensitivity cannot be represented by a textbase alone, but such information forms an integral part of a situation model. Apart from temporal information, a substantial amount of empirical evidence has emerged showing that people represent causal, temporal, spatial, motivational and protagonistic information in a discourse (see Zwaan & Radvansky, 1998, for a review).

More recently, some researchers have examined how perceptual information is represented when people read a text, and whether the entire discourse can be represented as a *perceptual simulation* (Barsalou, 1999; Zwaan et al., 2001). A perceptual simulation of a discourse is a recreation of perceptual experience; the verbal input is regarded as activating perceptual symbols in long-term memory, which are then integrated into a dynamic representation of the scenario. There is some evidence of perceptual effects in text comprehension that situation models are not equipped to deal with. Recent neuroimaging studies have shown that spatial processing regions of the brain can be stimulated by sentence comprehension tasks (Carpenter, Just, Keller, Eddy and Thulbom, 1999; Mellet, Tzourio, Crivello, Joliot, Denis & Mazoyer, 1996). For example, reading sentences such as "*The star is above the plus*" produced activation in the left parietal lobe around the intraparietal sulcus, a region usually activated during mental rotation tasks (Carpenter et al. 1999). Stanfield and Zwaan (2001; see also Zwaan, Stanfield & Yaxley, 2002) have also

tested the perceptual simulation view of representation by presenting sentences that mention objects with implied orientation, such as "*Rick put the pencil in the cup*" or "*Rick put the pencil in the drawer*", followed by a picture of an object (e.g., a pencil). People were faster to verify that a pencil had been mentioned in the sentence when it was pictured in the orientation implied by the sentence (i.e., pictured vertically for the *cup* sentence, pictured horizontally for the *drawer* sentence). Stanfield and Zwaan interpreted this as strong evidence that a perceptual simulation was represented, as the perceptual information (about object orientation) was only implied by the discourse and was not given explicitly.

Evaluation of the Role of Represented Information

Comprehending a text means that all necessary information must be brought together into a coherent representation. As earlier discussed, many plausibility judgements seem to be based upon comparing new information to existing knowledge. It is therefore essential to know what information is represented, as only then can we know exactly what is compared to our existing knowledge. The evidence discussed above shows that discourse is represented with quite a high degree of complexity, and that the representation includes verbal, referential, causal, spatial, temporal, motivational and perceptual information. Indeed, most theories and models of comprehension attempt to represent as wide a variety of information as possible (e.g., Gernsbacher, 1990; Kintsch, 1998; Zwaan et al., 2001). Any or all represented information may bear upon plausibility. In other words, for a text to be plausible it may have to be consistent with our existing knowledge on referential, causal, temporal, and other levels. The represented information plays a role in

plausibility by portraying a wide variety of the discourse's subtleties, so that we can accurately examine how well the discourse fits with our existing knowledge.

2.3.4 – Summary of Discourse Comprehension

The discourse comprehension literature agrees that the comprehension of a scenario is essentially the construction of a mental representation of the described situation, aided by the cues provided in the linguistic input. As earlier discussed, plausibility judgement involves examining this mental representation to ascertain how consistent the scenario is with our knowledge of the world. In general, accounts of comprehension assume that the mental representation integrates our knowledge of the world with the information in the scenario through the inferences that connect events. As this representation is being built, a lot of knowledge is activated in long-term memory. Some of this activated knowledge supports essential bridging and other inferences, whereas other activated knowledge may prove redundant, never being directly used for inferencing. The final form of the mental representation will include causal, temporal, motivational, and other information. However, the discourse comprehension literature tends to gloss over some important comprehension subprocesses. For example, how exactly do priming mechanisms activate relevant knowledge in long-term memory? If activated information affects plausibility, then an account of knowledge priming in comprehension must be specified. Similarly, are different types of inference (e.g., causal, temporal) represented differently? If inferences affect plausibility, then an account of the representation of different types of inference must be specified. In short, a full

account of human plausibility judgments should incorporate a full account of comprehension along with plausibility assessment criteria.

2.4 – Distributional Analysis

The distributional analysis of visual, auditory, and linguistic information has been implicated in many cognitive phenomena. This section discusses such research, and examines how linguistic distributional analysis in particular may be relevant to the analysis of plausibility.

Distributional analysis is the examination of emergent statistical patterns of structure. Our environment is full of structural regularities and a sensitivity to these regularities has been linked to children's extraordinarily fast rate of learning in the first years of life. In particular, it has been proposed that people can exploit statistical regularities in their environment to accomplish a range of conceptual and perceptual learning tasks. For example, it has been shown that infants as young as two months can learn implicit statistical patterns in visual sequences of shapes (e.g., yellow circle, pink diamond), and are sensitive to when the same shapes appear in new sequences (Kirkham, Slemmer & Johnson, 2002). These same learning mechanisms have also been shown to apply to tone sequences (Saffran, Johnson, Aslin & Newport, & Aslin, 1999). Most interesting, however, is the application of these statistical learning mechanisms to linguistic input. Infants from eights months old appear to learn word boundaries from listening to a continuous stream of nonsense syllables with underlying statistical patterns (Saffran, Aslin, & Newport,

1996; Saffran, Newport, & Aslin, 1996). The syntactic structure of language is also full of distributional patterns, and it has been suggested that children's sensitivity to these patterns contributes to their learning of syntactic categories (Redington & Chater, 1997; Redington, Chater & Finch, 1998). Furthermore, the semantic structure of language also contains distributional patterns. These patterns can be modelled computationally, and such distributional models of language have been shown to capture information that has been implicated in many areas of cognition.

2.4.1 – Distributional Models of Language

The distributional structure of a language can be seen in the knowledge of what words tend to occur in the context of others. For example, the word "scalpel" tends to be found in a discourse context with the word "surgery" and not with the word "gardening". Similar words are used in similar contexts, which allows two words to be linked even though they may never appear together. Distributional models operate on the principle that if a sufficiently large sample of a language is taken, it can provide useful information about the semantic properties of lexemes in that language. Several such models of English have been created, such as the Hyperspace Analogue to Language (HAL: Burgess & Lund, 1997) and Latent Semantic Analysis (LSA: Landauer & Dumais, 1997).

In essence, distributional information can be modelled using 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. A contextual distribution is calculated for a

word by counting the frequency with which it co-occurs with every other lexeme (that is, are used together within a particular context, such as a paragraph or moving-window) in the corpus being analysed. In this way, every word may be summarised as a vector – or point in high-dimensional distributional space – showing the frequency with which is it associated with other lexemes in the corpus.

While LSA and HAL both adopt the general approach outlined above to generate their distributional space, the exact parameters of the models are different. For example, HAL leaves its high-dimensional space intact at approximately 140,000 dimensions (i.e., from 140,000 unique lexemes in the corpus). LSA, on the other hand, reduces the dimensionality of its distributional space through singular value decomposition with principal components analysis. This process extracts the most relevant dimensions from the high-dimensional space (i.e., those dimensions with the most informative co-occurrence frequencies), and results in a distributional space of 300-400 dimensions. The advantage of dimensionality reduction is that it allows for a more parsimonious model of distributional information with lower computational overheads; indeed, the creators of HAL and LSA agree that the number of cognitively plausible dimensions lies in the hundreds rather than in the thousands (Landauer & Dumais, 1997; Burgess, Livesay & Lund, 1998). In LSA's distributional space of 300-400 dimensions, every word is still summarised as a unique vector, or point in high-dimensional space. Similarly, LSA may represent a whole sentence as a single point in distributional space by using the weighted sum of constituent word vectors to denote tracts of text. Whether or not a model performs dimensionality reduction, a high-dimensional representation means that two words that occur in similar linguistic contexts (i.e., that are distributionally similar) will be

positioned closer together in this space than two words that do not share as much distributional information.

This type of distributional information has been implicated in many cognitive phenomena. As already discussed, Lapata et al. (1999) showed that distributional similarity could predict people's plausibility ratings of adjective-noun pairs. It has also been shown that distances in distributional space can be used to predict priming effects (Landauer & Dumais, 1997; Lund, Burgess & Atchley, 1995). Additionally, Connell and Ramscar (2001a, 2001b; see also Connell, 2000) showed that the typicality of category membership could be predicted by distributional distance, as could the effects of context on typicality. For example, people tend to rate *robin* as the most typical *bird*, but in the context of a *bird that hunts mice* tend to rate *owl* as most typical. Distributional models have also been shown to select synonyms with accuracy high enough to pass a Test of English as a Foreign Language (TOEFL test) (Landauer & Dumais, 1997). When used in combination with existing models, distributional information has been shown to improve performance in modelling more complex phenomena, such as metaphor interpretation (Kintsch, 2001) and analogical retrieval (Ramscar & Yarlett, 2003). Indeed, some researchers have suggested that distributional information may provide a useful means of priming knowledge in long-term memory relevant to whatever task is in hand (Burgess, Livesay & Lund, 1998; Halldorson & Singer, 2002; Kintsch, 1998, 2000; Kintsch, Patel & Ericsson, 1999).

However, it is important not over overestimate the role of distributional information. When a task involves additional inferencing and relation-building, the predictive power of distributional models is much 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 point 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 grey hair and sharp teeth. Indeed, distributional models have difficulty "interpreting" any linguistic description as they are essentially blind to word order, and are therefore insensitive to important syntactic cues such as negation. Some of the strongest criticisms of distributional models surround the issue of symbol grounding; in a distributional model, words are defined by their relations with other abstract symbols and are not tied to perceptual experience or action. For example, Glenberg and Robertson (2000) showed that distributional information did not distinguish between sensible novel situations (such as using a sweater filled with leaves as a pillow) and nonsensical novel situation (such as using a sweater filled with water as a pillow).

2.4.2 – Summary of Distributional Analysis

People appear to be sensitive to distributional patterns in the environment across a range of modalities. However, it is linguistic distributional information that interests us from the perspective of plausibility. This kind of distributional information appears to be exceedingly useful across a wide variety of cognitive phenomena (much like plausibility itself), and its direct relevance to plausibility has been clearly shown by its ability to predict adjective-noun plausibility ratings (Lapata et al., 1999). The automated construction of distributional models from

corpora also offers them a degree of objectivity usually unavailable to hand-coded models of knowledge representation. Nevertheless, models of distributional information have received much valid criticism. Meaning is not grounded in distributional models, nor can inferencing or analogical mapping be performed with a solely distributional approach.

It therefore seems fair to conclude that distributional models are not models of meaning, but rather that they model a particular form of linguistic knowledge that reflects the distributional relationships between words. Many simple cognitive linguistic phenomena emerge from this distributional knowledge (e.g., priming effects, typicality), but tasks of greater complexity are beyond its scope. Yet the success in incorporating distributional models into models of larger cognitive processes (e.g., metaphor interpretation, analogical retrieval) suggests that distributional information can still play an important role in complex cognitive tasks, and may even provide a vital means of activating task-relevant knowledge in long-term memory. In order to capture all the different effects that have been shown to bear upon plausibility, it seems reasonable that a full account of human plausibility judgments should incorporate distributional information along with other more complex reasoning mechanisms.

2.5 – Literature Review Conclusions

The theoretical and empirical literature relevant to the study of plausibility spans a wide variety of fields in cognitive psychology. Plausibility has been examined from many perspectives, including memory research, reasoning, text

linguistics, and conceptual combination. In addition, we have argued that work in discourse comprehension and distributional analysis is also relevant to the processes involved in plausibility judgement. The breadth of coverage in this literature is important to the study of plausibility, as it offers us some valuable insights.

First, the empirical work reviewed in this chapter suggests that plausibility is affected by two distinctly different factors, concept-coherence and word-coherence. Some plausibility judgements are influenced by *concept-coherence* – that is, by the conceptual consistency of the events and other elements in a scenario (e.g., Costello & Keane, 2000; Reder, 1982; Smith et al., 1993). The discourse comprehension literature expands this notion, suggesting that concept-coherence may depend not only on the information given in the scenario, but also on the information drawn in from our own prior knowledge as we make inferences between events (e.g., Halldorson & Singer, 2002). However, other plausibility judgements are influenced by *word-coherence* – that is, by the distributional distance of the words in the description (e.g., Lapata et al., 1999). While the notion of word-coherence seems quite separate to that of concept-coherence, the distributional analysis literature suggests they are not mutually exclusive, and that distributional knowledge may play an important supporting role in complex cognitive processes (e.g., Kintsch et al., 1999; Ramscar & Yarlett, 2003). To establish exactly how concept-coherence and word-coherence influence plausibility, we should carefully examine the effects and interactions of these factors in our empirical work. The findings of these experiments can then inform our theory of plausibility, and allow us to detail the role of different types of knowledge in the processes involved in plausibility judgement.

Second, the literature suggests that plausibility should be examined using different experimental paradigms. Some plausibility judgements carefully determine exactly *how* plausible we find a scenario (e.g., Black et al., 1986; Lapata et al., 1999). Other judgements rapidly determine *whether* a scenario is plausible or implausible (e.g., Reder, 1982; Pickering & Traxler, 1998). In other words, these paradigms allow us to examine, respectively, the product and time course of the plausibility judgement process. To gain a proper understanding of plausibility, we should explore the influences of concept-coherence and word-coherence by using both of these experimental paradigms. The empirical findings can then inform our theory of plausibility, and allow us to describe how different influences bear upon the speed and product of plausibility judgement.

Third, the literature suggests that plausibility judgement is both an objective and subjective assessment. We have already discussed the suggestion that plausibility is influenced by concept-coherence and word-coherence; both objective factors based on assessing the events and their description. In addition, some types of plausibility have been modelled with objective computational metrics (Costello & Keane, 2000; Lapata et al., 1999). However, it can be argued that every human judgement is essentially subjective, and that people are influenced by subjective factors even when they strive to make an objective assessment of the facts. Rather than being a purely objective judgement, plausibility judgement also appears to take into account subjective factors such as prior opinions and trustworthiness (Pohl, 1998; Thompson & Kliegl, 1991). This suggests that plausibility judgement has an objective *core* of concept-coherence and word-coherence, and that people are influenced to different extents by various subjective influences. To establish the

importance of this objective core in plausibility judgement, we should test our theory of plausibility by implementing it computationally. The accuracy of this cognitive model can then further inform our theory of plausibility, and allow us to determine the importance of concept-coherence and word-coherence in plausibility judgement.

CHAPTER 3 – PLAUSIBILITY RATINGS

3.1 – Outline of Experiments

There are many different ways in which we can judge plausibility. Some of these judgements can carefully determine exactly *how* plausible we find a scenario, and so allow us to examine what makes one scenario more plausible than another. Other judgements can rapidly determine *whether* a scenario is plausible or implausible, and so allow us to examine what makes one scenario faster to judge than another. In this chapter, we investigate the former type of plausibility judgement using a paradigm involving plausibility ratings. This paradigm allows us to examine how concept-coherence and word-coherence may exert an influence on the perceived plausibility of a scenario.

In Chapter 1, we introduced the central ideas and components in the Knowledge-Fitting Theory of Plausibility. To recap, concept-coherence is concerned with how well the scenario fits with prior knowledge. Generally speaking, if the events in a scenario are connected in a way that fits closely with our knowledge of the world, then the scenario will appear quite plausible. The following experiments manipulate concept-coherence by using different inferential connections between events, which allows us to examine if people find some inference types more plausible than others. Word-coherence, on the other hand, is concerned with the distributional distance between the sentences in the scenario description. If the

sentences that describe a scenario are far apart in distributional space, then this may affect how plausible people find the scenario. In the experiments that follow, word-coherence is manipulated by describing the same scenario in different ways that vary distributional distance, which allows us to examine if people find distributionally close or distant sentences more plausible.

Three experiments are reported that ask people to rate the plausibility of simple event scenarios (e.g., "*The bottle fell off the shelf. The bottle smashed.*" See also Connell & Keane, in press, 2002a, 2002b). In Experiment 1, we manipulate the concept-coherence of scenarios (via their inferences) while holding the word-coherence of the descriptions constant. In Experiment 2, we manipulate the word-coherence of scenario descriptions (via the distributional distance between sentences) and cross it with a key concept-coherence manipulation (attributal versus causal inferences). Finally, Experiment 3 uses the same manipulations in an alternate design that capitalises on individual differences in distributional knowledge to maximise potential word-coherence effects.

3.2 – Experiment 1: Plausibility Ratings and Inference Type

In this experiment, the distributional distance of the event descriptions was held constant, and the types of inferences invited by the sentence pairs manipulated (see Appendix A for full set of materials). Four distinct categories of sentence pair were used: causal, attributal, temporal, and unrelated. As we described in Chapter 1, the Knowledge-Fitting Theory of Plausibility sees concept-coherence as being about the background knowledge that is needed to make a given scenario fit

what we know of the world. Broadly speaking, if the events described in the sentence pair are strongly connected – as determined by the closeness of fit with prior knowledge – then the scenario will appear more plausible. In contrast, weakly connected events will make the scenario seem less plausible. This simple explanatory model allows us to make certain predictions for the four inference conditions in this experiment; specifically, it predicts a decreasing trend in perceived plausibility with the following ordering causal > attributal > temporal > unrelated conditions (see Table 3.1 for sample materials). Let us consider how we arrive at these predictions.

Table 3.1 – 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

First, we assume that the more primed knowledge that was used to connect the events in the scenario, the stronger the actual connection between them. Some scenarios can be represented by using background knowledge that has already been primed by the description, while others can only be represented by retrieving different knowledge from long-term memory. For example, the causal scenario in Table 3.1 is likely to prime the background knowledge that fragile things smash when they fall on hard surfaces, which is then used to infer the connection that the bottle smashed because it was fragile and hit the floor (see e.g., Halldorson & Singer,

2002; Singer & Halldorson, 1996). In this respect, the events in the causal scenario are very strongly connected because the knowledge necessary to connect them has already been primed. On the other hand, the attributal and temporal scenarios are less likely to prime the knowledge that the inferential connection will need (e.g., that glass can be pretty, or that glass can be shiny and that shiny things sparkle when they catch the light). That is, attributal and temporal scenarios are less strongly connected than causal scenarios because they are less likely to prime such useful knowledge. Finally, the unrelated scenario in Table 3.1 is not at all likely to prime the knowledge that high temperatures can melt a bottle, that metal can be heated to high temperatures, and that metal needs a heat source to get that hot. One can only infer the connection between the events (the bottle melted because the floor was extremely hot, because it was made of metal and something had heated it up) by retrieving the necessary knowledge from long-term memory. That is, the events in the unrelated scenario are quite weakly connected. The greater the usage of primed information in making each scenario fit prior knowledge gives us the following ordering of inference types: causal > (attributal = temporal) > unrelated, with plausibility decreasing from left to right.

Second, the strength of the temporal and attributal connections is distinguished in an additional way; namely by the amount of background knowledge that was necessary to connect the events. For example, the temporal scenario in Table 3.1 is represented by inferring a temporal connection between the two events; namely, that the bottle sparkled because it caught the light after hitting the floor. This connection incorporates the background knowledge that bottles can be shiny, and that shiny things sparkle when they catch the light. In contrast, the attributal

scenario in Table 3.1 is represented merely by inferring a co-referent for the attribute *pretty*, namely *bottle*. The smaller quantity of extra information that was needed to make the *pretty* scenario fit prior knowledge means that it is more strongly connected than the *sparkled* scenario, and should seem more plausible.

So, this simple explanatory model of concept-coherence, using the strength of the connection between the two sentences, gives us the predicted ordering of decreasing plausibility across the causal > attributal > temporal > unrelated conditions.

3.2.1 – Method

Materials

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 for full set of materials), where the unrelated pairs provided a control in which no obvious inferences could be made. Rather than use free-generation of sentences to provide the inferential variants, we chose to use experimenter-derived sentence pairs in order to facilitate controls of distributional distance and word frequency (detailed below). The causal pairs were designed to invite a causal inference by using a second sentence (S2) that was a reasonably direct causal consequence of the first sentence (S1) (e.g., "*The bottle fell off the shelf. The bottle smashed*"). The attributal pairs invited an attributal inference by using an S2 that referred to an

attribute of its subject in a way that was not causally related to S1 (e.g., "*The bottle fell off the shelf. The bottle was pretty*"). The temporal pairs invited a temporal inference by using an S2 that could occur in the normal course of events, regardless of the occurrence of S1 (e.g., "*The bottle fell off the shelf. The bottle sparkled*"). The unrelated pairs used an S2 that described an event to that was unlikely to occur in the normal course of events and had no obvious causal link to S1 (e.g., "*The bottle fell off the shelf. The bottle melted*").

In addition, the second sentence of each of these 4 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 if the repetition of terms would facilitate participants' ability to construct inferences between the two sentences in the pair. Thus, the sentence pairs captured two variables: *inference type* (causal, attributal, temporal, unrelated) and *noun type* (repeated or alternate).

The distributional distance of each sentence pair was controlled by comparing their scores using Latent Semantic Analysis (LSA: Landauer & Dumais, 1997). All 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 first-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 of the distributional scores of the sentence pairs revealed a significant difference between noun types [Repeated M=0.72, Alternate M=0.29;

$F(1, 88)=217.780$, $p<0.0001$, $MSe=0.020$], as 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; $F<1$] (confirmed by pairwise comparisons using Bonferroni adjustments, all ps>0.9) and no interaction between noun type and inference type [$F<1$].

Word frequency was also controlled for 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 analysis of variance of the frequency scores showed no reliable difference between the inference types [Causal M=1462.5, Attributal M=3020.7, Temporal M=870.0, Unrelated M=1025.9; $F(3, 44)=2.007$, $p>0.1$, $MSe=5778177$; with Bonferroni adjustments, all pairwise comparison ps>0.2] or between noun types [Repeated M=6941.5, Alternate M=5833.2; $F<1$].

