Comprehension and Production in Conceptual Combination

Dermot Lynott
B.Sc. (Hons), M.Sc.

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

Department of Computer Science
Head of Department : Mr. Greg O’Hare
Supervisor of Research : Prof. Mark Keane
# Table of Contents

1. **INTRODUCTION** .................................................................................................................. 1
   1.1. **BACKGROUND AND OBJECTIVE OF RESEARCH** ......................................................... 1
   1.2. **CONCEPTUAL COMBINATION IN ISOLATION** ................................................................. 2
   1.3. **CONCEPTUAL COMBINATION IN CONTEXT** ................................................................. 4
   1.4. **COMPONENT PRODUCTION IN CONCEPTUAL COMBINATION** ................................. 4
   1.5. **OUTLINE** .................................................................................................................... 5

2. **REVIEW** ............................................................................................................................ 6
   2.1. **INTRODUCTION** ............................................................................................................ 6
   2.2. **THEORIES OF CONCEPTUAL COMBINATION** ............................................................. 7
   2.3. **CONCEPTUAL COMBINATION IN CONTEXT** ................................................................. 42
   2.4. **THEORIES OF NOVEL COMPONENT PRODUCTION** ..................................................... 52
   2.5. **REVIEW SUMMARY** .................................................................................................. 58

3. **COMPREHENSION IN CONCEPTUAL COMBINATION** .................. 60
   3.1. **INTRODUCTION** ........................................................................................................... 60
   3.2. **CONTEXT** .................................................................................................................... 61
   3.3. **EXPERIMENT 1 – OUT-OF-CONTEXT INTERPRETATIONS** ........................................... 65
   3.4. **EXPERIMENT 2 – ONLINE OUT-OF-CONTEXT INTERPRETATIONS** ............................. 69
   3.5. **EXPERIMENT 3 – OUT-OF-CONTEXT INFLUENCES ON IN-CONTEXT PROCESSING** ...... 73
   3.5. **EXPERIMENTS 4A AND 4B – BIAS POST-TESTS** ......................................................... 85
   3.7. **CONCLUSIONS** .......................................................................................................... 89

4. **COMPONENT PRODUCTION IN CONCEPTUAL COMBINATION** 91
   4.1. **INTRODUCTION** ........................................................................................................... 91
   4.2. **PLAUSIBILITY IN COMPONENT PRODUCTION** ............................................................. 92
   4.3. **A PARADIGM FOR COMPONENT PRODUCTION** .......................................................... 93
4.4. EXPERIMENT 5 – ESTABLISHING DIFFERENT LEVELS OF PLASIBILITY..........................97
4.5. EXPERIMENT 6 – PRODUCING NOVEL COMPOUNDS 1 ..............................................100
4.6. EXPERIMENT 7 – PRODUCING NOVEL COMPOUNDS 2 ..............................................105
4.7. OBSERVATIONS ON COMPOUND PRODUCTION IN EXPERIMENTS 6 AND 7 ..........108
4.8. EXPERIMENT 8 – COMPOUND PRODUCTION FROM INTERPRETATIONS..................111
4.9. CONCLUSION.............................................................................................................................116

5. THEORY AND MODEL ............................................................................................. 118

5.1. OUTLINE..............................................................................................................................118
5.2. IPAC THEORY OF CONCEPTUAL COMBINATION ........................................................119
5.3. THE PUNC MODEL............................................................................................................139
5.4. TESTING THE MODEL.........................................................................................................164
5.5. COMPARING PUNC TO OTHER MODELS..........................................................................179
5.6. CONCLUSION.............................................................................................................................184

6. CONCLUSION ......................................................................................................................... 185

6.1. SUMMARY OF ACCOMPLISHMENTS..................................................................................185
6.2. FURTHER RESEARCH...........................................................................................................189
6.3. CONCLUSIONS...........................................................................................................................193

APPENDIX A – EXPERIMENT 1 MATERIALS ............................................................... 195

APPENDIX B – EXPERIMENT 2 MATERIALS................................................................. 196

APPENDIX C – EXPERIMENT 3 & 4 MATERIALS ......................................................... 198

APPENDIX D – EXPERIMENT 5, 6 & 7 MATERIALS .................................................... 205

BIBLIOGRAPHY......................................................................................................................... 209
Summary

Everyday people produce and comprehend novel noun-noun compounds. These concept combinations play a fundamental role in the creativity and generativity of language. Not surprisingly, the importance of this phenomenon has lead to a great deal of interest, with recent years seeing the emergence of several competing views of conceptual combination. These views have focussed, almost exclusively, on the comprehension of these novel combinations in isolation, without any consideration for the effects of surrounding context or the relationship between the production and comprehension of these compounds.