Design

The design treated inference type as a within-participant and between-item variable, and treated noun type as within-participants and within-items. 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 three 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 the different groups in the experiment. All participants were student volunteers at University College Dublin.

Procedure

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 participants to note their 0-10 plausibility rating.

3.2.2 – Results and Discussion

The results show that plausibility is affected by subtle changes in the concept-coherence of simple event descriptions when different inferences are invited (see Figure 3.1). Importantly, these plausibility differences are purely due to concept-coherence, and occur when word-coherence is held constant (i.e., the distributional distance of sentence pairs were held constant across the different inference types). Table 3.2 gives the mean ratings for each condition – the causal

pairs were rated the most plausible (M=7.8), followed as 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 judgements, with the added finding that different inference types differentially affect plausibility.

All analyses of variance by participants and by items were performed by respectively treating participants (F_1) and sentences (F_2) as a random factor. A two-way analysis of variance by inference type and noun type found a significant effect of inference type on plausibility ratings [F_1 (3, 117)=84.57, p<0.0001, MSe=7.5721; F_2 (3, 44)=41.60, p<0.0001, MSe=16.746]. Planned pairwise comparisons revealed that all of the conditions were reliably different to one another using Bonferroni adjustments. No reliable effect of noun type was found [Fs<1]. There was also no significant interaction between inference type and noun type found [Fs<1], showing that repeating a term between the first and second sentences in a pair did not affect participants' ability to construct inferences between them. As predicted, the order of inference types (from causal > attributal > temporal > unrelated) was found to be reliable using Page's Trend Test by participants [L(40)=1147, p<0.0001] and by items [L(12)=341.5, p<0.0001].

Table 3.2 – Mean plausibility ratings for each condition in Experiment 1.

Noun Type	*Inference 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*

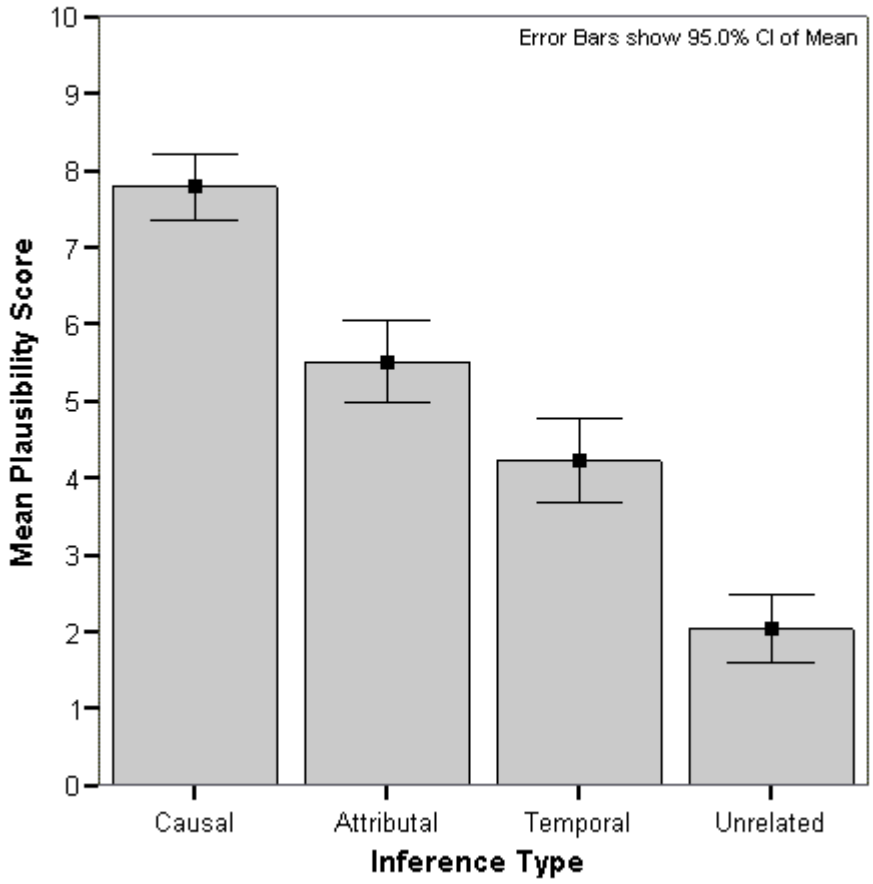


Figure 3.1 – Mean plausibility ratings per inference type in Experiment 1.

But, how can we be confident that the different sentence-pairs were indeed treated in accordance with the experimenter-defined categories? For example, if a temporal pair was interpreted using a causal relation, then the 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 75-100% agreement between the raters and the original classification of the pair. Of the 96 original sentence pairs, 72 met the criterion. A re-analysis 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). Again, as before, there was a significant effect of inference type [F_1 (3, 117)=124.98, p<0.0001, MSE=5.538; F_2 (3, 32)=21.02, p<0.0001, MSE=4.503], no effect of noun type [Repeated M=4.6, Alternate M=4.7; Fs<1], and no significant interaction [Fs<1].

To summarise, these results show that plausibility judgements are sensitive to the type of inference made between events in a scenario description. The strength of the inferential connection between the two sentences, and hence its perceived plausibility, is greatest in the causal pairs where primed knowledge is used to make the inference. The connection strength, and hence plausibility, is lowest in the unrelated pairs where no useful knowledge is primed and other background knowledge must be used to form the connection (which, indeed, may fail to be formed at all). Ranged in between are the attributal and temporal pairs, largely distinguished by the complexity of the inferential connection (i.e., the amount of prior knowledge that was necessary to connect events).

On the basis of previous research, one might not be surprised to learn that the plausibility of causal pairs was higher than that of the unrelated pairs, although we know of no previous study that has explicitly examined such a manipulation in a direct and controlled manner. The main novelty of the present experiment is the demonstration that there are four empirical distinguishable categories (causal, attributal, temporal, 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 specifically due to concept-coherence and not to the possible effects of word-coherence, to which we now turn in the next experiment.

Table 3.3 – Sample of sentence pairs used in Experiments 2 and 3.

Sentence 1	*Sentence 2*	*Inference Type*	*Distributional Distance*	*LSA Score*
The pack saw the fox.	The hounds growled.	Causal	Close	0.37
	The hounds snarled.	Causal	Distant	0.20
	The hounds were fierce.	Attributal	Close	0.19
	The hounds were vicious.	Attributal	Distant	0.12

3.3 – Experiment 2: Plausibility Ratings and Distributional Distance

In Experiment 1, we concentrated on the role of concept-coherence in plausibility, controlling for the possible influence of word-coherence. In this experiment, we examine word-coherence by crossing the variables of inference type (causal or attributal) and distributional distance (close or distant). As we described in Chapter 1, the Knowledge-Fitting Theory of Plausibility sees word-coherence as being about the distance between sentences in distributional space. In general, each sentence creates a distributional spotlight that activates a surrounding area of distributional space, which in turn primes relevant background knowledge in long-term memory. This means that the same scenario can be described in different ways with varying distributional distance. For example, although the two causal sentence pairs in Table 3.3 have essentially the same meaning (i.e., they both invite the same inference), they differ markedly in their distributional distance because *growled* is

much more likely to occur with the words in the first sentence than *snarled*. From a word-coherence perspective, the *snarled* sentence pair is more distributionally distant than the *growled* sentence pair, even though from a concept-coherence perspective both sentences invite the same causal inference. So, these two test items vary distributional distance while holding the inference type constant (see Table 3.3 for an example of attributal sentence pairs).

As in Experiment 1, we predict that the causal pairs will be rated as more plausible than the attributal pairs because of their stronger inferential connection. However, this experiment uses a different design that treats inference type as a between-participant variable, which would also lead us to anticipate that the difference between causal and attributal pairs may be smaller than in Experiment 1. Regarding the effect of distributional distance, previous research has shown that distributionally close adjective-noun combinations are rated as more plausible than the distributionally distant (Lapata et al., 1999). This would lead us to predict that people will rate the distributionally close pairs as more plausible than the distributionally distant pairs. 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 distributional measures 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 point 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 grey hair

and sharp teeth. Similarly, Glenberg and Robertson (2000) also find distributional information did not distinguish between sensible novel situations (such as using a sweater filled with leaves as a pillow) and nonsensical novel situation (such as using a sweater filled with water as a pillow). Since our sentence pairs involve causal, temporal, and other inferences, these studies would also lead us to anticipate that the difference between distributionally close and distant pairs may be smaller than that observed for adjective-noun combinations.

3.3.1 – Method

Materials

Fifteen basic sentence pairs were created and then modified to produce variants of the different materials. As in the last experiment, several causal and attributal variants were produced, each of which maintained the basic meaning of the original sentence pair (see Appendix B for full set of materials). Using LSA, the highest and lowest scoring pairs were selected as the distributionally close and distant pairs, respectively[1]. The causal sentence pairs had a mean LSA score of 0.42 for close pairs and 0.27 for distant pairs, while the attributal sentence pairs had a mean score of 0.35 for close pairs and 0.23 for distant pairs. A two-way analysis of variance showed a reliable difference between the distributionally close and distant scores of these materials [Close M=0.39, Distant M=0.25; $F(1, 56)=13.543$, $p<0.001$,

[1] LSA scores are the cosine of the angle between the two points being compared. This means that higher scores represent short distances (similarity), while lower scores represent long distances (dissimilarity).

MSe=0.020]. There was no reliable difference in the LSA scores for the different inference types of the pairs [Causal M=0.34, Attributal M=0.29; $F(1, 56)=2.067$, $p>0.15$, *MSe*=0.020] and no reliable interaction between inference type and distributional distance [$F<1$].

Frequency was controlled as in Experiment 1. Causal pairs had a mean frequency of 1327 for distributionally close pairs and 2517 for distant pairs, while attributal pairs had a mean frequency of 2909 for close pairs and 16285 for distant pairs. There was no difference in word frequency between the distributional distances [$F(1, 56)=2.508$, $p>0.1$, *MSe*=317171038], no difference between inference types [$F(1, 56)=2.786$, $p>0.1$, *MSe*=317171038] and no significant interaction between inference type and distributional distance [$F(1, 56)=1.756$, $p>0.15$, *MSe*=317171038].

Additionally, a pretest examined whether the basic meaning of the second sentence was maintained in the close and distant variants. A group of eighteen participants not used in any other experiment were asked to rate the appropriateness of the terms in the second sentence (e.g., the appropriateness of using *growled* or *snarled* in a sentence with *hounds*). On a scale from 1 (not appropriate) to 7 (very appropriate), this pretest showed little difference between the distributionally close and distant versions for noun/verb appropriateness in the causal pairs (Close M=5.8, Distant M=6.0) or noun/adjective appropriateness in the attributal pairs (Close M=6.0, Distant M=5.8). A two-way analysis of variance of the appropriateness ratings confirmed that there was no effect of either distributional distance [Fs<1] or inference type [Fs<1]. The interaction of the factors was also not significant [F_1 (1, 16)=2.38, $p>0.1$, *MSe*=1.821; F_2 (1, 29)=1.74, $p>0.15$, *MSe*=2.505].

Design

The design treated inference type as a between-participants and between-items variable and distributional distance as within-participants and within-items. The materials, 60 sentence pairs in all, were spilt by inference type into 30 causal and 30 attributal sentence pairs (i.e., close and distant versions of 15 basic sentence pairs). Two groups were formed per inference type, each of which contained fifteen sentence pairs counterbalanced between distributionally close and distant variants. All 15 sentence pairs within each group were presented in a random order, resampled for each participant.

Participants

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

Procedure

Participants read instructions that explained the 0-10 plausibility scale (0 being implausible and 10 being very plausible) with examples of the type of sentence pairs ahead (using a pair not featured in the experiment, appropriate to the inference type of the group). Those in the causal group saw the causal pair "*The waitress dropped the cup. The coffee spilled.*" while those in the attributal group saw the attributal pair "*The waitress dropped the cup. The coffee was hot.*" Participants

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 participants to note their 0-10 plausibility rating.

3.3.2 – Results and Discussion

The results showed no reliable main effects, of either inference type or distributional distance (see Figure 3.2 and Table 3.4). For attributal sentence pairs, close items were rated slightly more plausible than distant items (following the results of Lapata et al., 1999). However, for causal sentence pairs the opposite direction was found, with distant items rated slightly more plausible than close items. This resulted in a significant interaction between the factors.

A two-factor mixed design analysis of variance showed no significant difference between the plausibility ratings for causal and attributal items [Fs<1]. This was not unexpected, as inference type was a between-participants factor, meaning that causal and attributal items were rated independently from each other rather than being compared. In addition, the main effect of distributional distance was also not reliable [Fs<1]. Despite the lack of main effects, there was a reliable interaction between inference type and distributional distance [F_1 (1, 21)=4.36, p<0.05, MSe=10.055; F_2 (1, 28)=6.63, p<0.05, MSe=6.702]. However, planned pairwise comparisons for each inference type show that the distributional effect failed to achieve significance for either causal sentence pairs [F_1 (1, 11)=2.85, p>0.1, MSe=8.957; F_2 (1, 14)=3.14, p=0.098, MSe=7.244] or attributal sentence pairs [F_1 (1, 10)=1.66, p>0.2, MSe=11.263; F_2 (1, 14)=3.53, p=0.081, MSe=6.151].

Table 3.4 – Mean plausibility ratings for each condition in Experiment 2.

Distributional Distance	*Inference Type*		
	Causal	*Attributal*	*Overall Mean*
Close	6.23	7.02	*6.61*
Distant	6.94	6.30	*6.64*
Overall Mean	*6.59*	*6.66*	

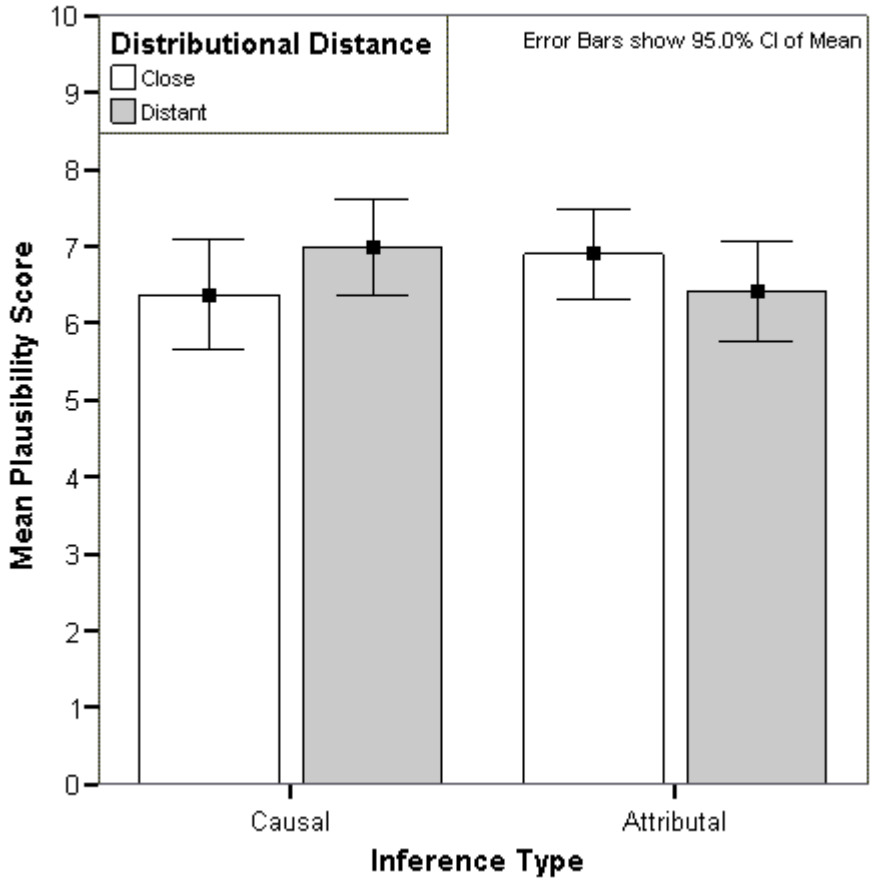


Figure 3.2 – Mean plausibility ratings per inference type and distributional distance in Experiment 2.

It could be argued that the failure to find an effect of distributional distance was due a poor rendering of the difference between the distributionally close and distant pairs. Some close-distant materials were further apart than others in terms of their distributional distance (i.e., their LSA scores). However, recall that our pretests of the materials showed that the two conditions were reliably different in distributional distance. In order to test this issue more stringently, we grouped the sentence pairs according to how extreme their LSA differences were and re-ran 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 close and distant forms), and examined participants' ratings for those sentence pairs only. Table 3.5 shows the mean LSA scores for these groups and their respective plausibility ratings in each condition. The results confirmed our earlier findings that there was no significant difference between distributionally close and distant sentence pairs, with distributional distance failing to achieve significance even in the group with the most extreme LSA score differences [all Fs<1].

In summary, this experiment shows that people were not sensitive to manipulations of word-coherence when they rated the plausibility of scenarios. Even the largest manipulations of word-coherence, where the distributional distance between the sentences was at its greatest, failed to produce any robust effect on plausibility ratings. The effects found by Lapata et al. (1999) for adjective-noun pairs were not found to generalise to sentence pairs. However, the individualistic nature of distributional knowledge means that inter-participant variation could have masked the effects of distributional distance. In the next experiment, we investigate this possibility.

Table 3.5 – Mean plausibility ratings for each condition in Experiment 2, subclassified by the extent of the difference in distributional scores.

LSA Score Difference	*Distributional Distance*	*Inference Type*		
		Causal	*Attributal*	*Overall Mean*
Least Extreme	Close	7.17	6.62	*6.88*
(M=0.09, N=26)†	Distant	7.97	6.34	*7.14*
Mid Extreme	Close	5.92	7.86	*6.85*
(M=0.13, N=16) †	Distant	6.42	7.14	*6.76*
Most Extreme	Close	5.37	6.90	*6.00*
(M=0.22, N=18) †	Distant	6.13	5.44	*5.83*

† M represents the mean difference of strong minus weak LSA scores for the N sentence pairs in that category. Different Ns per category result from tied LSA score differences.

3.4 – Experiment 3: Plausibility Ratings and Distributional Distance

In Experiment 2, we found that manipulating word-coherence had no effect on the plausibility ratings that people gave to scenario descriptions. In that experimental design, each participant saw either the distributionally close or distant form of each sentence pair. However, the Knowledge-Fitting Theory of Plausibility holds that distributional knowledge results from a cumulative lifetime exposure to language and as such is a highly individualistic phenomenon with much inter-participant variation. This means that the effect of distributional distance could have been masked in Experiment 2 by individual differences in distributional knowledge. In other words, a large distributional distance to one person may be a short distributional distance to another; the only consistency from person to person is the

relative difference between distributionally close and distant sentence pairs. In this experiment, therefore, we replicate Experiment 2 but present each participant with both the distributionally close and distant forms of each sentence pair together on the same page.

Our predictions are identical to those of Experiment 2. As in Experiment 1, we predict that the causal pairs will be rated as more plausible than the attributal pairs because of their stronger inferential connection. However, as in Experiment 2, this experiment uses a different design that treats inference type as a between-participant variable, which would also lead us to anticipate that the difference between causal and attributal pairs may be smaller than in Experiment 1. Regarding the effect of distributional distance, previous research (Lapata et al., 1999) would lead us to predict that people will rate the distributionally close pairs as more plausible than the distributionally distant pairs. However, as we noted in Experiment 2, several studies outside the area of plausibility have shown that when target tasks involve additional inferences the predictive power of distributional measures is reduced (French & Labiouse, 2002; Glenberg & Robertson, 2000; Lynott & Ramscar, 2001). Since our sentence pairs involve causal, temporal, and other inferences, these studies would also lead us to anticipate that the difference between distributionally close and distant pairs may be smaller than that observed for adjective-noun combinations.

3.4.1 – Method

Materials

The materials were those used in Experiment 2.

Design

As with Experiment 2, the design treated inference type as a between-participants and between-items variable and distributional distance as within-participants and within-items. The materials, 60 sentence pairs in all, were spilt by inference type into 30 causal and 30 attributal sentence pairs (i.e., close and distant versions of 15 basic sentence pairs). The matched close and distant 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 matched sentence pairs, one group received a close/distant order of presentation, while the other received a distant/close 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.

Participantsy

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

complete the experiment.

Procedure

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 ahead (using pairs not featured in the experiment, appropriate to the inference type of the group). Those in the causal group saw the distributionally close pair "*The chef poured the stew. The gravy dripped*" followed by the distributionally distant pair "*The chef poured the stew. The gravy dribbled*". The attributal group saw the close pair "*The chef poured the stew. The gravy was delicious*" followed by the distant pair "*The chef poured the stew. The gravy was tasty*". Participants were asked, if they found one sentence pair more plausible than the other, to make certain that their ratings reflected this fact. One distributionally close and one distant sentence pair (from the same matched set) were presented per page, each with the scale for participants to circle their plausibility rating. The position of the pairs on the page relative to one another (i.e., close or distant above or below) was counterbalanced in the experiment.

3.4.2 – Results and Discussion

The results replicated the causal-attributal effect found in Experiment 1, confirming the role of concept-coherence, but again showed no reliable effect of word-coherence (see Figure 3.3 and Table 3.6).

As found previously, there was a main effect of inference type with the causal sentence pairs yielding higher plausibility ratings than attributal pairs in the two-factor mixed design analysis of variance, though the by-participants analysis is outside significance [F_1 (1, 21)=1.17, p>0.2, MSe=35.978; F_2 (1, 672)=8.08, p<0.005, MSe=5.217]. 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<0.005] and by items [L(15)=71, p<0.00001].

In contrast, and in corroboration of Experiment 2's findings, the main effect of distributional distance was not reliable across analyses [F_1 (1, 21)=7.26, p<0.05, MSe=1.516; F_2 (1, 28)=0.86, p>0.3, MSe=12.789]. Planned comparisons for each inference type showed that the distributional effect was not significant for causal sentence pairs [F_1 (1, 10)=1.79, p>0.2, MSe=1.840; F_2 (1, 14)=0.65, p>0.6, MSe=15.664] and not significant by items for attributal sentence pairs [F_1 (1, 11)=6.88, p<0.05, MSe=1.221; F_2 (1, 14)=0.84, p>0.3, MSe=9.944]. Indeed, the direction of the difference in the mean was opposite to that predicted from previous work (i.e., distant was rated more plausible than close; see Lapata et al, 1999). No reliable interaction between inference type and distributional distance was found [Fs<1].

Table 3.6 – Mean plausibility ratings for each condition in Experiment 3.

Distributional Distance	*Inference Type*		
	Causal	*Attributal*	*Overall Mean*
Close	7.37	6.82	*7.08*
Distant	7.57	7.13	*7.34*
Overall Mean	*7.47*	*6.96*	

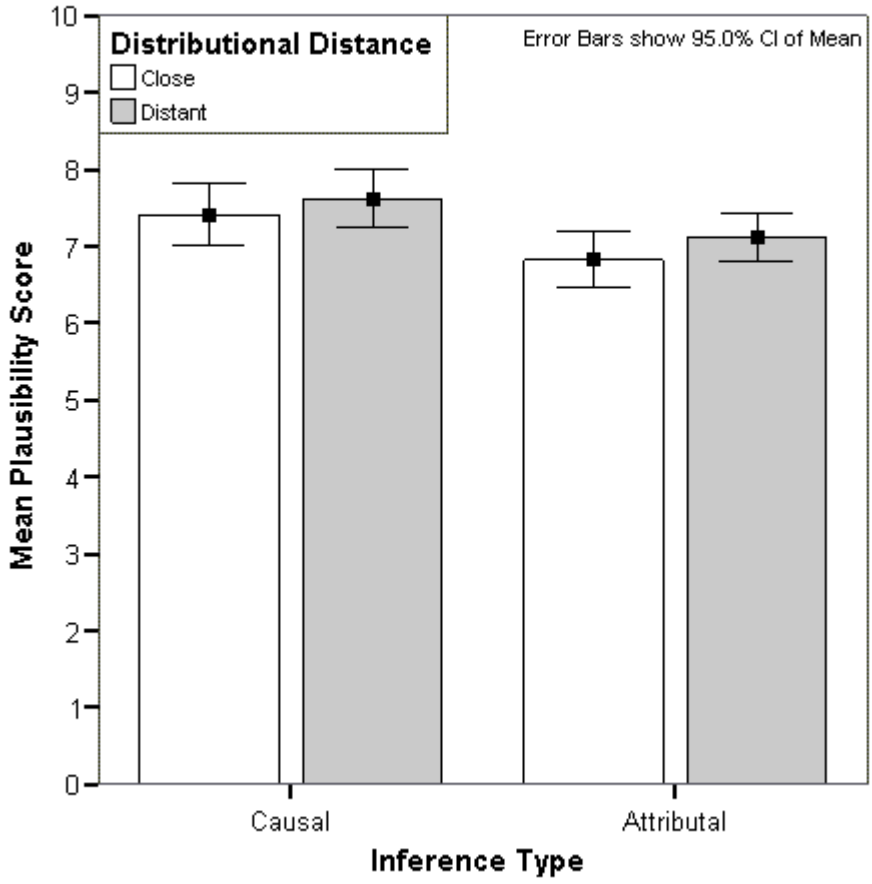


Figure 3.3 – Mean plausibility ratings per inference type and distributional distance in Experiment 3.

Although distributionally distant items were not reliably rated as more plausible than distributionally close items, it could again be argued that the failure to find a reliable effect was due a poor rendering of the difference between the distributionally close and distant pairs. Although a pretest showed that the two conditions were significantly different in distributionally distance (i.e., their LSA scores), some close-distant materials were further apart than others. However, regression analysis showed that the size of this difference between the close-distant variants had little effect on the differences in plausibility ratings that were provided by participants (adjusted $R^2 = -0.003$, $p > 0.7$).

We also carried out a more stringent test of this issue. As in Experiment 2, we grouped the sentence pairs according to how extreme their LSA differences were and re-ran 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 close and distant forms), and examined participants' ratings for those sentence pairs only. Table 3.7 shows the mean LSA scores for these groups and their respective plausibility ratings in each condition. The results confirmed our earlier findings that there was no significant difference between distributionally close and distant sentence pairs, with distributional distance failing to achieve significance even in the group with the most extreme LSA score differences [Least Extreme Group: F_1 (1, 21)=2.53, p>0.1, MSE=2.507; F_2 (1, 11)=1.17, p>0.3, MSE=5.398. Mid Extreme Group: Fs<1. Most Extreme Group: F_1 (1, 21)=4.03, p>0.05, MSE=3.412; F_2 (1, 7)=0.88, p>0.7, MSE=15.546]. None of these analyses showed any obvious grouping of the material set along the close-distant dimension that generated robust differences in plausibility ratings.

Table 3.7 – Mean plausibility ratings for each condition in Experiment 3, subclassified by the extent of the difference in distributional scores.

LSA Score Difference	*Distributional Distance*	*Inference Type*		
		Causal	*Attributal*	*Overall Mean*
Least Extreme	Close	7.31	7.19	*7.25*
(M=0.09, N=26)†	Distant	7.84	7.01	*7.44*
Mid Extreme	Close	7.44	6.58	*6.94*
(M=0.13, N=16) †	Distant	7.06	7.07	*7.06*
Most Extreme	Close	7.38	6.56	*7.00*
(M=0.22, N=18) †	Distant	7.49	7.38	*7.44*

† M represents the mean difference of strong minus weak LSA scores for the N sentence pairs in that category. Different Ns per category result from tied LSA score differences

In conclusion, this experiment replicates the concept-coherence effect found in Experiment 1 but shows no reliable effect of word-coherence. Even when presenting distributionally close and distant items together to maximise any potential word-coherence effects, people do not rate their plausibility differently. The results of further analyses verified that even the largest word-coherence 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. for adjective-noun pairs do not generalise to more complex descriptions.

3.5 – Experimental Conclusions

There are two novel findings in this chapter's empirical work. First, we have established not only that concept-coherence plays a role in plausibility, but 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 plausibility. Second, we have shown that word-coherence does not appear to play a role in rating plausibility: the distributional distance between sentences does not affect how plausible people find the scenario described.

Regarding previous research from the *concept-coherence view* of plausibility, the results presented here offer some interesting extensions of earlier findings. Black, Freeman and Johnson-Laird (1986) have shown that varying the number of possible inferences that could be drawn between sentences in a story has an effect on plausibility judgements. We have demonstrated that it is not only the presence but also the *type* of inference that is important to plausibility. Causal inferences are found to be the most plausible because the background knowledge they draw in provides 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 emphasises that plausibility is not just affected by the presence of inferences, but also by the type of inference in question. In other words, plausibility is influenced by the actual knowledge drawn in as each of the inferences is made.

With regard to the *word-coherence view*, the present results suggest Lapata et al.'s (1999) findings are limited to adjective-noun combinations. Lapata et al. gave their participants simple adjective-noun pairs, and found that distributionally close items were judged more plausible than distributionally distant items. It has been proposed that local word-level context is only useful when people are given no global concept-level context (Hess, Foss & Carroll, 1995), which would suggest that Lapata et al.'s participants based their plausibility ratings on the only useful information they had – distributional information. This represents a simple situation where distributional distance 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 conceptual representation on which to base a plausibility assessment. In this case, concept-coherence is the basis of the plausibility rating. Plausibility ratings of scenarios show no reliable effect of distributional distance because word-coherence is rendered somewhat irrelevant by concept-coherence. However, it should be remembered that the experiments in this chapter investigate the end product of plausibility judgement, which does not preclude an influence of word-coherence on the time course of making the judgement itself. It is to this issue that we turn in the next chapter.

CHAPTER 4 – PLAUSIBILITY JUDGEMENT TIMES

4.1 – Outline of Experiments

As we have earlier described, there are many different ways in which we can judge plausibility. Some of these judgements can carefully determine exactly *how* plausible we find a scenario, and so allow us to examine what makes one scenario more plausible than another. Other judgements can quickly determine *whether* a scenario is plausible or implausible, and so allow us to examine what makes one scenario faster to judge than another. In this chapter, we investigate the latter type of plausibility judgement using an online paradigm. We have seen in the previous chapter that the product of the plausibility judgement process (i.e., plausibility ratings) is influenced by concept-coherence, and not by word-coherence. However, this does not obviate a role for word-coherence during the judgement process itself; that is, word-coherence may contribute to the early stages of plausibility judgement but the effect is not discernable once concept-coherence comes into play. The online paradigm used in this chapter allows us to examine how concept-coherence and word-coherence may exert an influence on how quickly the plausibility of a scenario can be judged.

Chapter 1 gave a general introduction to the central ideas and components in the Knowledge-Fitting Theory of Plausibility. To recap, concept-coherence is concerned with how background knowledge is used to make the scenario fit with

what we know of the world. In general, the more background knowledge that is used to connect the events in a scenario, the slower the scenario will be to understand and to judge. The following experiments manipulate concept-coherence by using different inferential connections between events, which allows us to examine if people process some inference types more quickly than others. Word-coherence, on the other hand, is concerned with the distance between the distributional spotlights created by the sentences in the scenario description. As described in Chapter 1, if the spotlights fall far apart in distributional space, then more background knowledge will be activated in long-term memory. Generally speaking, the more background knowledge that has been primed, the faster the scenario will be to understand. In the experiments that follow, word-coherence is manipulated by describing the same scenario in different ways that vary distributional distance, which allows us to examine if people process distributionally distant sentences more quickly than the distributionally close.

Our experiments measure the time it takes people to process two-sentence descriptions of various events (e.g., "*The pack saw the fox. The hounds growled.*" See also Connell & Keane, 2002b, in prep.). Two experiments are reported that use this online paradigm. In Experiment 4, we cross the factors of inference type and distributional distance and measure comprehension times (i.e., how long it takes to read and understand the described scenario). Experiment 5 is identical except for the task performed, where we instead measure plausibility judgement times (i.e., how long it takes to judge if the described scenario is plausible or not). The time taken to actually assess plausibility (after the sentence is read) can then be calculated by subtracting the times for Experiment 4 from those of Experiment 5.

4.2 – Experiment 4: Comprehension Times

This experiment is focussed on the first stage of plausibility judgement, *comprehension*. In this stage, a scenario description must be read and understood, and a mental representation of the scenario created. We expect an effect of both distributional distance and inference type on comprehension times, and we now describe how we arrived at these predictions.

First, let us consider the distributional spotlight phenomenon described in Chapter 1, where larger spotlight coverage means that more background knowledge will be activated in long-term memory. If the sentences lie close together in distributional space, then their spotlights will overlap and coverage will be small. However, if the sentences are far apart, then each spotlight falls in different places and coverage will be large. For example, sentences in the *growled* pair given in Table 4.1 are more distributionally distant than the *snarled* pair, even though conceptually, both pairs involve the same inference and require the same background knowledge to connect the events (e.g., that the *hounds* are hunting, and hunting dogs *growl*/*snarl* at their prey, and the *fox* is their prey). This distributional distance gives the *growled* sentence pair an advantage because it primes more background knowledge, and thus is more likely to have primed the knowledge necessary to connect the sentences. We predict that this advantage will be reflected in faster comprehension times for distributionally distant sentence pairs.

Second, let us consider the amount of background knowledge needed to make the inferential connection between the sentences. For example, connecting the sentences in the attributal *vicious*/*fierce* sentence pair in Table 4.1 is simply a matter

of adding the attribute *vicious*/*fierce* to the *pack* mentioned in the first sentence. In contrast, connecting the sentences in the causal *growled*/*snarled* sentence pair involves more complex background knowledge about the behaviour of hunting dogs. This would lead us to predict that attributal inferences will be made more quickly than causal inferences, simply because causal inferences are more complex, and involve integrating more background knowledge than do attributal inferences.

Table 4.1 – Sample of sentence pairs used in Experiments 4 and 5.

Sentence 1	*Sentence 2*	*Inference Type*	*Distributional Distance*	*LSA Score*
The pack saw the fox.	The hounds growled.	Causal	Close	0.37
	The hounds snarled.	Causal	Distant	0.20
	The hounds were fierce.	Attributal	Close	0.19
	The hounds were vicious.	Attributal	Distant	0.12

4.2.1 – Method

Materials

The test items were the same as those used in the final ratings experiments in Chapter 3, and are described in Experiment 2. The materials consisted of sixty sentence pairs with crossed manipulations of both distributional distance and inference type (see Table 4.1, or Appendix B for a full set of materials).

The syllable length of the sentences was not controlled in the original paper-based experiments of Chapter 3. An analysis of variance of the syllable lengths of the second sentence in each pair showed a difference between inference types

[Causal M=3.7, Attributal M=5.2; $F(1, 56)=33.82$, $p<0.0001$, $MSe=1.043$]. However, since reading time increases with syllable length, the direction of this difference does not contribute to the predicted effect of inference type (i.e., we predict that attributal items will be comprehended more quickly than causal items despite their differences in syllable length). There was no difference in syllable length between the distributionally close and distant pairs [Close M=4.3, Distant M=4.3; $F<1$], and no interaction of inference type and distributional distance [$F(1, 56)=1.02$, $p>0.3$, $MSe=1.043$].

Sixty-eight filler items of the same form as the test items (i.e., with synonyms used to create two versions, but this time with no difference in LSA scores) were also created, thirty-four of each inference type. Of these, half described plausible scenarios and half described implausible scenarios. Four plausible and four implausible fillers were randomly selected for use in practise sessions. This gave the items in the test phase of the experiment a plausible:implausible ratio of 2:1.

Design

The design was the same as Experiment 3, and treated inference type (causal, attributal) as a between-participant and between-item variable, and distributional distance (close, distant) as a within-participant and within-item variable. Each participant saw both the close and distant forms of a sentence pair; this was done to ensure that individual differences in distributional knowledge did not mask the effect of the distributional manipulation, as found in Experiment 2 in the previous chapter. The order of presentation of the close and distant forms was randomised. Each participant was randomly assigned to one of the inference groups.

Participants

Twenty students of University College Dublin participated in this experiment. All were native English speakers, and received a nominal fee for participation.

Apparatus

The experiment was run on a laptop computer that recorded responses and response times. The sentences were displayed onscreen in black text on a white background, in standard upper- and lowercase typeface. Participants were seated in front of the screen with their dominant hand resting on the keyboard, and responded by pressing certain keys as specified in the instructions.

Procedure

Participants were instructed to read the first sentence of a pair when it appeared onscreen, and indicate that they understood it by pressing the spacebar. The second sentence would then appear underneath, and participants were asked to press the spacebar when they had understood it. Each sentence pair was followed by a short pause that consisted of "Wait…" being shown onscreen for two seconds before the first sentence of the next pair was displayed.

A short practice session followed the instructions, where participants received feedback if their responses were too slow (>5 seconds). The main experiment then commenced, where test and filler items were presented in a different random order for each participant.

4.2.2 – Results

Response times measured were those for the second sentence of each test pair. For this and the subsequent experiment, we regarded as an error any response time that was 3 standard deviations above or below the mean for a participant or item within a condition. Removing these responses resulted in a loss of 0.2% of this experiment's data.

Mean comprehension times per condition are presented in Table 4.2. A two-way mixed design analysis of variance was performed by participants (F_1) and by items (F_2), by treating participants and items as random factors, respectively.

A main effect of distributional distance on comprehension times was found, with distributionally distant items read 116ms faster than close items [F_1 (1, 18)=5.39, p<0.05, MSe=387215; F_2 (1, 28)=6.53, p<0.05, MSe=305301]. Planned comparisons showed that this distributional effect was slightly stronger for causal items (distant 123ms faster than close) [t_1 (19)=1.98, p=0.06; t_2 (29)=2.19, p=0.04] than for attributal items (distant 109ms faster than close) [t_1 (19)=1.90, p=0.07; t_2 (29)=1.98, p=0.06]. In addition, a main effect of inference type on comprehension times was significant by items – attributal items were read 165ms faster than causal items [F_1 <1; F_2 (1, 28)=8.17, p<0.01, MSe=503044]. There was no significant interaction by either participants or items (Fs<1).

As an additional test, we also performed a multiple regression analysis to compare the relative contribution of each factor, while including participants and items as predictors in the model. Both inference type [standardised coefficient $beta$=0.190, t(594)=2.378, p<0.05] and distributional distance [standardised

coefficient beta=0.087, $t(594)=2.182$, $p<0.05$] were shown to be significant predictors of comprehension times, with inference type the stronger predictor of the two.

Table 4.2 – Mean comprehension times (in milliseconds) per distributional distance and inference type in Experiment 4.

Distributional Distance	*Inference Type*		
	Causal	*Attributal*	*Total*
Close	1529	1356	*1442*
Distant	1405	1247	*1326*
Total	*1467*	*1302*	

4.2.3 – Discussion

Results were in line with our predictions. Distributional distance affects comprehension in the predicted direction, even when other factors (such as word frequency and noun/verb or noun/adjective appropriateness) have been controlled. Inference type also affects comprehension, with regression analysis showing it is an even stronger predictor of comprehension times than distributional distance.

When people read a sentence, they find it easier to understand if it is distant from the previous sentence in distributional space. While this finding may seem counterintuitive at first, as it is in the opposite direction to that found for the priming of single words (Landauer & Dumais, 1997; Lund, Burgess & Atchley, 1995), it is consistent with the predictions of the Knowledge-Fitting Theory. Simple lexical priming effects are seen in cognitively simple tasks that have no concept-coherence component (see Hess, Foss & Carroll, 1995). Presented words cause spotlights to

fall on areas of distributional space, and these spotlit areas have increased activation. In this way, priming effects result directly from spotlit distributional knowledge; distributionally similar words (high LSA scores) are close to each other in distributional space and so will prime each other.

However, full sentence comprehension is a more complex task that utilises concept-coherence, and as such, does not depend directly on distributional knowledge. Instead, it depends on what prior knowledge has been primed by the distributional spotlight. When creating the close and distant variants of a sentence pair, we were careful to maintain the same basic meaning so that both variants would invite the same inference. For example, "*The pack saw the fox. The hounds snarled*" invites the same causal inference as "*The pack saw the fox. The hounds growled*". Sentences that are close together in distributional space (the *snarled* pair) have largely overlapping spotlights, and hence prime only a limited amount of prior knowledge. Distant sentences (the *growled* pair) do not have (much) spotlight overlap, and hence prime more prior knowledge. As we increase the distributional distance between sentences, we increase the amount of background knowledge that is made available, and increase the chance that inference-making will be facilitated by primed knowledge. Thus, people can make the inference between distant sentences like the *growled* pair more quickly than between close sentence like the *snarled* pair. In other words, the distributional effects in Experiments 4 and 5 result directly from primed prior knowledge; distributionally similar sentences (high LSA scores) prime less background knowledge and so the inference will be slowed by having to retrieve un-primed knowledge from long-term memory.

In addition, attributal pairs were comprehended more quickly than causal pairs. The difference in comprehension time is quite intuitive, as there is obviously more background knowledge to be retrieved and integrated in representing a causal connection than an attributal connection. The slower comprehension times for causal pairs result from the extra complexity in building their representations.

Interestingly, causal sentence pairs exhibited a slightly stronger distributional distance effect than attributal sentence pairs. This was an unexpected finding, but one that is consistent with our predictions. We suggest that causal inferences benefit more from distributional distance because they use more of the prior knowledge that was primed by the distributional spotlight. Causal inferences (e.g., the hounds snarling/growling *was caused by* the pack seeing their prey) require more background knowledge to be retrieved and integrated than attributal inferences (e.g., the hounds being fierce/vicious *adds an attribute to* the pack). This means that causal inferences have more to gain if the knowledge they need has been primed. In other words, the greater spotlight coverage of distant sentences is of greater benefit to complex causal inferences than to simple attributal inferences.

4.3 – Experiment 5: Plausibility Judgement Times

As Experiment 4 examined the comprehension stage of the plausibility judgement process, the aim of this experiment is to shift the emphasis to plausibility *per se*, and measure the time taken to judge if a scenario is or is not plausible. The Knowledge-Fitting Theory assumes that plausibility judgement times are the sum of the times taken to comprehend and then assess the scenario, and so the differences

we predicted for comprehension should carry through. Specifically, first, plausibility judgement times should show an effect of distributional distance, namely that distant sentence pairs should be judged more quickly than close pairs. Second, plausibility judgement times should also show an effect of inference type, namely that attributal sentence pairs should be judged more quickly than causal pairs. Third, the plausibility judgement times should be slower than comprehension in all conditions, reflecting the additional time taken for assessment.