This thesis bridges this gap by taking an inclusive view of conceptual combination, in all its facets. The thesis focuses on the three pillars of conceptual combination: the comprehension of novel compounds in isolation, their comprehension in context and finally, the production of novel compounds. A series of experiments examines each of these pillars, providing a more comprehensive view of conceptual combination that has hitherto been absent from cognitive science. The Integrated Production and Comprehension (IPAC) theory draws from these results and existing research to specify a core set of factors that contribute to both comprehension and production. The IPAC theory is implemented computationally as the PUNC (Producing and Understanding Novel Compounds) model. In simulations, PUNC’s performance closely reflects human responses in a variety of tasks.
Acknowledgements

There are a host of people who deserve my thanks for their help and encouragement over the last few years.

I would like to thank Prof. Mark Keane, my supervisor, for his time, encouragement, insights and advice, which he gave selflessly. Without his guidance, this thesis could not have been written.

A resounding ‘Thank You’ to Louise Connell, for everything. Insanity would have crept in a lot sooner had it not been for her support, constructive comments, patience and other attributes too numerous to mention, all of which were available in abundance.

Thanks to the UCD-TCD Cognitive Science group for helpful comments throughout the course of this work.

Thanks also to all the postgrads of the Computer Science department in UCD, who often made my life a lot easier with their help and support.

And lastly, to my family for all of their never-ending support and encouragement down through the years, even when I didn’t 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.
1. **INTRODUCTION**

“The term *soccer mom* is thrown about all too much these days”,

- Krusty the Clown, The Simpsons (Episode 8F20).

1.1. **Background and Objective of Research**

How do we understand things we’ve never heard before? How do we describe things we’ve never encountered? Combining existing concepts to create new ones is both a means of extending the language and of encapsulating more complex ideas in simple word combinations. This thesis is designed to provide a wider understanding of conceptual combination than has previously been offered. The thesis is structured around three central pillars; the first representing conceptual combination comprehension in isolation, the second comprehension in context, the third the production of novel compounds.

The objective of this thesis is to build a broader picture around these pillars. We propose a theory of conceptual combination, Integrated Production and Comprehension (IPAC), that considers conceptual combination in a more inclusive manner than has previously been done. Additionally, we describe a computational implementation of this theory, PUNC (Producing and Understanding Novel Combinations), and show how the model parallels many aspects of human performance in conceptual combination tasks. In the following section, we introduce
each of the three pillars, which focus on the treatment of conceptual combination in its various forms.

1.2. Conceptual combination in Isolation

Such compounds are a ubiquitous feature of the language. People encounter newly combined concepts in a range of circumstances; from everyday conversations to books, newspaper articles, movies and TV shows. *Truck doctor* (a person who fixes trucks), *guitar figure* (referring to someone’s curvy physique) and *yawn factory* (a boring place) are all examples of conceptual combination, which occur with great frequency in a variety of situations. Christopher Isherwood’s “Down there on a visit” (1968) contains almost 200 unique compounds (Leonard, 1984). In 500 sentences of Newsweek and the New York Times, a previously unseen noun compound is encountered on average in every second sentence (McDonald, 1982). More contemporary sources like The Simpsons cartoons use novel compounds in almost every episode (e.g., *floor pie*, *trash cookies* and *imagination Christmas*). Although, these compounds may appear on the surface to be quite complex, their usage is not restricted to adult language usage. Children as young as two are able to understand novel word combinations in isolation and by the age of 6 are able to produce them without grammatical errors (Clark & Barron, 1988; Clark & Berman, 1987). It is clear that these combinations play a central role, not only in the extension of the language by adult users, but also in the course of child language development.