We also planned a meta-analysis based on subtracting comprehension times (Experiment 4) from plausibility judgement times (this experiment) to determine the relative contribution of the assessment stage. Our fourth prediction is that these net assessment times will show a distinct effect of inference type, namely that causal items will take longer to assess than attributal items. In assessment, causal inferences (e.g., the hounds growling *was caused by* the pack seeing their prey) have a more complex representation than attributal inferences because more background knowledge has been drawn in (e.g., the hounds being fierce *adds an attribute to* the pack), and this extra complexity takes longer to assess. Our final prediction is that there will be no effect of distributional distance for these net assessment times. Since the distributional spotlight affects the ease of understanding the scenario during comprehension, there is no eventual difference between the representations of close and distant items, and we therefore expect no effect of distributional distance when it comes to assessing the scenario.

4.3.1 – Method

Materials and Design

We used the same materials, both test and filler items, as in Experiment 4. Items in the test phase of the experiment had a plausible:implausible ratio of 2:1

Design was also the same as in last experiment, with inference type (causal, attributal) treated as a between-participant variable, distributional distance (close, distant) as a within-participant variable. Each participant was randomly assigned to one of the inference groups.

Participants

A new set of twenty students of University College Dublin participated in this experiment. All were native English speakers, and received a nominal fee for participation.

Procedure

The experimental procedure was identical to that of Experiment 4, except that participants were asked to judge the plausibility of the scenario rather than indicate their comprehension. When the second sentence appeared onscreen, participants were instructed to press a key labelled "yes" if they thought the sentence pair was plausible, or a key labelled "no" if they thought it was not plausible. The key "b" was labelled "yes", and the key "n" was labelled "no". Feedback was given during

the practice session if a response was too slow (>5 seconds) or if it was incorrect. Test and filler items were presented in a different random order for each participant.

4.3.2 – Results

Response times measured were the judgement times for the second sentence of each test pair. Removing outlier responses as in Experiment 4 resulted in a loss of 0.6% of this experiment's data. Error rates for test items in all conditions were 0% (i.e., participants correctly judged the sentence pairs as plausible). Filler items, both plausible and implausible, also had 0% error rates.

Factors Affecting Plausibility Judgement

Mean plausibility judgement times in each condition are shown in Table 4.3, and analysis was carried out as in Experiment 4. A main effect of distributional distance on plausibility judgement times was found, with distributionally distant items judged to be plausible 204ms faster than close items [F_1 (1, 18)=8.24, p<0.05, MSe=792332; F_2 (1, 28)=6.20, p<0.005, MSe=985360]. As found in Experiment 4, planned comparisons showed that this distributional effect was slightly stronger for causal items (distant 227ms faster than close) [t_1 (19)=2.27, p=0.04; t_2 (29)=2.12, p=0.04] than for attributal items (distant 180ms faster than close) [t_1 (19)=2.26, p=0.04; t_2 (29)=1.97, p=0.06]. In addition, a main effect of inference type on plausibility judgement times was significant by items, with attributal items judged to be plausible 379ms faster than causal items [F_1 (1, 18)=1.44, p>0.2, MSe=14730911;

F_2 (1, 28)=8.28, p<0.01, *MSe*=2635409]. There was no significant interaction by either participants or items (Fs<1).

As an additional test, we also performed a multiple regression analysis to compare the relative contribution of each factor, while including participants and items as predictors in the model. Both inference type [standardised coefficient *beta*=0.440, t(591)=5.489, p<0.0001] and distributional distance [standardised coefficient *beta*=0.083, t(591)=2.084, p<0.05] were shown to be significant predictors of plausibility judgement times, with inference type the stronger predictor of the two.

Comparing Comprehension and Plausibility Judgement

We also compared the data for comprehension times (Experiment 4) and plausibility judgement times (Experiment 5), and found that plausibility judgement took on average 622ms longer than comprehension [F_1 (1, 36)=13.57, p<0.001, *MSe*=10125356; F_2 (1, 28)=121.81, p<0.0001, *MSe*=956847]. There was no significant interaction of task type with any other factor.

It should be noted that there is an implicit task difference between comprehension and plausibility judgement, namely that plausibility judgement involves a yes/no decision while comprehension does not. The procedure of the experiments thus differed in that Experiment 4 participants had to press a single key to indicate comprehension, while Experiment 5 participants had to press one of two keys to indicate plausibility judgement. This means that the extra 622ms that plausibility judgement takes after comprehension will include time required for the processing of this motor response. While it may be possible in future experiments to

eliminate this task difference, it is not of concern in the present analyses. Participants saw the same proportion of plausible:implausible scenarios in each condition (causal close, causal distant, attributal close, attributal distant), and so the extra time required for processing the yes/no response is constant across conditions and could not have contributed to the observed effects.

Meta-Analysis of Assessment Times

In order to isolate the extra time taken after comprehension to perform the plausibility judgement process, we created a set of net response times by subtracting the mean comprehension time for each item in each condition from the corresponding mean plausibility judgement time. These net assessment times therefore represent the time taken to perform the assessment stage, and are shown in Table 4.4. A two-way analysis of variance was performed for the factors of inference type and distributional distance. There was a significant effect of inference type on plausibility assessment times, with a 214ms difference between attributal and causal items [$F(1, 56)=4.26$, $p<0.05$, $MSe=165672$]. However, there was no effect of distributional distance on assessment times, as the 87ms difference between distributionally close and distant items was not significant ($F<1$). There was no interaction of inference type and distributional distance ($F<1$). As an additional test, we also performed a multiple regression analysis to compare the relative contribution of each factor, while including items as a predictor in the model. Inference type was shown to be a significant predictor of assessment times [standardised *beta*=0.264, $t(56)=2.063$, $p<0.05$], while distributional distance was not [standardised *beta*=0.106, $t(56)=0.827$, $p>0.4$].

Table 4.3 – Mean plausibility judgement times (in milliseconds) per distributional distance and inference type in Experiment 5.

Distributional Distance	*Inference Type*		
	Causal	*Attributal*	*Total*
Close	2310	1907	*2108*
Distant	2083	1728	*1904*
Total	*2196*	*1817*	

Table 4.4 – Mean net assessment times (in milliseconds), representing mean plausibility judgement times (Experiment 5) minus mean comprehension times (Experiment 4), per distributional distance and inference type.

Distributional Distance	*Inference Type*		
	Causal	*Attributal*	*Total*
Close	781	551	*666*
Distant	678	481	*580*
Total	*730*	*516*	

4.3.3 – Discussion

The results of this experiment were again in line with our predictions. The pattern of effects for comprehension times found in Experiment 4 was echoed in the effects found for plausibility judgement times in Experiment 5. Inference type and distributional distance were shown to exert a significant influence on plausibility judgement times. Furthermore, our meta-analysis showed that the assessment stage of plausibility judgement was influenced only by inference type. In short, concept-coherence and word-coherence affect the comprehension stage of plausibility judgement, but the assessment stage is purely influenced by concept-coherence.

Figure 4.1 nicely illustrates the comparative results of this chapter's experiments. The graph shows the standardised regression (*beta*) coefficients for each factor in each experiment. As reported in the regression analyses, inference type appears as a stronger predictor than distributional distance, but we can gain greater insight by comparing the experiments. Inference type can be seen to contribute more to plausibility judgement times than to comprehension times, while distributional distance contributes the same amount to both. In other words, the extra time that plausibility judgement takes after comprehension (i.e., assessment) is affected by the type of inference that links events, and not by the distributional properties of the description.

In the assessment stage of a plausibility judgement, people examine their representation of the scenario in question. As we have earlier described, complex representations take longer to examine. Since causal inferences (e.g., the hounds snarling/growling *was caused by* the pack seeing their prey) require more background knowledge to be retrieved and integrated than attributal inferences (e.g., the hounds being fierce/vicious *adds an attribute to* the pack), they lead to more complex representations. Thus, people take longer to assess scenarios with causal inferential connections than with attributal inferential connections. However, there will be little difference in the representation between the distributionally close and distant variants of a sentence pair, since both variants (e.g., *snarled*/*growled*) were designed to invite the same inference. Thus, the distributional distance of the original sentences has little bearing on how long it takes to examine the representation.

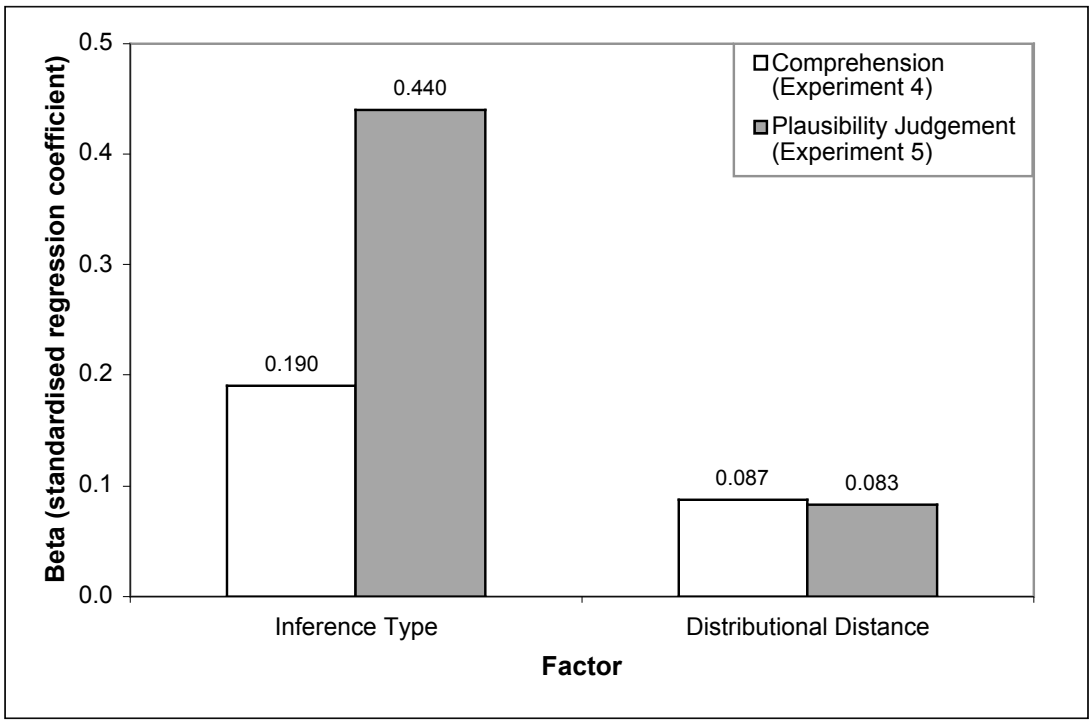


Figure 4.1 – Standardised regression (beta) coefficients for each factor of inference type and distributional distance, as calculated for comprehension times and plausibility judgement times; the higher the beta coefficient, the greater the predictive value of that factor.

4.4 – Experimental Conclusions

This chapter's experiments show a number of novel findings. First, word-coherence has been shown to affect comprehension times, with distributionally distant sentences being processed faster than distributionally close sentences. This phenomenon occurs even when controlling for other factors such as word frequency and noun/verb or noun/adjective appropriateness. Second, concept-coherence has been shown to affect both the processes of comprehension and plausibility assessment, with people taking longer to understand and judge the plausibility of scenarios with more complex causal inferential connections than less complex attributal inferential connections. Third, we have established a dichotomy between

the comprehension and assessment stages of plausibility judgement, as shown by the separability of the influences that act on each stage.

Regarding the *concept-coherence* view of plausibility, the present experiments offer some interesting extensions to our earlier findings. We have seen in Chapter 3 that the product of the plausibility judgement process (i.e., plausibility ratings) is influenced by the type of inference that connects events, and by the actual knowledge drawn in as an inference is made. The experiments in this chapter support this finding, and show that the time course of the plausibility judgement process (i.e., the length of time required to judge plausibility) is also influenced by the type of inference that connects events. Specifically, we found that the complexity of an inference affects both the time taken to comprehend the sentence (see also McKoon & Ratcliff, 1992; Singer, Graesser & Trabasso, 1994), and also the time taken to assess plausibility. In other words, both the comprehension and assessment stages of judging plausibility are influenced by the amount of knowledge drawn in as an inference is made.

With regard to the *word-coherence* view, our results show that the word-coherence of the description does indeed play a role in plausibility judgement. As we saw in Chapter 3, the product of the plausibility judgement process (i.e., plausibility ratings) is not influenced by word-coherence because distributional effects are rendered somewhat irrelevant by concept-coherence. However, this chapter's experiments show that word-coherence contributes to the plausibility judgement process itself (i.e., the length of time required to judge plausibility) during the comprehension stage. Specifically, distributionally distant sentences are understood more quickly than distributionally close sentences, because they prime

more relevant background knowledge in long-term memory. In short, distributional effects are confined to the building of the scenario representation during the comprehension stage, and do not affect how quickly the representation is examined in the assessment stage.

CHAPTER 5 – THEORY & MODEL

5.1 – Outline

The empirical work in Chapters 3 and 4 has given us a better understanding of how people make plausibility judgements. We examined the plausibility judgement process in two ways, firstly by asking people *how* plausible they find a particular scenario, and secondly by measuring the length of *time* it takes people to judge if a scenario is plausible or not. From these experiments, we concluded that the plausibility judgement process is composed of two stages, comprehension and assessment. We also identified two factors that influence plausibility judgements; the word-coherence of the scenario's description and the concept-coherence of the scenario's representation. Word-coherence effects were found to be confined to the comprehension stage, and only affected the time course of plausibility judgement. Concept-coherence effects were found in both stages, and affected both the time course of plausibility judgement and the level of plausibility a scenario was judged to have.

In this chapter, we present a well-specified theory of plausibility based on our empirical findings – the Knowledge-Fitting Theory of Plausibility (see also Connell & Keane, 2003a, 2003b, in prep.). This theory provides an account of how we represent and analyse scenarios to determine their plausibility, and how this process is influenced by word- and concept-coherence. In addition, we describe the

computational implementation of this theory, the Plausibility Analysis Model (PAM), and show how the model's performance parallels human plausibility responses.

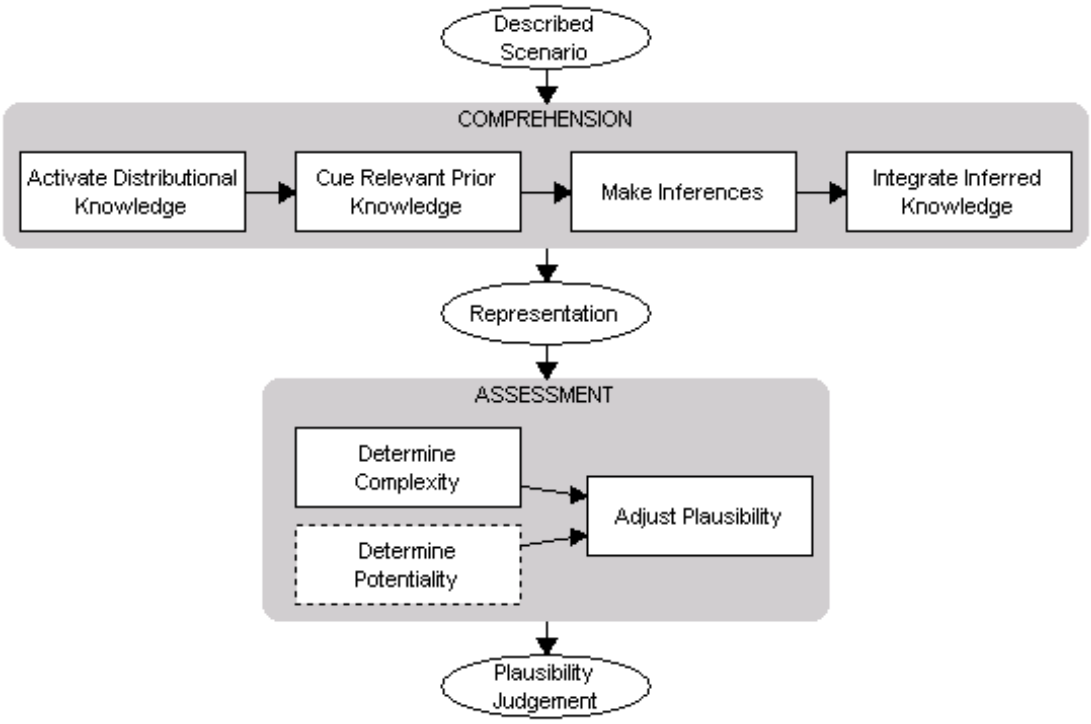


Figure 5.1 – The Knowledge-Fitting Theory of Plausibility, showing both the comprehension and assessment stages and their constituent processes.

5.2 – The Knowledge-Fitting Theory of Plausibility

The Knowledge-Fitting Theory of Plausibility asserts that, when people make a plausibility judgement, they fit what they have been told with what they know about the world. In other words, plausibility is a function of the degree of fit between some presented scenario and prior knowledge, where prior knowledge includes both distributional word knowledge and background conceptual knowledge.

In this theory, plausibility judgments involve two main processing stages: a comprehension stage and an assessment stage. During the *comprehension stage*, a mental representation of the presented scenario is created from the verbal description and inferences made from prior knowledge. During the *assessment stage,* this mental representation is evaluated to determine its fit to prior knowledge. The plausibility judgement is based on the results of this analysis in the assessment stage. Figure 5.1 shows the two stages of the theory and some of the sub-components that come into play when a plausibility judgement is being made.

5.2.1 – Making a Plausibility Judgement: Comprehension

We assume a fairly standard account of comprehension based on the consensus in the literature, with a few added twists (e.g., Gernsbacher, 1990, 1997; Singer, Graesser & Trabasso, 1994; Johnson-Laird, 1983; Kintsch, 1998; van Dijk & Kintsch, 1983; Zwaan, Kaup, Stanfield & Madden, 2001). This model characterises the comprehension of a scenario as the construction of a mental representation of the described situation, aided by the cues provided in the linguistic input. This mental representation integrates our knowledge of the world with the information in the scenario through the inferences that connect events. We assume this representation is built in working memory and that pieces of knowledge are retrieved from long-term memory to construct essential bridging and other inferences. As this representation is being built, a lot of knowledge is activated in long-term memory. Some of this activated knowledge supports the inferences made, whereas other activated knowledge may prove redundant, never being directly used for inferencing. Given the sheer amount of knowledge in human long-term memory, a major problem

facing any comprehension mechanism is the recruitment of a relevant subset of knowledge, in real time, from which to make appropriate inferences. One of the key comprehension assumptions of the Knowledge-Fitting Theory is that this rapid recruitment of knowledge for inferencing is influenced by the distributional properties of a scenario's words.

We assume three specific ideas in our comprehension account: that inferences are diverse and differ in complexity, that aspects of concepts are primed by distributional information, and that inferencing is influenced by distributional information. The first two of these proposals have been made by other researchers, though the last is more novel.

Inferences Are Diverse and Differ in Complexity

During the comprehension of discourse, we assume that many different types of inference are tracked and made. Most discourse cannot be comprehended solely from the presented text but rather contains conceptual gaps that must be filled by making suitable inferences. Imagine the following scenario:

1) *John poured water on the fire. The fire went out.*

To comprehend this scenario, the connection between the two events (i.e., that the water *caused* the fire to extinguish) must be inferred and then integrated into the representation of the scenario (see e.g., Halldorson & Singer, 2002; Keenan, Baillet & Brown, 1984; Kintsch & van Dijk, 1978; Singer & Halldorson, 1996). There is considerable diversity in the types of inferences that can be made. Apart from causal

information, people also monitor and integrate temporal, spatial and motivational aspects of the situation (see Zwaan & Radvansky, 1998, for a review).

Furthermore, we assume that some inferences are more complex than others, and hence, require more time to process. For example, the following two scenarios are very similar but invite inferences of differing complexity:

2a) *The bottle fell off the shelf. The glass smashed.* (causal)

2b) *The bottle fell off the shelf. The glass was pretty.* (attributal)

The former (2a) requires background knowledge to make the causal inference between the two events (e.g., the inference that the bottle smashed *because* it fell requires the knowledge that fragile things break when they hit hard surfaces). In contrast, the latter (2b) merely requires finding a co-referent for the attribute *pretty*, namely *bottle*. These differences in complexity have empirical consequences for ease of comprehension, and ultimately plausibility judgement.

We have seen that such inferential differences lead to different perceptions of plausibility; Chapter 3 has shown that, all things being equal, people will rate causal scenarios (like 2a) as being reliably more plausible than attributal scenarios (like 2b). In addition, we have seen that these inferential differences affect processing speeds: Chapter 4 has shown that less complex inferences take less time to comprehend, and less time to assess their plausibility.

Distributional Information Primes Aspects of Concepts

During comprehension, we also assume that different aspects of concepts are activated (in part) by the distributional properties of the words. The manifold effects

of different types of priming are well established, showing how words can activate other related concepts (e.g., Duffy, Henderson & Morris, 1989; Meyer & Schvaneveldt, 1976; Swinney, 1979). Indeed, words can act in combination to prime items that they would be unable to prime individually. For example, Duffy et al. (1989) presented sentences such as "*the barber trimmed the mustache*" one word at a time, and measured the time it took people to name the last word. Even though neither *barber* nor *trimmed* are strongly related to *mustache*, Duffy et al. found that the naming time for *mustache* was facilitated by the preceding context. Similarly, many studies have shown that different aspects of a concept can be highlighted by subtle changes in its discourse context (e.g., Anderson & Ortony, 1975; Barclay, Bransford, Franks, McCarrell & Nitsch, 1974; Halff, Ortony & Anderson, 1976; Keane, 1985; McKoon & Ratcliff, 1988). For example, McKoon and Ratcliff (1988) found that after being presented with the sentence "*The little girl found a tomato to roll across the floor with her nose*", people were faster to confirm that tomatoes were "round" than that they were "red".

Recently, it has been show that such priming effects can be predicted by the distributional properties of the words involved (Landauer & Dumais, 1997; Lund, Burgess & Atchley, 1995). For example, Lund et al. (1995) used their Hyperspace Analogue to Language (HAL) model to produce pairs of distributionally similar words (e.g., "ant" and "flea"), and showed that these words did indeed prime each other when given to people in a lexical decision task. In the above tomato-rolling scenario, the distributional explanation of the phenomenon would propose that the context activates words that are distributionally similar, resulting in higher activation levels for *round* over *red*. The sentence "*The little girl found a tomato to roll across*

the floor with her nose" activates a whole set of nearby words in the surrounding distributional space, and since the word "round" is distributionally closer than "red" to the context sentence, the activation of *round* increases relative to the activation of *red*. If the scenario described someone painting a still life of a ripe tomato, then the activation of *red* would increase relative to that of *round*. This activation can be pictured as a spotlight being directed at a region of words in distributional space; the context directs the spotlight to light up its distributionally similar neighbours, and the relative brightness of these neighbours fades out at the edge of the lighted area. In the rolling scenario, the word "round" is highly illuminated at the centre of the spotlight, whereas "red' is less illuminated towards the edge. In the painting scenario, the spotlight has shifted so that "red" is much brighter than "round".

Distributional Information Influences Inferencing

Finally, we assume that during comprehension the inferences made are influenced (in part) by what is activated by the distributional properties of the words. Other researchers have suggested that the distributional properties of the linguistic input may help to prime knowledge relevant to the present context (see Burgess, Livesay & Lund, 1998; Halldorson & Singer, 2002; Kintsch, 1998, 2000; Kintsch, Patel & Ericsson, 1999). Also, many studies have shown that some inferences can be made more rapidly than others (Halldorson & Singer, 2002; Keenan, Baillet & Brown, 1984; Myers, Shinjo & Duffy, 1987; Singer & Halldorson, 1996). For instance, take the following two scenarios:

3a) *The hiker shot the injured deer. The deer died.* (causal)

3b) *The hiker examined the injured deer. The deer died.* (temporal)

It has been shown that people are quicker to infer the connection between the events in 3a than the events in 3b (Halldorson & Singer, 2002). Even though the causal scenario is arguably more complex inferentially, it is understood faster than the temporal scenario because it is better able to set up the knowledge necessary to connect events. That is, the *shot-died* sentences in 3a prime relevant knowledge (e.g., about bullets killing animals), which is then used to make the causal inference between the events. This facilitation does not happen for the *examined-died* sentences in 3b. While part of this effect must arise from conceptual connections in background knowledge about these events, we also believe that distributional information plays a key role in facilitating the inferences being made. Distributional knowledge provides an important means of activating relevant knowledge, which eases the problem of making inferences in real time from a vast repository of prior knowledge.

The way in which distributional information supports inferencing can be appreciated as an extension of how it activates aspects of concepts. When people read a sentence, the words in the sentence activate an area of distributional space, which in turn primes related knowledge. The proximity and nature of the regions activated by different sentences will affect the knowledge primed. This, in turn, will affect the ease with which inferences are made, because an inference can be made more quickly if the knowledge it needs has already been primed. In a sentence pair, each sentence can be seen as producing separate spotlights, which may be close (i.e., overlap) or distant (i.e., not overlap) in the distributional space (see Figure 5.2). If there is a large overlap between the spotlight regions of the two sentences, then a very circumscribed portion of knowledge is primed, which is less likely to include

the knowledge that an inference will need. If there is little overlap between the spotlights of the two sentences, then a much more extensive portion of relevant knowledge is primed, which is more likely to include the knowledge required by an inference. For example, take the following sentence pairs:

4a) *The dress snagged on a nail. The satin ripped.* (close)

4b) *The dress snagged on a nail. The satin tore.* (distant)

These scenarios are practically identical – both describe the same event and both invite the same inference between the sentences. However, the sentences in the *ripped* pair (4a) are closer together in distributional space than the sentences in the *tore* pair (4b) (as shown in Chapter 3); that is, the spotlights for the *ripped* pair have a large overlap while the spotlights for the *tore* pair have a small overlap. This means that, although both pairs prime a certain portion of relevant knowledge, the distant *tore* pair will prime more knowledge than the close *ripped* pair. This gives the *tore* pair an advantage over the *ripped* pair, because it has a greater chance of priming the knowledge that the inferential process will need. In other words, people should be faster to make the inference for the distant *tore* pair because its distributional properties are more likely to prime useful knowledge.

In general, we assume that distributional distance supports inference making; the further apart the activated regions, the more knowledge will be primed, and the more likely it is that inferential connections will be facilitated by primed knowledge. In contrast, as the regions come closer together, less knowledge is primed and, hence, it becomes more likely that inferential connections will have to be made without help from primed knowledge.

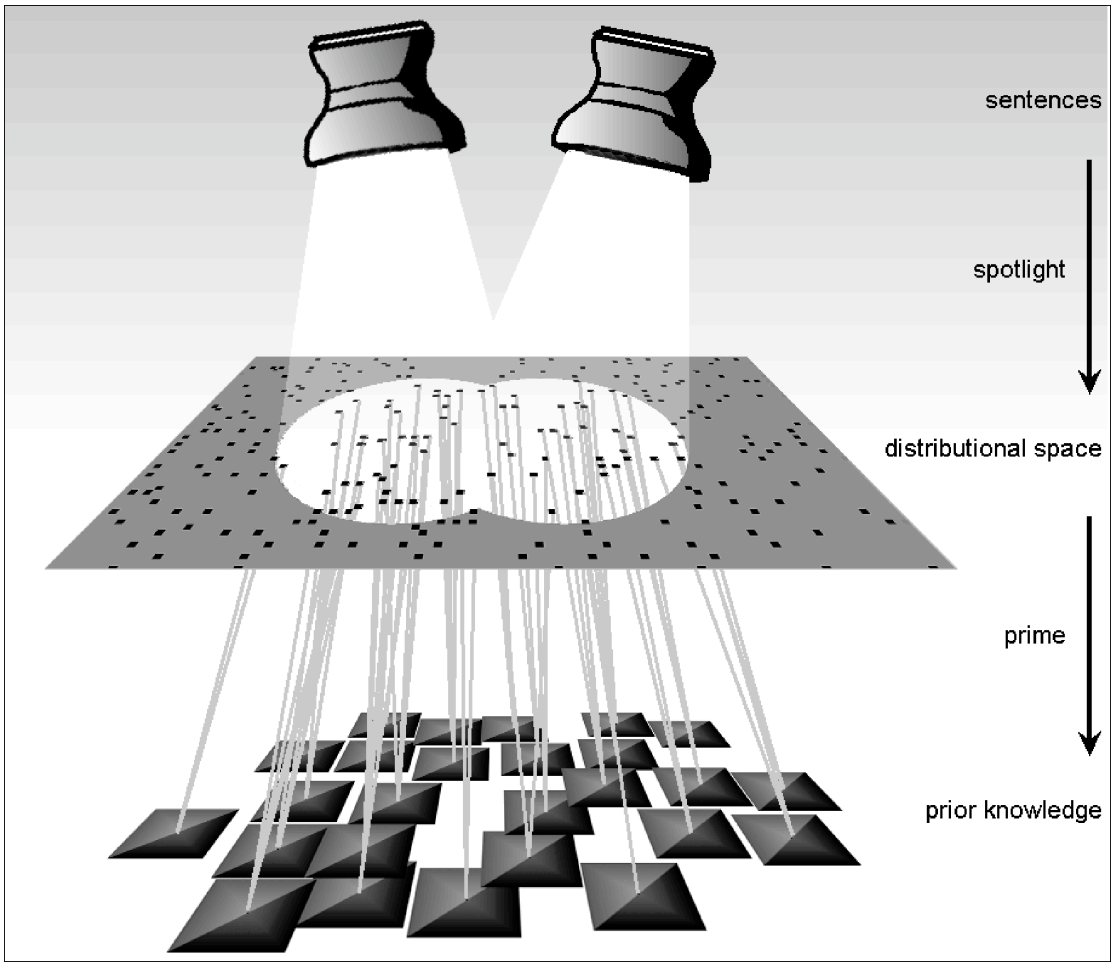


Figure 5.2 – Each sentence spotlights (activates) a surrounding area of distributional space, which in turn primes pieces of prior knowledge.

Summarising Comprehension

To summarise, we assume a core model of comprehension in which scenarios are represented by making inferences from prior knowledge. We add a number of additional assumptions to this model. In this extended model, the time taken for comprehension is affected by the complexity of the inferences being made and by the distributional properties of the words used in the description. Distributional information helps to prime relevant knowledge for inferencing by activating different regions of knowledge in long-term memory. Distant distributional regions are more likely to facilitate inference making, whereas close regions constrict what is primed

and are less likely to aid inference making. In short, comprehension is affected by both word-coherence and concept-coherence. As we shall see in the next section, the assessment stage differs in that it is purely concerned with concept-coherence.

5.2.2 – Making a Plausibility Judgement: Assessment

The second stage of the plausibility judgement process involves an assessment of the representation built during comprehension. The assessment stage is the core of the plausibility judgement *per se*. If the events in a scenario could not be connected in the representation (e.g., if we failed to come up with an explanation for how a bottle could melt after falling off a shelf), then the absence of an inferential connection makes the scenario implausible. If an inferential connection has been found, then the scenario's representation is analysed to determine how well it fits our knowledge of the world. Specifically, assessment measures how much extra information from background knowledge was necessary to connect the events described (the *complexity*), and how much of this knowledge had been primed by distributional information (the *potentiality*).

Measuring Complexity

As stated previously, we assume that some inferences are more complex than others. For example, the following two sentence pairs are very similar but are represented with inferences of differing complexity:

5a) *The bottle fell off the shelf. The glass sparkled.* (temporal)

5b) *The bottle fell off the shelf. The glass was pretty.* (attributal)

The sentences in 5a are represented by inferring a temporal connection between the two events; namely, that the bottle sparkled because it caught the light after hitting the floor. This connection incorporates the background knowledge that bottles can be shiny, and that shiny things sparkle when they catch the light. In contrast, the sentences in 5b are represented merely by inferring a co-referent for the attribute *pretty*, namely *bottle*. The greater amount of extra information that was needed to make the *sparkled* scenario fit prior knowledge means that its representation should take longer to assess than the *pretty* scenario. It also means that the *sparkled* scenario should seem less plausible than the *pretty* scenario, although potentiality must also be taken into account.

Measuring Potentiality

As we earlier described, we assume that primed knowledge can potentially support the inferential process. Some scenarios can be represented by using knowledge that distributional information has primed, while others can only be represented by retrieving different knowledge from long-term memory. For example, the following sentence pairs are very similar but use primed knowledge to different extents:

6a) *The bottle fell off the shelf. The glass smashed.* (causal)

6b) *The bottle fell off the shelf. The glass melted.* (unrelated)

The scenario in 6a has high potentiality, as it is likely to prime the background knowledge that fragile things smash when they fall on hard surfaces (see Halldorson & Singer, 2002; Singer & Halldorson, 1996). This knowledge is then used to infer

the connection that the bottle smashed because it was fragile and hit the floor. In contrast, the scenario in 6b has low potentiality, as it is not likely to prime the knowledge that high temperatures can melt a bottle, that metal can be heated to high temperatures, and that metal needs a heat source to get that hot. One can only infer the connection between the events (the bottle melted because the floor was extremely hot, because it was made of metal and something had heated it up) by retrieving the necessary knowledge from long-term memory. The greater amount of primed information that was used to make the *smashed* scenario fit prior knowledge means that it should seem more plausible than the *melted* scenario, although complexity must also be taken into account.

Summarising Assessment

To summarise, we assume that the representation formed during comprehension is analysed during assessment to determine its fit to prior knowledge. The more background knowledge that one must incorporate in building the representation (assuming that any inferences can be made in the first place), the longer it should take to assess and the less plausible it should seem. The more useful background knowledge that one can find in primed knowledge, the more plausible the scenario should seem. So, concept-coherence affects assessment as well as comprehension. In contrast, word-coherence does not affect assessment, and influences the plausibility judgement process only during the comprehension stage.

5.2.3 – Accounting for Plausibility Effects

The Knowledge-Fitting Theory of Plausibility accounts for the empirical results reported in Chapters 3 and 4, and resolves some of the contradictions found in the literature. First, the theory provides us with a specific proposal on the nature of concept-coherence; it is based on the amount of background knowledge needed to properly comprehend and represent the scenario. Later, we cash out this idea in a computational model of the comprehension and assessment stages. Second, the theory makes allowances for differential influences of word-coherence and concept-coherence at different stages of the plausibility judgement process.

Accounting for Effects on Plausibility Ratings

In the Knowledge-Fitting Theory's account, when people rate the plausibility of a sentence pair that describes events, they do this by assessing the concept-coherence of the representation. So, the theory would predict only an effect of concept-coherence on plausibility ratings (Chapter 3, Experiments 1-3). We saw that causal scenarios are rated the most plausible, followed by attributal, temporal, and unrelated scenarios. The Knowledge-Fitting theory explains this finding by describing how the concept-coherence of the representation is determined by the interaction of potentiality and complexity. Causal scenarios have high potentiality, because they are likely to prime the knowledge that the inferential process will need to make the connection between the events. Attributal and temporal scenarios have medium potentiality, because they are less likely to prime such useful knowledge. Finally, unrelated scenarios have low potentiality, because they are unlikely to prime much knowledge that could help to connect the events. This makes causal scenarios

the most plausible, followed by attributal and temporal scenarios tied on medium plausibility, and unrelated scenarios the least plausible. However, we must also take complexity into account, although is plays a lesser role than potentiality in determining concept-coherence. Causal and temporal scenarios have high complexity, because a lot of background knowledge must be folded into the representation to make the inference. The same is true of unrelated scenarios, which will have high complexity if any inference can be made to connect the events. Attributal scenarios, on the other hand, have low complexity because they simply involve inferring a co-referent (i.e., attaching the adjective to an entity in the first sentence). When we allow potentiality and complexity to interact, we find that concept-coherence (and therefore plausibility) is highest for causal scenarios, followed by attributal, temporal, and finally unrelated scenarios.

Accounting for Effects on Comprehension Times

The Knowledge-Fitting Theory holds that both word-coherence and concept-coherence influence the comprehension stage, and so would predict effects of both on comprehension times (Chapter 4, Experiment 4). First, we saw that distributionally distant sentences are understood more quickly. The Knowledge-Fitting Theory explains this by its account of the distributional spotlight. Distributionally distant sentences activate a greater region of distributional information, which in turn primes more prior knowledge for use in inferencing. This gives distributionally distant sentences an advantage over close sentences, because they are more likely to prime knowledge that the inferential process will need. Second, we saw that less complex inferences can be made more quickly. The theory

explains this by its account of the complexity of inferences. Comprehension times will be shorter for inferences that do not require so much background knowledge to be incorporated into the representation. This gives attributal inferences an advantage over causal inferences, because less extra knowledge has to be retrieved and integrated to represent an attributal connection.

Accounting for Effects on Plausibility Judgement Times

According to the Knowledge-Fitting Theory, the plausibility judgement process is made up of the comprehension and assessment stages. Thus, when we measure how long it takes to judge if a scenario is plausible, we are in fact measuring the time taken to perform both these stages. This means that the Knowledge-Fitting Theory would predict that the effects on comprehension times will carry through to plausibility judgement times. In this way, the theory explains, first, why we saw that plausibility judgement times are shorter for distributionally distant sentences, and second, why we saw that plausibility judgement times are shorter for less complex inferences. However, the Knowledge-Fitting Theory also holds that concept-coherence influences the assessment stage; that is, that it would be faster to assess inferences that did not incorporate much background knowledge into the representation. This means that, third, the theory explains why plausibility judgement times show an even greater effect of concept-coherence than comprehension times.

Resolving Apparent Contradictions in Distributional Effects

Distributional distance has been found to affect the plausibility ratings of adjective-noun pairs (Lapata, McDonald & Keller, 1999) but not event scenarios (Chapter 3, Experiments 2 and 3). This apparent contradiction of effects is resolved within the Knowledge-Fitting Theory. When people rate the plausibility of a sentence pair that describes events, they do this by assessing the concept-coherence of the representation. Word-coherence does not play a role in the assessment stage, and hence no distributional effects are found in the plausibility ratings of Experiments 2 and 3. In contrast, when people rate the plausibility of an adjective-noun pair, they have no scenario to represent and no inferences to make, and therefore have no means of assessing concept-coherence. This forces people to base their plausibility ratings on whatever information they have. All that is available is the word-coherence of the adjective and noun, and hence, we see a distributional effect in Lapata et al's plausibility ratings.

Resolving the Quickness of Plausibility Over Retrieval

It has been shown that plausibility is used as a kind of cognitive shortcut in place of direct retrieval from long-term memory, especially once verbatim memory has faded (e.g., Reder, 1982; Reder & Ross, 1983; Reder, Wible & Martin, 1986). For example, Reder (1982) asked people to read short stories, and after a delay presented them with a sentence. She found that people were faster to confirm that the sentence was plausible given the story than confirm that the sentence appeared in the story. This effect can also be accounted for by the Knowledge-Fitting Theory. When people are asked to judge whether they have seen a particular item before,

they must search through their representation of the scenario it refers to. In addition, they must try to separate what they inferred in the scenario from what they were given (i.e., the actual story), as only the latter information is relevant to the task. If they find a match between the item and their memory of the scenario, then they can stop searching. All this takes time. In contrast, when people are asked to judge whether a particular item is plausible, all they must do is check if the item fits the scenario. If the item can be added to the representation with an inferential connection, then it is plausible. There is no need to painstakingly search the representation for a particular item and no need to distinguish between given and inferred information; hence, plausibility judgement can be performed more quickly than direct retrieval.

Resolving the Inverse Fan Effect in Plausibility Judgements

When people are asked whether they have seen a particular fact before, their response times exhibit a fan effect; that is, the more one knows about a particular topic, the slower the retrieval of any given fact (e.g., Anderson, 1974; Radvansky, Spieler & Zacks, 1993; Reder & Anderson, 1980). However, it has been shown that the fan effect is inverted for plausibility judgements; that is, the more one knows about a particular topic, the faster one can judge if a given fact is plausible / consistent (Reder & Anderson, 1980; Reder & Ross, 1983). The Knowledge-Fitting Theory can account for this phenomenon in the same way it accounts for the quickness of plausibility over retrieval. When people are asked to judge whether they have seen a particular fact before, they must search through their representation of the scenario it refers to. The more facts that are in their representation, the longer

it takes to find a match for the item in question. In contrast, when people are asked to judge whether a given fact is plausible, all they must do is check if it fits the scenario. In this case, the more facts that are in the representation, the more opportunities that one has to create an inferential connection with the item in question. Unlike retrieval, known facts do not have to be painstakingly searched and do not interfere with the given item; hence, plausibility judgement is facilitated rather than hampered by knowing a lot about a particular topic.

5.3 – PAM: The Plausibility Analysis Model

One of our key criticisms of the plausibility literature to date is that ideas of concept-coherence were not well specified. In our Knowledge-Fitting Theory we have tried to be more precise about the exact nature of concept-coherence. It is not yet clear that these proposals are much better than what came before, without showing that the theory can be instantiated in a set of effective procedures. In the remainder of this chapter, we describe the Plausibility Analysis Model (PAM); a cognitive model of human plausibility judgements that implements both the comprehension and assessment stages of the theory taking into account both distributional distance and inference type. We also show how well simulations run on this model correspond to our plausibility judgement data.

PAM takes sentence inputs and outputs a plausibility rating (from 0 – 10) and an estimated plausibility judgement time (in milliseconds) for the scenario described in the sentences (see Appendix C for system diagrams). The comprehension stage is modelled by representing a scenario through a combination of distributional analysis

and knowledge-fitting, and the assessment stage is modelled by examining the quality of this knowledge fit in the representation of the described scenario.

5.3.1 – Modelling the Comprehension Stage

The comprehension stage takes a sentences pair as input and outputs to the assessment stage a representation of the scenario described in those sentences. During comprehension, PAM carries out a distributional analysis of the sentence pair and makes appropriate inferences by fitting the scenario to relevant prior knowledge

Distributional Analysis

When a sentence is first read, it spotlights (activates) a neighbourhood of distributional knowledge. PAM models this process by the use of a model of linguistic distributional knowledge: LSA (Landauer & Dumais, 1997). LSA (in the form used by PAM) is a statistical model of the distributional patterns of English words, which works by passing a window over a corpus that represents the lifetime readings of an American first-year university student [2]. Each sentence in the input pair is represented in LSA as a point in high-dimensional distributional space. PAM uses LSA to calculate the 50 words that are the nearest neighbours of each sentence. These 50 words are PAM's distributional spotlight for that sentence. When the distributional spotlight has been found for each sentence in the pair, the union of the

[2] In LSA parlance, the analysis was 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, any words with a frequency in the corpus of less than 5 were excluded.

two sets of words is found. The number of unique words in this set union is a measure of the distributional distance of the two sentences. If the sentence spotlights are far apart, there will be large number of unique words in the union (no overlap = 100 words). If the sentence spotlights are close together, there will be fewer words in the union (perfect overlap = 50 words). This *distributional word count* represents the coverage of the distributional spotlight, and is used by PAM in estimating the plausibility judgement time for the sentence pair.