When we talk about novel combinations (e.g., *guitar figure*), the constituents are generally referred to as the head (i.e., *figure*; the second element) and modifier concepts (i.e., *guitar*; the first element). Any modifier-head combination can give rise to a multitude of possible meanings and many different types of interpretation. Take
for example, the compound *whale seal* – there are several different ways in which it can be interpreted. One possible meaning might be “a very large seal”. This kind of interpretation is called a property interpretation, where a property of the modifier (i.e., the property of “large” from whale) is transferred to the head concept to create an interpretation. Another possible meaning might be “a seal that eats whales”. This interpretation is a relational interpretation, as the relation “eats” links the modifier and head concepts. Although these two types of interpretation describe the vast majority of interpretations that people tend to produce for novel combinations (Costello & Keane, 2001; Wisniewski & Gentner), there are other types that occur less often. These include conjunctions (e.g., a *pet fish* can be both a pet and a fish), reversals (e.g., a *chair ladder* could be a chair that is used as a ladder) – where the referent is not a member of the head concept’s category - and known concept versions (e.g., a *pencil bed* could refer to a pencil case). Thus, the types of interpretations people generate generally fall into the following categories:

- Property-based (e.g., a *whale seal* is a large seal)
- Relation-based (e.g., a *whale seal* is a whale that eats seals)
- Conjunctions (e.g., a *pet fish* is both a pet and a fish)
- Reversals (e.g., a *chair ladder* is a chair being used as a ladder)
- Known concepts (e.g., a *pencil bed* is a pencil case)

Despite the inherent polysemy of novel compounds and the variety of potential interpretation types that could possibly be given, people display great efficiency and consistency in their ease comprehension.
1.3. Conceptual Combination in Context

Traditionally, conceptual combination research has focussed on people’s interpretations of novel compounds in isolation from any contextual effects. For example, people would be presented with a compound on a screen and asked to press a key when they had thought of a meaning. The implicit assumption of this research is that the processes people utilise and interpretations they produce out-of-context, will still have a meaningful role for in-context comprehension (see e.g., Wisniewski, 1996). However, there has been little direct empirical analysis of this belief. Consequently, there is no clear understanding of the relationship between how people interpret compounds out-of-context and how they interpret them in-context. Is it the case that if an interpretation is more easily processed out-of-context that it will still be processed quickly in-context? Will different contextual environments have different effects? These questions form the basis of our empirical investigation into conceptual combination in and out of context in Chapter 3.

1.4. Compound Production in Conceptual Combination

The lack of empirical studies into context effects is not the only imbalance facing conceptual combination research. While much of the literature has examined the comprehension side of conceptual combination, little has been said on the creation of novel word combinations. Work by Eve Clark and colleagues (e.g., Clark, 1982; Clark & Hecht, 1982; Clark, Hecht & Mulford, 1986) has provided some insights into child compound production in terms of their language development, but there is little known about the factors that influence adult speakers in the production of such compounds. In Chapter 4, we develop a novel paradigm to illicit novel compounds from people. We examine the compounds people produce in response to different
types of description, specifically with reference to the plausibility of these
descriptions. We discuss how plausibility impacts the people’s production of
compounds and on the perceived acceptability of those compounds.

1.5. Outline

From our empirical investigations in Chapters 3 and 4, we draw together our
findings and develop a theory of compound comprehension and production anchored
in the high-level principles of Costello and Keane’s Constraint Theory of Conceptual
Combination (2000, 2001). In Chapter 5, we provide a computational level account
of the comprehension and production sides of conceptual combination (The IPAC
theory – Integrated Production and Comprehension) before outlining an algorithmic
instantiation of this theory, the PUNC model (Producing and Understanding Novel
Combinations). Our model shows how common representations and factors can be
used in modelling empirical phenomena from both sides of conceptual combination.
Then, using PUNC, we perform several simulations which reflect the findings of our
previous psychological experiments. Finally we review the thesis as a whole and
consider relevant areas for future research in conceptual combination and related
areas.
2. Review