Knowledge-Fitting

To represent the scenario, PAM must break down each sentence into propositional form. First, each sentence is converted into a simplified form with the aid of a synonym and morphological lookup table. This replaces words with their more common synonyms, singularises plural nouns, and changes verbs into the present tense third person singular form; for example, the sentences "*the hounds growled*" and "*the dogs snarled*" are both converted to the same simplified form "*the dog growl*". Next, the simplified sentences are parsed according to a set of basic grammatical rules and converted into propositions. To do this, PAM passes the sentences through a simple syntactic parser that extracts verbs and adjectives to use as predicate names, and extracts nouns to use as arguments; for example, the simplified sentence "*the dog growl*" is converted into the proposition *growl(dog)*. It should be noted that the ease with which PAM can break sentences into propositional form is due to the regular syntax of the sentence pairs. Although the automatic conversion of text into propositions is not a trivial task (see Kintsch,

1998), the current syntactic form of PAM's input lends itself quite well to automation.

Once the sentences are in propositional form, PAM makes the inferences between the sentences by fitting their propositions to information in the knowledge base. PAM's knowledge base is organised as a predicate set, where each entity (noun) is defined as part of a type hierarchy and each predicate (verb) is defined by the conditions of its constituent arguments in PAM's knowledge base. All entries in the knowledge base were added in a "blind" fashion; that is, each entity and predicate was defined as thoroughly as possible in terms of argument conditions without reference to the original sentence pairs. In order to represent a particular scenario, PAM must check the conditions of each proposition as it is defined in the knowledge base. For example, Table 5.1 gives an example of the knowledge base representation of the scenario *see(pack, fox), growl(hounds)*. In representing this scenario, PAM must first check the predicate *see* in the knowledge base to determine if the arguments meet the conditions specified. The *see* predicate requires that its first argument be an *animal* (i.e., something must be an animal to see). As the definition of *pack* shows that it contains *dogs*, and the type hierarchy for *dog* shows that it is an *animal*, the first condition of *see* is met. Also, the *see* predicate requires that its second argument must be a non-abstract entity (i.e., something must be non-abstract to be seen). Since the type hierarchy of *fox* shows that it is an *animal* and not an *abstract entity*, the second condition of the *see* predicate is met. The way in which each condition is met is listed, and if all conditions are fulfilled, PAM returns this list as a path (as given in Table 5.1).

Table 5.1 – Example of PAM's representation of a sentence pair, giving conditions in a single sample path.

Sentence	*Proposition*	*Path*
The pack saw the fox.	see(pack, fox)	BECAUSE pack is animal AND fox is non-abstract
The hounds growled.	growl(dog)	BECAUSE dog is growling at fox → BECAUSE dog is hunting fox → BECAUSE dog is predator AND fox is prey

When the first proposition has been represented, PAM moves on to processing the second proposition, *growl(dog)*, and searches for ways to meet the conditions of the *growl* predicate. Table 5.1 shows one of the paths that PAM finds for this proposition; it represents the ideas that the dogs are growling because they are growling at the fox, because they are hunting it, because dogs are predators and foxes are prey. However, there are many other reasons for dogs to growl, such as because they are generally aggressive or because they are fighting amongst themselves. Some of these conditions lead to other predicates which have their own conditions attached, such as *hunt(dog)* which requires that *dog* must be a predator and that the *fox* of the first sentence must be prey. More often than not, there are several paths in the knowledge base that could be followed to fulfil the conditions of a particular predicate, and PAM will record all these alternative paths. Sometimes, a path may involve conjecture; that is, the path contains a condition that could only be fulfilled by assuming the existence of a hypothetical entity not explicitly mentioned. For example, the *dogs* may *growl* because they are afraid, but that would involve assuming the existence of something to frighten them. PAM also records these

hypothetical paths, and marks them as such. In this respect, PAM models group behaviour in plausibility judgement; rather than limit the representation to a single path that one individual may consider, PAM represents the set of paths that a group may consider and averages out the differences.

The scenario is therefore represented by PAM in the form shown in Figure 5.3. A representation consists of several distinct paths, each of which consists of a set of one or more conditions. There is no hard-coded distinction between different types of inference; PAM simply tries to build a path by drawing in whatever information is necessary to fulfil the conditions in the predicate. The structure of this representation is analysed to determine its inferential complexity, and is used by PAM in estimating the plausibility rating and judgement time for the sentence pair.

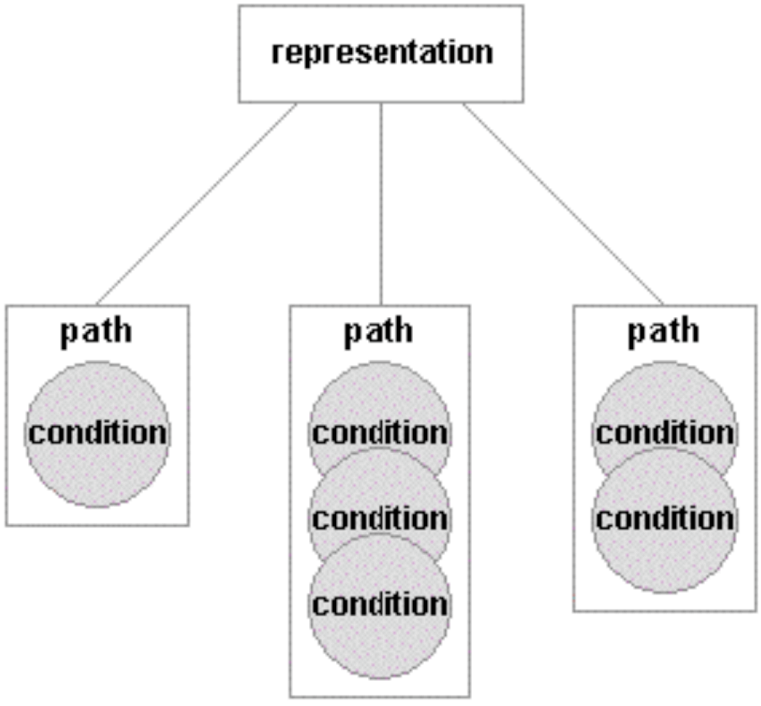


Figure 5.3 – Form of scenario representation created by PAM in the comprehension stage. It is then analysed in the assessment stage to determine plausibility.

5.3.2 – Modelling the Assessment Stage

Once a scenario has been comprehended, the representation is taken as input to the assessment stage. PAM uses this representation to model two different types of plausibility judgement: plausibility ratings and plausibility judgement times. Each of these types of judgement extracts different variables to ascertain the quality of the knowledge fit in the representation.

Plausibility Ratings

PAM analyses the representation to estimate the plausibility of the scenario, returning a rating between 0 (not plausible) and 10 (completely plausible). In this analysis, PAM extracts three main variables from the representation:

1. *Total Number of Paths (P).* This is quantified as the number of different paths in the representation. It partially represents the potentiality of the representation.

2. *Mean Path Length (L).* This is quantified as the sum of all path lengths in the representation (i.e., all conditions across all paths) divided by *P*. It represents the complexity of the representation.

3. *Proportion of Hypothetical Paths (H).* This is quantified as the number of paths that contain a condition met by conjecture, divided by *P*. It represents the case where a scenario can be represented only by assuming the existence of something not explicitly mentioned – for example, "*The bottle fell off the shelf. The bottle melted*" can return a plausible path if we allow that the bottle may have fallen into a hypothetical furnace. It partially represents the potentiality of the representation.

4. *Distributional Word Count (D).* This measures the number of terms in the unified set of the sentences' nearest neighbours, and ranges from 50-100. It partially represents the potentiality of the representation.

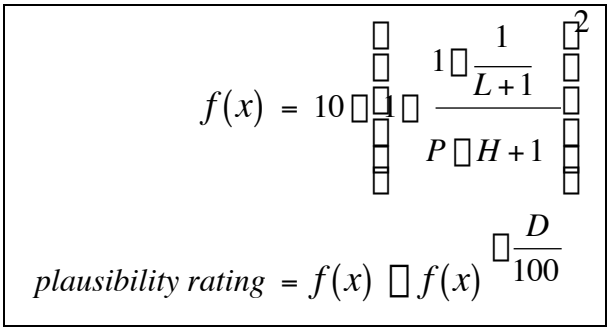


Figure 5.4 – PAM's formula for plausibility ratings (variables are described in the text).

The exact rating is calculated according to the asymptotic functions of Figure 5.4. In short, a high number of paths (*P*) and a high distributional word count (*D*) means higher plausibility, because there are more ways that the representation can use primed knowledge. A high mean path length (*L*) means lower plausibility, because more conditions had to be met to connect the sentence events. Finally, a high proportion of hypothetical paths (*H*) means lower plausibility, because it assumes the existence of entities that may not be present.

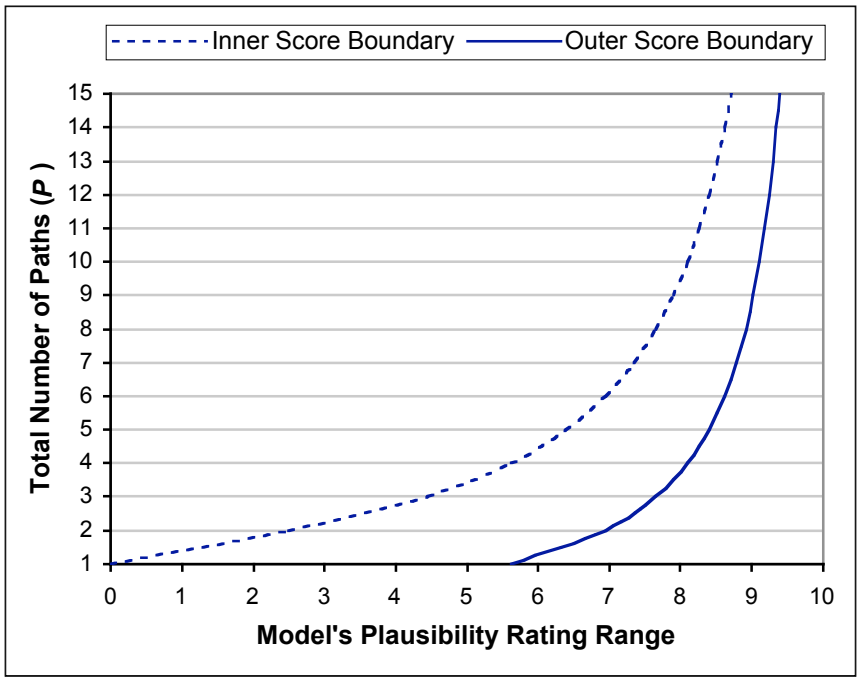


Figure 5.5 – Graph showing the inner and outer score boundaries of PAM's asymptotic plausibility rating function.

Figure 5.5 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, every path is hypothetical, and the sentences are distributionally close together. The solid line represents the outer (upper) score boundary, where the mean path length is one, no path is hypothetical, and the sentences are distributionally far apart. For example, take a sentence pair with a best-case distributional word count of 100. A set of four (non-hypothetical) paths with a mean length of three will then have a rating of 7.1 out of 10, while a set of three paths (again with a mean length of three) will have a rating of 6.5 out of 10. Had the distributional word count been 80 (still with three paths), then the rating would be lower at 6.4 out of 10. If even one of those three paths were hypothetical, then the rating would drop to 6.1 out of 10.

Plausibility Judgement Times

PAM analyses both the representation and the sentence itself to estimate the time required to judge plausibility, returning the judgement time in milliseconds. In this analysis, PAM extracts four main variables from the representation and sentence:

1. *Mean Path Length (L).* This is quantified as the sum of all path lengths in the representation (i.e., all conditions across all paths) divided by the number of different paths. It represents the complexity of the representation.
2. *Distributional Word Count (D).* This measures the number of terms in the unified set of the sentences' nearest neighbours, and ranges from 50-100. It represents the coverage of the distributional spotlight.
3. *Number of Syllables (S).* This is quantified as the number of syllables in the second sentence.
4. Or*thographic Length (O).* This is quantified as the number of characters (letters) in the second sentence.

$$plausibility\ judgement\ time = a + bL + cS + dO - eD$$

Figure 5.6 – PAM's formula for plausibility judgement time estimation (variables are described in the text).

The exact plausibility judgement time estimate is calculated according to the linear function shown in Figure 5.6, where each lowercase letter (*a*-*e*) represents a constant. In brief, a high path length (*L*) means a longer judgement time, because

there are more conditions to be met to connect the sentence events, and then examined in the assessment stage. A high distributional word count (*D*) means a faster judgement time, because of the greater amount of knowledge that is primed by the coverage of the distributional spotlight. Finally, a high number of syllables (*S*) or orthographic length (*O*) also means a longer judgement time, because of the increased time needed to read the sentence [3].

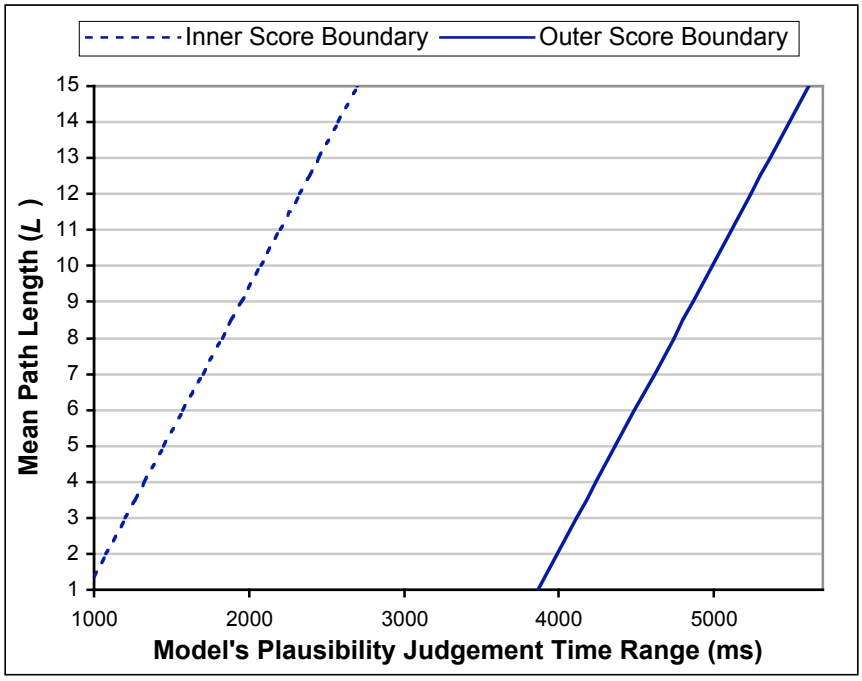


Figure 5.7 – Graph showing the inner and outer score boundaries of PAM's linear plausibility judgement time estimation function.

Figure 5.7 shows the boundaries of plausibility judgement times that PAM generates for increasing path lengths. The dotted line represents the inner (lower)

[3] The variables of syllable count and orthographic length have been included because it is generally accepted that these factors affect comprehension times. However, in PAM's estimation of plausibility judgement times, these variables are weighted much more lightly than the key variables of mean path length and distributional word count, and so do not in themselves deliver the main effects observed in the model.

score boundary, which is the best-case situation where the second sentence contains the minimum number of syllables and characters, and sentences are distributionally far apart. The solid line represents the outer (upper) score boundary, where there is a large number of syllables and characters in the second sentence, and the sentences are distributionally close together. For example, let us take the sentence "*The hounds were fierce*" that has 4 syllables and 19 characters. If this sentence had a mean path length of three and a distributional word count of 80, it will have an estimated judgement time of 2033ms. A mean path length of four would increase this estimate to 2158ms. If the distributional word count was 100 (still for a path length of four), then the estimate would drop to 1978ms.

5.4 – Model Evaluation

To evaluate the model, we compared PAM's output to human responses in two simulations. Simulation 1 compares PAM's estimated plausibility ratings to the ratings produced by people in Chapter 3's experiments. Simulation 2 compares the estimated plausibility judgement times produced by PAM to those found in Chapter 4, Experiment 5. This means that the model was run on the same sentence pairs presented to the human participants. As previously noted, the knowledge base used in PAM was built in a "blind" fashion; each of the individual predicates was defined using simple definitions of argument conditions, without checking possible path lengths that might emerge from combining these words in a sentence. Such knowledge bases will always be a crude approximation of the knowledge of a particular individual, but they should be closer to the aggregate knowledge that a

group of participants bring to the task. The critical point was that the knowledge base was not modified in any iterative way to fit the data. Also, the test sentence pairs used in this simulation represented a different subset of materials to those used as PAM's training items, and thus could test the generalisability of the model.

5.4.1 – Simulation 1: Plausibility Ratings

Method

Materials. The materials for this simulation consisted of 60 sentence pairs with manipulations of concept-coherence, and were drawn from the experiments in Chapter 3. Each sentence pair had one of four inference types connecting the sentences (causal, attributal, temporal and unrelated). There were more causal and attributal sentence pairs than temporal and unrelated pairs, due to the unequal distribution of inference types across the experiments, with 22 causal, 20 attributal, 9 temporal and 9 unrelated sentence pairs.

Procedure. The human procedure is detailed in Experiments 1, 2 and 3 of Chapter 3. Participants were 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. For the purposes of this simulation, the mean plausibility rating for each sentence pair (in each condition) was used. The procedure for PAM was to enter each natural language sentence pair and note the estimated plausibility rating. Each rating (0 – 10) was output rounded to one decimal place.

Results and Discussion

The simulation shows that PAM's output accurately reflects the product of human plausibility judgements. Inference type effects on plausibility ratings are accurately modelled, with plausibility decreasing from causal>attributal>temporal>unrelated.

Table 5.2 gives the mean ratings per condition compared to the human responses from Experiment 1 in Chapter 3. PAM's estimates correlate strongly with participants' plausibility judgements, and regression analysis suggests that the model could be used as a successful predictor of human plausibility ratings [r=0.788, r^2=0.621, N=60, p<0.0001). Furthermore, PAM's estimates reveal the same response patterns found for human plausibility judgements. An analysis of PAM's ratings showed a reliable effect of inference type [F(3, 56)=115.644, p<0.0001, MSe=0.943].

Table 5.2 – Mean plausibility ratings per inference type as produced by participants and by PAM in Simulation 1, on a scale from 0-10 where 0 is implausible and 10 is very plausible.

Inference Type	*Human Ratings*	*Model Ratings*
Causal	7.8	8.1
Attributal	5.5	5.8
Temporal	4.2	5.2
Unrelated	2.0	1.0

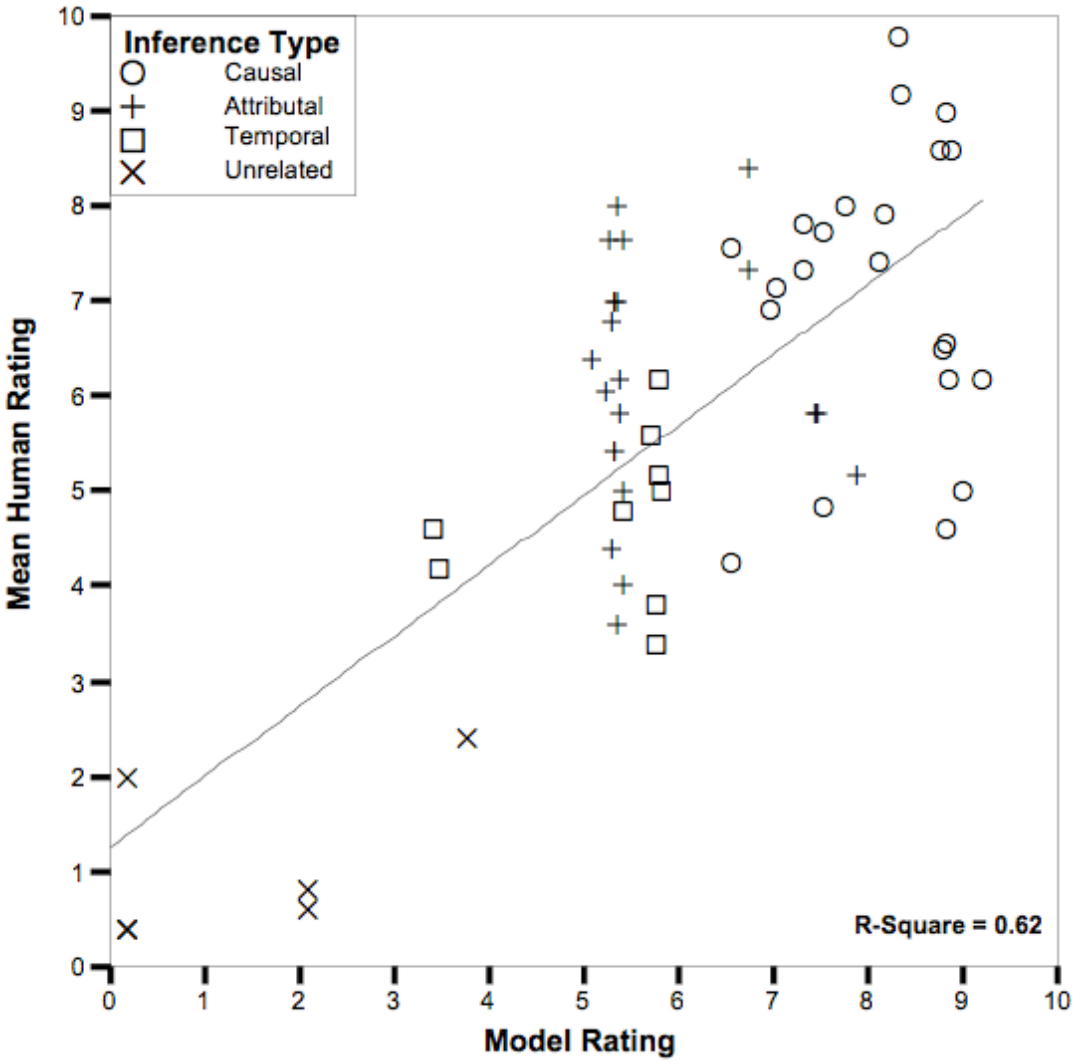


Figure 5.8 – Scatterplot of relationship between plausibility ratings produced by PAM in Simulation 1 and by participants (r=0.788), with each sentence pair distinguished by inference type.

Additionally, a multiple regression analysis (including item as a predictor) showed that inference type was a significant predictor of the estimated plausibility ratings [standardised coefficient *beta*=0.890, $t(57)$=14.980, p<0.0001], even though separate inference types were not explicitly encoded in the model. Figure 5.8 shows a scatterplot of the relationship between model output and participant means with each inference type distinguished.

5.4.2 – Simulation 2: Plausibility Judgement Times

Method

Materials. The materials for this simulation were those used in experiments in Chapter 4, and consisted of 60 sentence pairs with manipulations of word- and concept-coherence. Each sentence pair had one of two inference types connecting the sentences (causal and attributal), and one of two distributional distances (close and distant). There were 15 sentence pairs in each of the four conditions: that is, causal close, causal distant, attributal close, and attributal distant.

Procedure. The human procedure is detailed in Experiment 5, Chapter 4. For the purposes of this simulation, the mean plausibility judgement time for each sentence pair (in each condition) was used. The procedure for PAM was to enter each natural language sentence pair and note the estimate of plausibility judgement time. Each estimate was output rounded to the nearest millisecond.

Results and Discussion

The simulation shows that PAM's output accurately reflects the time course of human plausibility judgements. Distributional distance effects on plausibility judgement times are accurately modelled, as are the effects of different inference types.

Table 5.3 gives the mean response times per condition compared to the human responses from Experiment 5 in Chapter 4. PAM's estimates correlate

strongly with participants' plausibility judgements, and regression analysis suggests that the model could be used as a successful predictor of human plausibility judgement times [r=0.633, r^2=0.401, N=60, p<0.0001). Furthermore, PAM's estimates reveal the same response patterns found for human plausibility judgements. An analysis of PAM's estimates showed a reliable effect of distributional distance, with distant items faster than close items [F(1, 56)=4.382, p<0.05, *MSe*=118884.0], and a reliable effect of inference type, with attributal items faster than causal items [F(1, 56)=23.009, p<0.0001, *MSe*=118884.0].

Additionally, a multiple regression analysis (including item as a predictor) showed that both inference type [standardised coefficient *beta*=0.522, t(56)=4.773, p<0.0001] and distributional distance [standardised coefficient *beta*=0.229, t(57)=2.098, p<0.05] were significant predictors of the estimated plausibility judgement times, with inference type the stronger predictor of the two. Figure 5.9 shows a scatterplot of the relationship between model output and participant means with each inference type and distributional distance distinguished.

Table 5.3 – Mean plausibility judgement times (in milliseconds) per inference type and distributional distance, as produced by participants and by PAM in Simulation 2.

Inference Type	*Distributional Distance*	*Human Times*	*Model Times*
Causal	Close	2310	2325
	Distant	2083	2139
Attributal	Close	1907	1898
	Distant	1728	1712

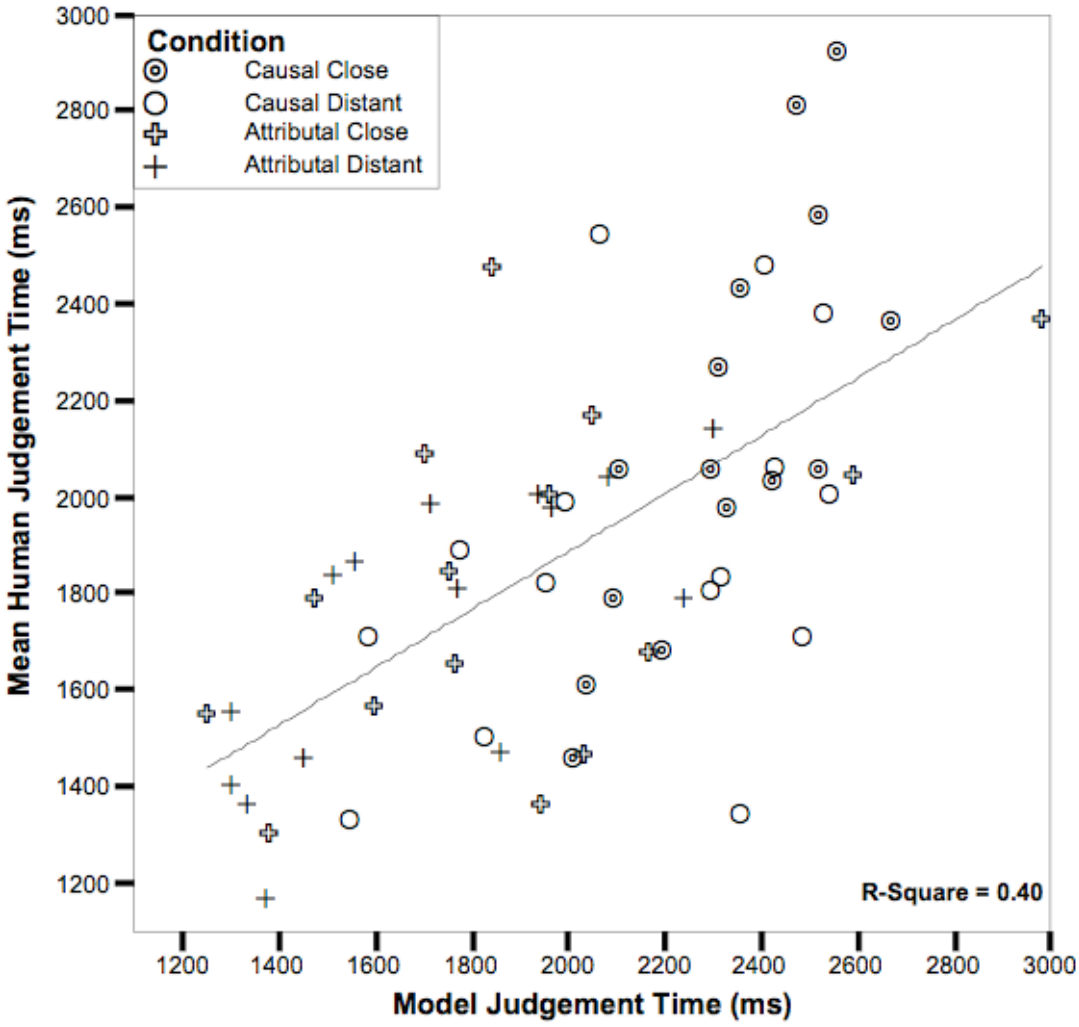


Figure 5.9 – Scatterplot of relationship between plausibility judgement times produced by PAM in Simulation 2 and by participants (r=0.633), with each sentence pair distinguished by inference type and by distributional distance.

5.4.3 – Simulation Discussion

The simulations show that PAM's output accurately reflects human plausibility judgements. Distributional distance effects on plausibility judgement times are accurately modelled, as are the effects of different inference type on plausibility judgement times and plausibility ratings.

Distributional distance effects are modelled by the use of LSA, where the distributional spotlight of the sentence pair is calculated. This number of words in the spotlight – the *distributional word count* – is then one of the variables in the formula for estimating plausibility judgement times. Sentences that lie close to one another in distributional space have smaller spotlight coverage, and hence a smaller distributional word count than sentences that lie far apart. The greater the distributional distance between sentences, the higher the distributional word count and the faster the judgement time estimate. This gives rise to the significant difference between distributionally close and distant sentence pairs in the simulation.

Inference type effects are modelled by extracting variables from the representation formed by PAM. The formula for plausibility ratings uses three extracted variables (number of paths, mean path length, and proportion of hypothetical paths), while the formula for plausibility judgement times uses one (mean path length). However, PAM does not distinguish between the different types of inferences that may connect sentences; there is no hard-coded differentiation in either the knowledge base or the model framework. Yet PAM produces distinctly different plausibility ratings and judgement times for different types of inference. So how does this happen? The answer lies in how each inference type tends towards certain values for each of the extracted variables.

Table 5.4 – How each inference type tends towards high or low values for the variables extracted from PAM's representation. High *P* values increase plausibility ratings, while high *H* and *L* values decrease plausibility. High *L* values also serve to increase plausibility judgement times.

Inference Type	*Extracted Variable*		
	Number of Paths (P)	*Proportion of Hypothetical Paths (H)*	*Mean Path Length (L)*
Causal	High	Low	High
Attributal	Low	Low	Low
Temporal	High	High	High
Unrelated	Low	High	Low

Table 5.4 illustrates how each inference type tends towards high or low values for each of the extracted variables (number of paths *P*, proportion of hypothetical paths *H*, and mean path length *L*). For plausibility ratings, the most plausible scenario will have a high number of paths, a low proportion of hypothetical paths, and a low path length. By the interaction of the first two variables shown in Table 5.4 (number of paths and proportion of hypothetical paths), causal inferences are the most plausible, attributal and temporal inferences are tied with medium plausibility, and unrelated inferences have the lowest plausibility. Taking the last variable (mean path length) into account, attributal inferences become more plausible than temporal inferences. This gives rise to the causal>attributal>temporal>unrelated trend in plausibility ratings seen in Simulation 1. For plausibility judgement times, slower estimated responses result from having a high mean path length (i.e., a high number of conditions connecting the events). This gives rise to the significantly faster estimates for attributal pairs than causal pairs seen in Simulation 2.

For any cognitive model, it is important that the key parameters are motivated by the theory and not motivated simply by the need to make the model work (i.e., the so-called the A|B distinction for cognitive models: Cooper, Fox, Farringdon & Shallice, 1996). In PAM, the variables used in calculating plausibility ratings and plausibility judgement time are critical to the behaviour of the model as a whole. In this sense, all three variables are 'A' components that are relevant to the theoretical rather than implementational aspects of the model. This is not to say that PAM is the only means of computationally modelling the Knowledge-Fitting Theory of Plausibility (i.e., we make no claims that PAM's mathematical formulae represent literal psychological reality). However, PAM's performance in modelling human plausibility responses shows that it is a successful computational implementation of the Knowledge-Fitting Theory.

CHAPTER 6 – CONCLUSIONS

6.1 – Summary of Accomplishments

The objective of this thesis was to explain plausibility in and of itself. To this end, we have presented a new theory of how people make plausibility judgements, called the Knowledge-Fitting Theory of Plausibility. Empirically, as well as resolving some apparent conflicts in the literature, the theory has been supported by a series of experiments on comprehension and plausibility judgement. Theoretically, we advance a fully specified and implemented account of plausibility that explains how key factors act to influence the plausibility judgement process.

6.1.1 – Empirical Novelties

In a series of experiments, we investigated plausibility judgement using two different paradigms. Chapter 3 used a ratings paradigm to examine plausibility judgements that are made with considered thought, and asked people to determine exactly *how* plausible they found each scenario. Chapter 4 used an online paradigm to examine plausibility judgements that are made quickly, and asked people to rapidly determine whether each scenario was plausible or implausible. A number of empirical novelties resulted from these experiments:

- ❖ Plausibility ratings are affected by concept-coherence, as shown by Experiments 1 and 3. Sentences with no obvious inferential link between them are rated as barely plausible, with temporal, attributal and causal inferences ranged in increasing plausibility.

- ❖ Plausibility ratings are not affected by word-coherence, as shown by Experiments 2 and 3. When a scenario requires an inferential connection to be made between events, distributionally close sentences are no more or less plausible than distributionally distant sentences.

- ❖ Comprehension times are affected by word-coherence, as shown by Experiment 4. A sentence that is distributionally distant from its predecessor is understood more quickly than a sentence that is distributionally close to its predecessor.

- ❖ Comprehension times involving different inference types are affected to different extents by word-coherence, as shown by Experiment 4. When understanding a scenario involves making a causal connection between events, there is an even greater effect of distributional distance than when an attributal connection is involved.

- ❖ Assessment times (i.e., the time spent assessing a scenario's plausibility) are not affected by word-coherence, as shown by the meta-analysis of Experiments 4 and 5. A scenario described with distributionally distant sentences is assessed no more quickly or slowly than a scenario described with distributionally close sentences.

- Assessment times (i.e., the time spent assessing a scenario's plausibility) are affected by concept-coherence, as shown by the meta-analysis of Experiments 4 and 5. A scenario that involves an attributal connection between events is assessed more quickly than a scenario that involves a causal connection.

6.1.2 – Theoretical Novelties

The Knowledge-Fitting Theory of Plausibility is based on our empirical findings regarding the nature of plausibility. In our theory, we describe the process of making a plausibility judgement, and detail how both word-coherence and concept-coherence influence this process. The key theoretical principles are validated by the computational implementation of the theory, the Plausibility Analysis Model (PAM), where simulations show a close correspondence between the model and our plausibility judgement data. There are a number of theoretical novelties in this thesis, which may be summarised as follows:

- Plausibility judgement is described as spanning two stages: *comprehension* (where a mental representation of the scenario is created) and *assessment* (where the representation is examined to determine how well the scenario fits prior knowledge).

- The comprehension of a scenario is affected by both *word-coherence* (the distributional properties of the words in the description) and *concept-coherence* (the background knowledge needed to make the scenario fit what we know about the world).

- The assessment of plausibility is affected by *concept-coherence* alone (the background knowledge that determines the quality of fit between the scenario and what we know of the world).

- Word-coherence effects are defined in terms of the *distributional spotlight*: each sentence spotlights an area of distributional knowledge, and each spotlit area then primes associated prior knowledge. Thus, comprehension is best facilitated by distributionally distant sentences that make available the greatest amount of background knowledge. In this way, the distributional spotlight offers an explanation of how linguistic cues can activate relevant knowledge in long-term memory.

- Concept-coherence effects are defined in terms of the prior knowledge needed to represent the scenario: different inferential connections require different background knowledge to link events. Thus, a scenario can be comprehended and assessed more easily if it is of low *complexity* (i.e., if its inferential connections need little background knowledge). Also, a scenario will be judged to be most plausible if it has low complexity and high *potentiality* (i.e., if most of this knowledge had been primed by the scenario description). In this way, complexity and potentiality offer a well-specified explanation of concept-coherence that was previously absent from the literature.

6.2 – Further Research

As they currently stand, the Knowledge-Fitting Theory of Plausibility and its computational implementation, the Plausibility Analysis Model (PAM), represent a well-specified account of plausibility judgement. However, this work suggests several further issues connected to the extension of the current analysis in a number of directions.

6.2.1 – Extending PAM's Dialectal Capacity

PAM implements the Knowledge-Fitting Theory using Latent Semantic Analysis to model distributional effects. The version of LSA used was trained on a corpus representing the cumulative lifetime readings of an eighteen-year-old American college student. This makes LSA a model of American English, with the inevitable result that it is less accurate for other dialects of English. For example, *football* in American English refers to American football, with related words like quarterback, touchdown, and superbowl. However, *football* in British English generally refers to soccer, with related words like goalie, offside, and premiership. This is complicated further in Hiberno-English and Australian English, where *football* can again refer to completely different sports, thus having different sets of related words. Thus, words that are distributionally close for speakers of one dialect can be distributionally distant for speakers of other dialects. Yet despite LSA's focus on American English, PAM can successfully model the distributional effects on Hiberno-English speakers' plausibility judgements. Given that distributional knowledge results from exposure to language, and that languages can differ

markedly between dialects, it is notable that PAM performed so well in the face of a dialectal mismatch. If alternative distributional models were created by running the LSA algorithm on corpora of other dialects such as Hiberno-English or British English, then PAM could more accurately model the distributional influence on plausibility for speakers of those dialects. This extension of PAM could also allow us to test how speakers of different dialects may differentially perform plausibility judgements.

6.2.2 – Extending PAM to Longer Discourse

In the present work, PAM was tested solely on the sentence pairs used in our experiments. This raises the question about the extendibility of the model to longer pieces of discourse. We see no obstacle, in principle, to the extension of the model to longer discourse. In the comprehension stage, each subsequent sentence would be folded into the representation in exactly the same way that the second sentence of the pair was processed, as there is no constraint on the size of the path-based representation. In the assessment stage, there is similarly no constraint on the size of the representation that could be subsequently analysed. With longer pieces of discourse, there may nonetheless be a need for some additional functionality. For example, we might want the distributional activation of regions from older sentences to decay over time, or we might need to optimise the analysis of the representation in some way were it to grow to several hundred paths. However, these additions are fine-tunings of PAM; they do not represent changes to the fundamental precepts of the theory.

6.2.3 – Extending the Theory to Other Input Modalities

At present, the Knowledge-Fitting Theory of Plausibility is firmly tied to the processing of verbal discourse, whether that be spoken or written text. Ultimately, we would like to extend the account to other forms of input such as pictures. It has been shown that people can judge the plausibility of a sentence just as quickly when a picture substitutes for one of its words as when the sentence contains only words or they read or hear text-based stories (Gernsbacher 1985, cited in Gernsbacher, 1997). It has also been shown that people draw the same inferences about a sequence of events whether the events were presented as pictures or sentences (e.g., Baggett, 1975). These findings suggest that the same processes of distributional spotlighting and knowledge priming could occur for linguistic and pictorial input. The challenge here is theoretical; to relate the pictorial input to the distributional knowledge that is primarily concerned with language. It is possible that an image (e.g., of an apple) could cause the activation of a distributional spotlight around its referent word (e.g., "apple"), as images have been shown to prime related words (Brown, Neblett, Jones & Mitchell, 1991; Stenberg, Radeborg & Hedman, 1994; Koivisto & Revonsuo, 2000)). However, it is not yet clear how this would operate for images of polynomial objects – would a picture of a sofa activate a spotlight around "sofa", "couch", "seat", or all three? These issues would need to be resolved before the Knowledge-Fitting theory could be extended to pictorial input.

6.3 – General Implications

The empirical, theoretical and computational issues raised in this work have important implications across a variety of fields. In our review of the relevant literature in Chapter 2, three main areas of research were discussed – previous treatments of plausibility, discourse comprehension, and distributional analysis – each of which is impacted by the word reported here.

6.3.1 – Implications for Previous Treatments of Plausibility

Across the wider cognitive psychology literature, there has been a shared view that plausibility has something to do with conceptual consistency with existing knowledge (e.g., Black, Freeman & Johnson-Laird, 1986; Costello & Keane, 2000; Reder, 1982). However, other research has suggested that plausibility may be concerned with word-level distributional information (Lapata, McDonald & Keller, 1999). The Knowledge-Fitting Theory of Plausibility resolves these apparently dichotomous views of plausibility, and offers a fully-specified account of how both conceptual and distributional factors affect the judgement process.

In the Knowledge-Fitting Theory, concept-coherence is defined in terms of the background knowledge needed to make inferential connections between events, and fit the scenario to what we know of the world. This explanation is both more specific and broader-reaching than any previous account. For example, it impacts upon areas such as conceptual combination (Costello & Keane, 2000, 2001) and argument evaluation (Smith, Shafir & Osherson, 1993) by describing plausibility as something more complex than just feature or proposition overlap. Also, it illustrates

that plausibility is not just resultant from the ability to make inference between events (Black et al., 1986), but rather is concerned with the background knowledge that these inferences require.

Word-coherence is defined by the Knowledge-Fitting Theory in terms of the distributional spotlight; that is, how distributionally distant sentences prime more related terms and knowledge than distributionally close sentences. Again, this account is more specific and far-reaching than others in the literature. Distributional information contributes more to plausibility judgement than just a word similarity metric (Lapata et al., 1999), and plays a greater role than simply speeding sentence parsing (e.g., Pickering & Traxler, 1998; Speer & Clifton, 1998). In addition, the distributional spotlight account provides an explanation for how conceptual and distributional factors can combine to affect the plausibility judgement process.

6.3.2 – Implications for Discourse Comprehension

In the discourse literature, the comprehension of a scenario involves constructing a mental representation of the described situation, aided by cues provided by the linguistic input and using inferences from prior knowledge (e.g., Gernsbacher, 1990; Johnson-Laird, 1983; Kintsch, 1998; Singer, Graesser & Trabasso, 1994; van Dijk & Kintsch, 1983; Zwaan, Kaup, Stanfield & Madden, 2001). The Knowledge-Fitting Theory impacts upon this field from both an empirical and a theoretical perspective.

In our empirical investigations of word-coherence, we have found that the distributional distance between sentences impacts on their comprehension and

plausibility judgement time, independent of other factors. As much experimental work in discourse comprehension centres on online data, this finding has important implications for the design of response time experiments, as the distributional distance between sentences could be a hitherto unconsidered confounding factor. Our findings suggest that online studies should control the distributional properties of experimental materials with the same care as word frequency, syllable count, and other factors commonly recognised as affecting comprehension times.

Our theoretical account of the comprehension process describes how the inferencing process is influenced by the distributional spotlight. This account has implications for the fields of discourse analysis and memory research, as it provides an explanation for how the distributional properties of words activate relevant knowledge from the vast store in long-term memory. While other work has suggested that the linguistic input may help to prime knowledge relevant to the present context (e.g., Burgess, Livesay & Lund, 1998; Kintsch, 2000; Kintsch, Patel & Ericsson, 1999), our account adds an extra level of specificity by emphasising the importance of distributional distance. Discourse comprehension is affected by the distributional distance between sentences, because greater distance primes more knowledge and better facilitates the inferential process.