2.1. Introduction

Languages change. The emergence of new words and phrases constantly alters the content of language. While some words are created neologisms (i.e., newly coined words), many new entrants to the language are formed by combining concepts we already know. Notebook computers, junk bonds and soccer moms are all examples of such phrases. Canon (1987) has said that over 55% of new terms in the English language are compounds of existing words. This process of conceptual combination is a fundamental form of language generativity; a process that somehow combines the existing meanings of words to produce new meanings for compound terms. People encounter such novel compounds in conversation, newspapers, and television programmes, where they are used to fill a variety of roles from temporary labels (e.g., the coffee chair is the chair with coffee spilt on it), to lexical shortcuts (e.g., vermin boys referring to men working for the pest control company), to alleviating tip-of-the-tongue states (e.g., saying truck doctor instead of mechanic). Since these occurrences of novel combinations are so frequent, it is not surprising that this phenomenon has received a great deal of attention.

In this chapter we look at the three pillars of conceptual combination; theories of comprehension in isolation, comprehension in context and production. We first examine theories of comprehension and weigh up the available empirical support for
each. We look at the question of context and consider whether existing theories can contribute to an account of conceptual combination in context. We then examine the existing work on compound production, an area that has received less attention in the literature.

2.2. Theories of Conceptual Combination

In recent years, four theories of conceptual combination in isolation have come to the fore, namely Concept Specialisation (Cohen & Murphy, 1984; Murphy, 1988), Dual-Process Theory (Wisniewski, 1997a, 1997b, 1998), Constraint Theory (Costello & Keane, 2000, 2001), and the CARIN theory (Gagné, 2000, 2001, 2001; Gagné & Shoben, 1997). These theories have superseded earlier attempts such as Selective Modification (Smith & Osherson, 1984; Smith, Osherson, Rips & Keane, 1988) and Fuzzy Set Theory (Zadeh, 1965, 1982), which lacked sufficient generality to account for many cognitive phenomena (Dunbar & Myers, 1988). For example, the Selective Modification approach was overly simplistic, focussing primarily on adjective-noun combinations, while Fuzzy Set Theory was not considered a psychological theory of conceptual combination (Murphy, 1988).

More recent theories provide a clear development of ideas, with each theory attempting to capture aspects of conceptual combination in a more parsimonious fashion than its predecessor. For example, Murphy’s theory of Concept Specialisation made an important leap towards dealing with “complex concepts” (e.g., noun-noun compounds) compared to previous theories that only examined simpler constructs such as adjective-noun combinations (see Smith et al, 1984). In turn, Wisniewski’s Dual-Process Theory may be considered a successor to the Concept Specialisation theory, with Wisniewski providing certain extensions that offer greater coverage of
empirical phenomena. Dual-Process Theory describes a holistic view of conceptual combination, and outlines several psychological processes that allow people to successfully interpret novel word combinations. In contrast, Costello and Keane’s Constraint Theory describes a unitary mechanism of conceptual combination, and is based on a set of core constraints that deal with the creativity and efficiency of the conceptual combination process. Finally, Gagné’s CARIN theory provides a mathematical model of conceptual combination that uses our prior experience of words in compounds to predict what interpretations people will produce, and what compounds people will find easiest to understand.

While each theory offers an insight into the problem of describing how people understand combinations of words they have never heard before, they each carry their own baggage and have many issues that have yet to be addressed. In the following sections, we describe each theory, evaluate the evidence for and against them, and discuss why each theory is found wanting in some way.

2.2.1. Concept Specialisation

Concept Specialisation was developed by Cohen and Murphy (Cohen & Murphy, 1984; Murphy, 1988) to explain how people interpreted noun-noun compounds, as opposed to less complex combinations (e.g., adjective-noun combinations). The theory describes two stages in the interpretation process; the first is a slot-filling mechanism where the modifier is inserted into the head concept to form an interpretation, and the second is an elaboration mechanism that uses world knowledge in order to expand these interpretations.
Concept Specialisation views compounds as being asymmetrical. The first concept of a compound acts as a modifier on the second, head concept. Thus, the compound *apartment dog* is not the same as the compound *dog apartment* – one is a type of dog, while the other is a type of apartment. To interpret such combinations, each concept must be represented in some way - Concept Specialisation assumes each concept is represented as a schema (see Rumelhart, 1980). Schemas contain a series of slots and fillers, which provide a description and the internal structure of a concept.

Table 2.1 shows an example schema for the concept *dog*. Examples of slots are “Habitat” and “Functions”, which provide general information about the concept, with fillers being specific values that occupy these slots (e.g., “house” and “sleeps”). To interpret the compound *apartment dog*, the modifying concept is used to fill some slot in the head concept. In this case, *apartment* would be classified as a type of habitat and so would fill the “Habitat” slot in the head concept. This provides the interpretation of “a dog that lives in an apartment”.

Table 2.1 A simplified schema for the concept "dog" (adapted from Murphy, 1988).

<table>
<thead>
<tr>
<th>NAME:</th>
<th>“dog”</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODY PARTS:</td>
<td></td>
</tr>
<tr>
<td>LEGS:</td>
<td>4, 3</td>
</tr>
<tr>
<td>HEAD:</td>
<td>1</td>
</tr>
<tr>
<td>HAIR:</td>
<td></td>
</tr>
<tr>
<td>EYES:</td>
<td>2</td>
</tr>
<tr>
<td>COLOUR:</td>
<td>Brown, black etc.</td>
</tr>
<tr>
<td>HABITAT:</td>
<td>Home, streets</td>
</tr>
<tr>
<td>FUNCTIONS:</td>
<td>Best friend, guard home</td>
</tr>
<tr>
<td>BEHAVIOURS:</td>
<td>Bark, bite, eat, sleep, chase cats</td>
</tr>
</tbody>
</table>
From this interpretation, Murphy (1988; see also Cohen & Murphy, 1984) describes a process of elaboration that is used to incorporate world knowledge in order to “clean up” interpretations. This process seeks to make an interpretation more coherent and complete by retrieving information from prior knowledge that is relevant to the interpretation. For example, people might consider that an *apartment dog* is smaller and quieter than other dogs. Thus, the elaboration process allows people to use their knowledge of the world to refine and augment interpretations. This knowledge can refer to both specific instances of concepts (e.g., a specific dog that one may have encountered) and also domain theory knowledge (i.e., a general understanding of particular domains). For example, the knowledge that empty stores lose money might stem from a general theory of buying and selling, rather than by reference to specific shops. In this way, features that were not part of the original concepts can become part of the final interpretation. Indeed, these features are considered to be emergent in the interpretation process.

Thus, Cohen and Murphy have provided a mechanism to explain how complex concepts can be combined and interpreted. Specifically, Concept Specialisation accounts for how people form interpretations using a mediating relation between concepts (e.g., “a dog that lives in an apartment”) and how an interpretation can take on additional properties by referencing general knowledge and specific examples (e.g., an *empty shop* is more likely to lose money). Below, we consider the strengths of the Concept Specialisation account and also outline some possible weaknesses, taking into account the available empirical evidence.

**Evidence for Concept Specialisation**

Concept Specialisation represents an important step forward in understanding conceptual combination, as previous theories (e.g., Selective Modification) had failed
to account for such phenomena as outlined previously. In addition, Murphy (1990) has reported empirical evidence in support of the theory. He has found that complex concepts where the modifier targets a specific slot of the head concept (e.g., adjective-noun combinations such as *flexible pipe*), are easier to understand than those where no specific slot is identified (e.g., noun-noun combinations like *fish pipe*). This is evidenced by the fact that people are significantly slower at interpreting the latter type of combination (Murphy, 1990).

According to Concept Specialisation, the elaboration mechanism allows new features to emerge and form part of the final interpretation. Medin and Shoben (1988) have shown that people often incorporate emergent features when there is a correlation between the attributes of specific instances. For example, *wooden spoon* is considered a better member of the category “Large Spoon” than of the category “Small Spoon”, because the “large” feature often occurs with the “wooden” feature in specific instances of spoons. This attribute correlation is exploited by Concept Specialisation during its elaboration process.

**Shortcomings of Concept Specialisation**

Although Concept Specialisation provides a description of how complex concepts can be interpreted, it has some limitations. While Concept Specialisation discusses the importance of world knowledge in interpreting novel combinations, the means by which this is achieved is fairly vague. As it stands, Concept Specialisation allows for unconstrained activation of world knowledge. This means that any potential advantage from using world knowledge to aid the interpretation process is minimised by the computational cost of filtering irrelevant information.