6.3.3 – Implications for Distributional Analysis

Many simple cognitive linguistic phenomena have been shown to emerge from linguistic distributional knowledge, including priming effects (Lund, Burgess & Atchley, 1995), typicality of category members in and out of context (Connell &

Ramscar, 2001a), and synonym matching (Landauer & Dumais, 1997). In the Knowledge-Fitting Theory, we describe how distributional knowledge can also play an essential role in a more complex cognitive task. Our account proposes how distributional knowledge and other background knowledge can operate in tandem, which impacts upon traditional criticisms of distributional models. For example, distributional knowledge alone may be insufficient to interpret analogies (French & Labiouse, 2002), but it may contribute to the activation of background knowledge relevant to the analogical mapping task (see also Ramscar & Yarlett, 2003).

Finally, the Plausibility Analysis Model (PAM) illustrates how distributional models of language may be enfolded into larger cognitive models. The success with which the two contrasting approaches of objective distributional models and hand-coded reasoning mechanisms can be combined has implications for distributional and cognitive modelling. Such a synergistic paradigm can exploit the strengths of both approaches and produce a more accurate model of human cognitive processing than could be achieved by separate efforts.

6.4 – Conclusions

When people judge the plausibility of a scenario, excuse or idea, they are susceptible to influences from both the word level and concept level. As political speechwriters have long known, the language used to describe an idea can be as important as the idea itself. Given the pervasiveness of plausibility in many cognitive phenomena, the empirical and theoretical work we describe here impacts upon many areas of research. The Knowledge-Fitting Theory of Plausibility can

explain how people make plausibility judgements, and it can explain exactly what affects this judgement process. In short, plausibility now has a clarity of definition that was previously absent from the literature. Existing research that has utilised plausibility judgements – in fields including memory, discourse comprehension, reasoning and conceptual combination – can now be re-examined in a new light. More importantly, however, the Knowledge-Fitting Theory now offers a theoretical touchstone to future research that makes use of plausibility.

BIBLIOGRAPHY

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6*, 451-474.

Anderson, J. R. (1983). *The Architecture of Cognition*. Harvard UP

Anderson, R. C., & Ortony, A. (1975). On putting apples in bottles: A problem of polysemy. *Cognitive Psychology, 7*, 167-180.

Baggett, P. (1975). Memory for explicit and implicit information in picture stories. *Journal of Verbal Learning and Verbal Behavior, 18*, 333-356.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior, 13*, 471-481.

Barsalou, L.W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences, 22*, 577-660.

Black, A., Freeman, P., & Johnson-Laird, P. N. (1986). Plausibility and the comprehension of text. *British Journal of Psychology, 77*, 51-60.

Brown, A. S., Neblett, D. R., Jones, T. C., & Mitchell, D. B. (1991). Transfer appropriate processing in repetition priming: Some inappropriate findings. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 514-525.

Burgess, C. & Lund, K., (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12*, 1-34.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes, 25*, 211-257.

Carpenter, P.A., Just, M.A., Keller, T.A., Eddy, W.F., Thulborn, K.R. (1999) Time course of fMRI-activation in language and spatial networks during sentence comprehension. *NeuroImage, 10*, 216-224.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning, *Cognitive Psychology, 38*, 191-258.

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

Connell, L. (2000). Categories, Concepts and Co-occurrence: Modelling Categorisation Effects with LSA. *M.Sc. Thesis*, School of Cognitive Science, University of Edinburgh, Scotland.

Connell, L., & Keane, M. T. (2002a). 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.* Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M. T. (2002b). The influence of inference type and distributional information on plausibility judgements. *Proceedings of the Nineteenth Annual Conference of the British Psychological Society Cognitive Psychology Section.*

Connell, L., & Keane, M. T. (2003a). PAM: A cognitive model of plausibility. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M. T. (2003b). The knowledge-fitting theory of plausibility. *Proceedings of the Fourteenth Irish Conference on Artificial Intelligence and Cognitive Science.* Dublin, Ireland: Trinity College.

Connell, L., & Keane, M. T. (in press). What plausibly affects plausibility? Concept-coherence & distributional word-coherence as factors influencing plausibility judgements. *Memory and Cognition.*

Connell, L., & Keane, M. T. (in prep.). The knowledge-fitting theory of plausibility: how people judge what is plausible. *Manuscript in preparation.*

Connell, L., & Ramscar, M. (2001a). Using distributional measures to model typicality in categorization. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Connell, L., & Ramscar, M. (2001b). Modeling canonical and contextual typicality using distributional measures. *Proceedings of the Third International Conference on Cognitive Science*.

Cooper, R., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence, 85*, 3-44.

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 and Cognition, 27*, 255-271.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, 30*, 205-247.

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

Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology, 14*, 590-607.

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*. Hillsdale, NJ; Erlbaum.

Friedman, N., & Halpern, J. Y. (1996). Plausibility measures and default reasoning. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 2. Menlo Park, CA: AAAI Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes, 23*, 265-304.

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*, 379-401.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse Comprehension. *Annual Review of Psychology, 48*, 163-189.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Halff, H. M., Ortony, A., & Anderson, R. C. (1976). A context-sensitive representation of word meanings. *Memory and Cognition, 4*, 378-383.

Halldorson, M. & Singer, M. (2002). Inference processes: integrating relevant knowledge and text information. *Discourse Processes, 34*, 145-161.

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

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. (1985). On drawing analogies when solving problems: A theory and test of solution generation in an analogical problem solving task. *British Journal of Psychology, 76*, 449-459.

Keenan, J. M., Baillet, S. D. & Brown, P. (1984). The effect of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior, 23*, 115-126.

Kintsch, W. & van Dijk, T. A. (1978). Towards a model of text comprehension. *Psychological Review, 85*, 363-394.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kintsch, W., (2000). The control of knowledge activation in discourse comprehension. In W. Perrig & A. Grob (Eds.). *Control of human behaviour, mental processes, and consciousness*. Mahweh, NJ: Erlbaum.

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

Kintsch, W., Patel, V. L., & Ericsson, K. A. (1999). The role of Long-Term Working Memory in text comprehension. *Psychologia, 42*, 186–198.

Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*, 35-42.

Koivisto, M., & Revonsuo, A. (2000). Semantic priming by pictures and words in the cerebral hemispheres. *Cognitive Brain Research, 10*, 91-98.

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 Ninth Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco, 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 and Cognition, 23*, 34-48.

Loftus, E. F. (1979) . Reacting to blatantly contradictory information. *Memory and Cognition*, 7, 368-374.

Lund, K., Burgess, C. & Atchley, R. A. (1995) Semantic and associative priming in high-dimensional semantic space. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

McKoon, G. & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*, 82-91.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440-466.

McKoon, G., & Ratcliff, R. (1988) Contextually relevant aspects of meaning. *Journal of Experimental Psychology: Learning, Memory and Cognition 4*, 331-343.

Mellet, E., Tzourio, N., Crivello, F., Joliot, M., Denis, M., & Mazoyer, B. (1996). Functional anatomy of spatial mental imagery generated from verbal instruction. *The Journal of Neuroscience, 16*, 6504-6512.

Meyer, D. E. & Schvaneveldt, R. W. (1976). Meaning, memory structure, and mental processes. *Science, 197*, 27-33.

Myers, J. L., & O'Brien, E. J. (1998) Accessing the discourse representation during reading. *Discourse Processes 26*, 131-157.

Myers, J. L., Shinjo, M. & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Verbal Learning and Verbal Behavior, 26*, 453-465.

Norton, S. W. (1988). An explanation mechanism for Bayesian inferencing systems. In J. F. Lemmer & L. N. Kanal (Eds.). *Uncertainty in Artificial Intelligence 2*. Amsterdam: North-Holland.

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

Pohl, R. (1998). The effects of feedback source and plausibility of hindsight bias. *European Journal of Cognitive Psychology, 10*, 191-212.

Radvansky, G. A., Spieler, D. H., & Zacks, R. T. (1993). Mental model organization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 95-114.l

Ramscar, M. & Yarlett, D. (2003). Semantic grounding in models of analogy: An environmental approach. *Cognitive Science, 27*, 41-71.

Reder, L. M. & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. *Cognitive Psychology, 12*, 447-472.

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., & Anderson, J. R. (1980). A partial resolution of the paradox of interference: The role of integrating knowledge. *Cognitive Psychology, 12*, 447-472.

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

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 and Cognition, 9*, 55-72.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

Redington, M., Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences, 1*, 273-281.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606-621.

Saffran, J. R., Johnson, E. K., Aslin, R. A. & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Singer, M. & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology, 30*, 1-38.

Singer, M. & Kintsch, W. (2001). Text retrieval: A theoretical exploration. *Discourse Processes, 31*, 27-59.

Singer, M., Graesser, A. C., & Trabasso, T. (1994). Minimal or global inferences during reading. *Journal of Memory and Language, 33*, 421-441.

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*, 965-978.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1996). Computation of Bayesian graphical models. *Bayesian Statistics, 5*, 407-425.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*, 153-156.

Stenberg, G., Radeborg, K., & Hedman, L. R. (1994). The picture superiority effect in a cross-modality recognition task. *Memory and Cognition, 23*, 425-441.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)Consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

Tennenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629-640.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thompson, L. A. & Kliegl, R. (1991). Adult age effects of plausibility on memory: The role of time constraints during encoding. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 542-555.

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

van Dijk, T. A., & Kintsch, W. (1983) *Strategies of Discourse Comprehension*. Academic Press.

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 1196-1207.

Zwaan, R. A. & Radvansky, G. A. (1998) Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Zwaan, R. A., Kaup, B., Stanfield, R. A., & Madden C. J. (2001). Language comprehension as guided experience. *Retrieved January 20, 2003, from http://cogprints.ecs.soton.ac.uk/archive/00000949/00/lc_as_guided-exp*

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.

Zwaan, R. A., Stanfield, R. A. & Yaxley, R. H. (2002). Language Comprehenders mentally represent the shapes of objects. *Psychological Science, 13*, 168-171.

APPENDIX A – MATERIALS FOR EXPERIMENT 1

This appendix shows the materials used in Experiment 1, including the LSA scores between each Sentence 2 and its corresponding Sentence 1. All LSA comparisons were performed on the LSA website at http://lsa.colorado.edu using the General Reading up to 1st Year College semantic space, with document-to-document comparison at maximum factors. The materials take 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 (0.93);	The thumb bled (0.13)
Attributal:	The boy was ugly (0.92);	The thumb was ugly (0.17)
Temporal:	The boy twitched (0.94);	The thumb twitched (0.17)
Unrelated:	The boy appeared (0.92);	The thumb appeared (0.10)

The cat pounced on the bird.

Causal:	The cat gripped (0.74);	The claws gripped (0.43)
Attributal:	The cat was scary (0.74);	The claws were scary (0.40)
Temporal:	The cat relaxed (0.74);	The claws relaxed (0.45)
Unrelated:	The cat hovered (0.75);	The claws hovered (0.51)

The woman swiped at his face.

Causal:	The woman missed (0.78);	The hand missed (0.15)
Attributal:	The woman was petite (0.79);	The hand was petite (0.19)
Temporal:	The woman waved (0.80);	The hand waved (0.19)
Unrelated:	The woman vanished (0.79);	The hand vanished (0.17)

The bottle fell off the shelf.

Causal:	The bottle smashed (0.60);	The glass smashed (0.22)
Attributal:	The bottle was pretty (0.47);	The glass was pretty (0.18)
Temporal:	The bottle sparkled (0.58);	The glass sparkled (0.20)
Unrelated:	The bottle melted (0.50);	The glass melted (0.23)

The dress snagged on a nail.

Causal:	The dress ripped (0.79);	The silk ripped (0.26)
Attributal:	The dress was costly (0.73);	The silk was costly (0.26)
Temporal:	The dress glittered (0.79);	The silk glittered (0.29)
Unrelated:	The dress shrank (0.81);	The silk shrank (0.28)

The knife caught on the fork.

Causal:	The knife snapped (0.73);	The blade snapped (0.47)
Attributal:	The knife was sharp (0.67);	The blade was sharp (0.44)
Temporal:	The knife gleamed (0.73);	The blade gleamed (0.40)
Unrelated:	The knife bubbled (0.74);	The blade bubbled (0.40)

The girl shook the box.

Causal:	The box rattled (0.81);	The lid rattled (0.47)
Attributal:	The box was wooden (0.78);	The lid was wooden (0.38)
Temporal:	The box gleamed (0.81);	The lid gleamed (0.44)
Unrelated:	The box floated (0.81);	The lid floated (0.40)

The girl hit the mirror.

Causal:	The mirror cracked (0.67);	The glass cracked (0.25)
Attributal:	The mirror was huge (0.61);	The glass was huge (0.23)
Temporal:	The mirror shone (0.65);	The glass shone (0.24)
Unrelated:	The mirror bubbled (0.66);	The glass bubbled (0.24)

The waitress dropped the cup.

Causal:	The cup smashed (0.80);	The handle smashed (0.21)
Attributal:	The cup was delicate (0.70);	The handle was delicate (0.19)
Temporal:	The cup glistened (0.78);	The handle glistened (0.21)
Unrelated:	The cup floated (0.77);	The handle floated (0.19)

The lightning struck the tree.

Causal:	The tree fell (0.95);	The branch fell (0.57)
Attributal:	The tree was huge (0.93);	The branch was huge (0.46)
Temporal:	The tree grew (0.91);	The branch grew (0.45)
Unrelated:	The tree melted (0.92);	The branch melted (0.50)

The breeze hit the candle.

Causal:	The candle flickered (0.43);	The flame flickered (0.26)
Attributal:	The candle was pretty (0.35);	The flame was pretty (0.25)
Temporal:	The candle shone (0.43);	The flame shone (0.29)
Unrelated:	The candle drowned (0.44);	The flame drowned (0.27)

The lever closed the cage.

Causal:	The cage rattled (0.81);	The bars rattled (0.47)
Attributal:	The cage was rusty (0.78);	The bars were rusty (0.38)
Temporal:	The cage tilted (0.81);	The bars tilted (0.44)
Unrelated:	The cage crumbled (0.81);	The bars crumbled (0.40)

APPENDIX B – MATERIALS FOR EXPERIMENTS 2-5

This appendix shows the materials used in Experiments 2-5, including the LSA scores between each Sentence 2 and its corresponding Sentence 1. All LSA comparisons were performed on the LSA website at http://lsa.colorado.edu using the General Reading up to 1st Year College semantic space, with document-to-document comparison at maximum factors. The materials take the following format:

Sentence 1.

Causal:	Sentence 2.	(*Close* LSA score)
	Sentence 2.	(*Distant* LSA score)
Attributal:	Sentence 2.	(*Close* LSA score)
	Sentence 2.	(*Distant* LSA score)

The opposition scored a penalty.

Causal:	The goalie wept.	(*Close* 0.29)
	The goalie cried.	(*Distant* 0.04)
Attributal:	The goalie was sluggish.	(*Close* 0.29)
	The goalie was slow.	(*Distant* 0.13)

The cat pounced on the bird.

Causal:	The claws tore.	(*Close* 0.46)
	The claws cut.	(*Distant* 0.04)
Attributal:	The claws were sharp.	(*Close* 0.36)
	The claws were pointy.	(*Distant* 0.27)

The woman swiped at his face.

Causal:	The hand slapped.	(*Close* 0.18)
	The hand hit.	(*Distant* 0.11)
Attributal:	The hand was petite.	(*Close* 0.19)
	The hand was little.	(*Distant* 0.13)

The pack saw the fox.

Causal:	The hounds growled.	(*Close* 0.37)
	The hounds snarled.	(*Distant* 0.20)
Attributal:	The hounds were fierce.	(*Close* 0.19)
	The hounds were vicious.	(*Distant* 0.12)

The flowers wilted in the vase.

Causal:	The petals dropped.	(*Close* 0.63)
	The petals fell.	(*Distant* 0.53)
Attributal:	The petals were velvety.	(*Close* 0.70)
	The petals were soft.	(*Distant* 0.53)

The dress snagged on a nail.

Causal:	The satin ripped.	(*Close* 0.29)
	The satin tore.	(*Distant* 0.17)
Attributal:	The satin was priceless.	(*Close* 0.26)
	The satin was valuable.	(*Distant* 0.13)

The knife caught on the fork.

Causal:	The blade bent.	(*Close* 0.44)
	The blade curved.	(*Distant* 0.31)
Attributal:	The blade was broad.	(*Close* 0.37)
	The blade was wide.	(*Distant* 0.28)

The sail caught the wind.

Causal:	The canvas flapped.	(*Close* 0.36)
	The canvas fluttered.	(*Distant* 0.44)
Attributal:	The canvas was strong.	(*Close* 0.28)
	The canvas was durable.	(*Distant* 0.18)

The girl shook the box.

Causal:	The lid hopped.	(*Close* 0.49)
	The lid jumped.	(*Distant* 0.32)
Attributal:	The lid was flimsy.	(*Close* 0.43)
	The lid was weak.	(*Distant* 0.22)

The girl hit the mirror.

Causal:	The reflection quivered.	(*Close* 0.52)
	The reflection shook.	(*Distant* 0.41)
Attributal:	The reflection was indistinct.	(*Close* 0.50)
	The reflection was faint.	(*Distant* 0.40)

The wolf raced towards the flock.

Causal:	The sheep ran.	(*Close* 0.52)
	The sheep fled.	(*Distant* 0.45)
Attributal:	The sheep were uneasy.	(*Close* 0.45)
	The sheep were nervous.	(*Distant* 0.34)

The lightning struck the tree.

Causal:	The branch scorched.	(*Close* 0.53)
	The branch burned.	(*Distant* 0.44)
Attributal:	The branch was huge.	(*Close* 0.46)
	The branch was big.	(*Distant* 0.32)

The breeze hit the candle.

Causal:	The flame flared.	(*Close* 0.26)
	The flame grew.	(*Distant* 0.17)
Attributal:	The flame was hot.	(*Close* 0.18)
	The flame was warm.	(*Distant* 0.08)

The lever shut the cage.

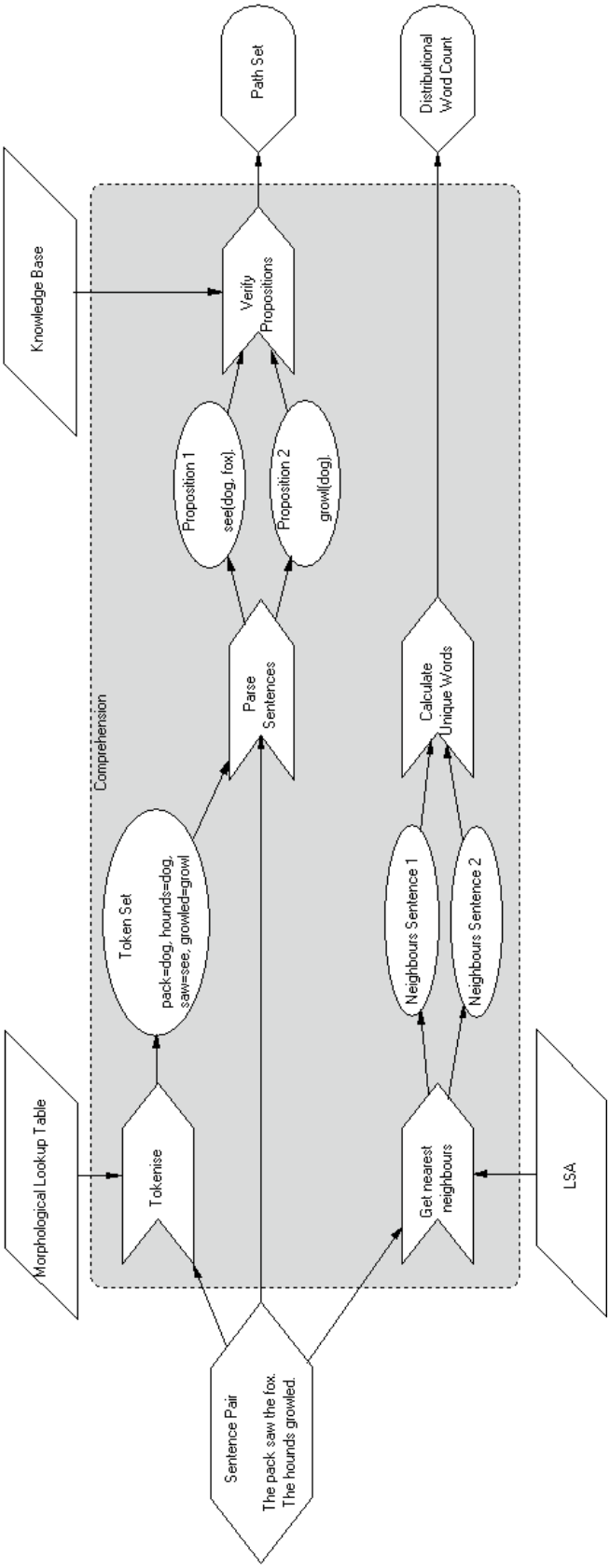
Causal:	The bars rang.	(*Close* 0.34)
	The bars resonated.	(*Distant* 0.25)
Attributal:	The bars were rigid.	(*Close* 0.17)
	The bars were solid.	(*Distant* 0.04)

The wave crashed against the ship.

Causal:	The vessel keeled.	(*Close* 0.54)
	The vessel tilted.	(*Distant* 0.39)
Attributal:	The vessel was antique.	(*Close* 0.48)
	The vessel was ancient.	(*Distant* 0.24)

APPENDIX C – PAM PROCESS DIAGRAMS

Process Diagram of PAM's Comprehension Stage



Process Diagram of PAM's Assessment Stage