A related criticism can be applied to the type of knowledge Concept Specialisation uses to construct interpretations. The knowledge encoded in the slots
and fillers is very general in nature and does not allow for certain aspects of a concept to be more important than any other. For example, in the concept dog (see Table 2.1), the information that a dog has two eyes and that a dog barks are equally important, even though most people would probably consider barking to be more relevant to the concept of dog. This means that the knowledge is not prioritised in any way, and Concept Specialisation cannot predict the order in which people will produce interpretations, or why some interpretations are considered more acceptable than others.

Despite the uncertainty of how general knowledge is incorporated and activated in the interpretation process, the biggest criticism of Concept Specialisation lies in the limited types of interpretation it can account for. Concept Specialisation produces only interpretations where the head and modifier concepts are linked by some relation and ignores the possibility of property-based ones. For example, Concept Specialisation generates interpretations by filling slots in the head concept with the modifier name e.g., *robin hawk* could be interpreted as “a hawk that preys on robins”, by filling the PREYS slot in the schema representation of hawk with the modifier name. However, it does not allow for properties of the modifier to be instantiated into the head representation. This means that an interpretation such as “a hawk with a red breast” cannot be produced by this model. As property interpretations account for up to 50% of interpretations (see Costello & Keane, 2000; Wisniewski & Gentner, 1993), this may be considered Concept Specialisation’s primary failing. However, the subsequently developed Dual-Process Theory, which can be viewed as a successor to Concept Specialisation, describes additional processes to account for the production of different types of interpretation. In the following section, we look at Dual-Process theory in more detail.
2.2.2. Dual Process Theory

Dual-process Theory (Wisniewski, 1997a, 1997b; Wisniewski & Love, 1998) proposes two primary mechanisms to describe comprehension in conceptual combination, namely Scenario Creation and Alignment. Scenario Creation is essentially the same as the previously described Concept Specialisation view, where a linking relation between the head and modifier concepts is used to create an interpretation. Alignment (also known as Alignment and Comparison) describes a process by which people can arrive at property-based and hybrid (e.g., pet bird where the entity is both a pet and a bird) interpretations. Briefly stated, the Dual-Process Theory can be seen as a combination of the Concept Specialisation and Alignment views of conceptual combination. The Alignment view holds that concepts must be structurally aligned in order to construct interpretations. In fact, this idea is based on a broader theory of structural alignment, which was developed primarily by Gentner (1983) as Structure Mapping Theory, and is also implemented as a computational model, the Structure Mapping Engine (Falkenhainer, Forbus & Gentner, 1989). Evidence has been provided for the mechanisms of structural alignment in analogy (Gentner & Clement, 1991; Gentner & Markman, 1997; Gentner & Toupin, 1986), metaphor comprehension (Gentner & Wolff, 1997) and learning (Kurtz, Miao & Gentner, 2001). While the title “Dual Process” implies that only the processes of Scenario Creation and Alignment are needed to describe how conceptual combination takes place, many sub-processes are recruited to fully explain various phenomena, which we discuss in more detail below.

Scenario Creation

Scenario Creation is described as the process that gives rise to relation-based interpretations of novel compounds. According to Wisniewski, people create a
scenario in which the head and modifier concepts can be bound to particular roles, thus giving a plausible interpretation. For example, the compound truck soap (Wisniewski, 1997a) could be interpreted as “soap for cleaning a truck”. In this case, “truck” fills the recipient roll of cleaning (i.e., the thing being cleaned), while “soap” fills the instrument role (i.e., the thing being used to clean).

These relational interpretations are generated in a similar fashion to Murphy’s Concept Specialisation (Murphy, 1988), where a slot of the head concept is filled using the modifier concept. For the compound house parrot, filling the “lives-in” slot of “parrot” with the modifier “house” would produce the interpretation “a parrot that lives in a house”. Costello and Keane (2000) use the example compound of horse knife to elucidate the scenario creation process. The concept “knife” is associated with “cutting” scenarios. Such a scenario will consist of various roles (e.g., agent, object and instrument) where something will perform the cutting, something is used to do the cutting and something is actually cut. Using this scenario, the interpretation “a knife used for butchering horses” can be created. In this way, a modifying concept will fill some role of the head concept as long as certain preconditions are met. What these preconditions might be is not spelled out; Wisniewski (1997a) states that “the exact details of such a role assignment process are not well understood, and they remain an issue for future research”. Thus, the specifics of Scenario Creation are left open for debate, which contrasts with the more tightly specified process of Alignment.

Alignment

The processes of Alignment and Comparison are used to construct property-based and hybrid interpretations (see Chapter 1 for definitions of interpretation types). Alignment can be seen as a process of structural alignment between the two concepts of a novel compound. Interpretations that would typically emerge from the alignment
process would include *robin snake* as “a snake with a red breast” and *robin snake* as “a cross between a snake and a robin”. An example of a hybrid interpretation for the compound *painter musician* might be “a painter who is also a musician”. Wisniewski (1997b) points out that these hybridisations are far rarer than property-based interpretations, so here we focus primarily on property-based interpretations.

Table 2.2 Illustrating commonalities, alignable and non-alignable differences between two concepts (adapted from Keane & Costello, 2001)

<table>
<thead>
<tr>
<th>Commonalities</th>
<th>Elephant</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class:</td>
<td>Living thing</td>
<td>Living thing</td>
</tr>
<tr>
<td>Size:</td>
<td>Big</td>
<td>Small</td>
</tr>
<tr>
<td>Colour:</td>
<td>Grey</td>
<td>Silver</td>
</tr>
<tr>
<td>Head:</td>
<td>Has Trunk</td>
<td></td>
</tr>
<tr>
<td>Body:</td>
<td></td>
<td>Has Fins</td>
</tr>
</tbody>
</table>

At its simplest, Dual-Process theory proposes that a property interpretation requires finding an alignable difference between the two constituents of the compound and transferring that property to the head concept. So what makes an alignable difference different from any other difference? And how are these differences identified? The two stages in this process are alignment and comparison. The two concepts are first aligned using commonalities in their relational structure (e.g., matched slots such as habitat, diet, function etc.). Once aligned, people can then compare relational and property commonalities between the concepts. From this
comparison process, two types of information can be extracted: alignable differences and non-alignable differences. Alignable differences are those differences that are linked to the common relational structure of the two concepts (Markman & Gentner, 1993; Markman & Wisniewski, 1997), while non-alignable differences are those differences that are not linked to the common structure. Table 2.2 shows the two concepts elephant and fish, and how they might be aligned in memory. For example, aligning their representations might reveal that they are both living things, both have eyes etc. Having made this alignment, the differences between the concepts are made apparent (e.g., the size of the elephant relative to the fish, the absence of a trunk on the fish). In determining the differences between the concepts, the Alignment procedure specifies which properties can be transferred from the modifier to the head. In other words, alignable differences will be used to map information from the modifier concept to the head. In the compound elephant fish, an alignable difference might be the property “big” as both concepts could have slots relating to size, with elephant having “big” as the slot value, and fish having the value of “small”. The value from the modifier would then be transferred to the head concept to give the interpretation of “an elephant fish is a big fish”.

In addition to the processes of Alignment and Comparison, Dual-Process specifies additional processes to explain other empirical phenomena; namely Construction and Construal. These processes are closely related to the elaboration procedures in Concept Specialisation and in Scenario Creation, except that they are applied to property-based interpretations rather than relation-based ones.

The Construction mechanism involves using world knowledge in order to guide property-based interpretations, within certain constraints. The motivation for the Construction process stems from the idea that when a property is transferred from
a modifier to a head concept, that property may undergo some changes. For example, a *zebra clam* may be “a clam with black and white stripes”, but the stripes on the clam will not be exactly the same as those on the zebra – they maybe narrower, shorter etc. Wisniewski (1998, p.1342) says that there are two important constraints inherent in this process. Firstly, the new version of a property must bear some resemblance to the original modifier property. Secondly, the final combination must still refer to a member of the head noun’s category. In other words, a *zebra clam* must still be a clam of some sort.

Construal is a process by which an interpretation of a compound can refer to some entity apart from the head concept of the combination. For example, an *artist collector* can be “a person that collects the work of an artist”. Construal is mediated by a process known as extensional feedback (Hampton, 1987; Murphy, 1988) that allows people to refer to existing representations in the real world. So, people can interpret a compound such as *firetruck onion* to a specific instance of a red onion that they have encountered. The Construal mechanism is required when interpretations point to some referent in the real world (e.g., *cheetah bike* as “a fast bike”), while Construction is used when no such referent exists (e.g., *cactus pig* as “a prickly pig”). Although Construction and Construal are presented as distinct processes, Wisniewski allows that they may be interdependent. The Construction process may first create a representation that then uses extensional feedback to be mapped to a real-world entity.

Dual-Process Theory extends the Concept Specialisation view of conceptual combination, providing much broader scope for the type of interpretations people create (i.e., property-based, relation-based and hybrids) and greater detail regarding the mechanisms people can use to in comprehension (i.e., alignment, comparison,
construction and construal). This theory has a certain amount of evidence in its favour but is nonetheless vulnerable to criticism, as we describe below.

**Evidence for Dual-Process Theory**

Much evidence has been garnered for the Dual-Process account of conceptual combination and its various predictions, particularly with reference to people’s use of alignable differences. Markman and Gentner (1993) found that when people are asked to list commonalities and differences between word pairs, they list more commonalities for similar words than for dissimilar words, which is not surprising. However, participants also list more differences for the similar pairs and these differences are conceptually related to the commonalities. For example, in comparing the concepts of *car* and *motorcycle*, “has wheels” is often given as a commonality, but equally “has four wheels” is offered as a difference (Wisniewski, 1997). By considering commonalities and alignable differences, people can identify the correct features to transfer from the modifier to the head concept. So, the compound *car motorcycle* could give the interpretation “a motorcycle with four wheels”. On the other hand, properties that have been found to be common to both concepts during alignment will not be used in interpretations. For example, the property “is coloured yellow” is common to both *banana* and *lemon* and so will not form part of an interpretation for *lemon banana*. Additionally, Wisniewski and Markman (1993) have shown that alignable differences are far more prevalent in interpretations than non-alignable differences, with the former occurring 78% of the time, compared to only 3% for the latter. In the same vein, Markman and Wisniewski (1997) showed that compounds where both concepts had the same superordinate category (e.g., basic-level categories such as pig and cow) were judged as more similar than compounds whose constituents have different superordinate categories (e.g., pig, tulip), or are
themselves differing superordinate categories (e.g., animal, plant). People produced more alignable differences for “basic-basic” compounds (e.g., *cow pig*), and consequently generated more property-based interpretations.

From the Alignment process arises the prediction that if the constituents of a novel compound are highly similar, then the resulting interpretation will more likely be property-based. This is due to the fact that more alignable differences will be found between the concepts. Wisniewski (1996, Experiment 2) found this to be the case, with 92% of 320 interpretations of 32 highly similar combinations interpreted using either property-based or hybridisation (e.g. *magazine paper* as a paper with glossy pictures).

Thus, there is compelling evidence to suggest that the mechanisms of Dual-Process theory have an important role in interpreting novel compounds. However, the theory is not without its failings.

**Evidence Against Dual-Process Theory**

A number of criticisms have been directed against the Dual-Process account. Primarily, its lack of definition of several critical components has been challenged, as has the necessity of some of these components in the interpretation process.

Aspects of the Dual-Process account, namely the structural alignment component, have been well-defined. In fact, this part of the theory has been implemented in several computer models, including the Structure Mapping Engine (Falkenhainer et al, 1989) and the MAC-FAC model of similarity-based analogical retrieval (Forbus, Gentner & Law, 1995). However, other aspects, more specifically the posited elaboration mechanisms (i.e., construal and construction), are lacking the same level of detail that would allow them to be implemented computationally.
Costello and Keane (2000, p.334) say that “The elaboration or construction process is to be singled out in this respect, as it is clearly a very complex process that is, as yet, unspecified”. In other words, describing a role for general knowledge is all well and good, but its theoretical merit is limited if no precise predictions can be made from the description of this role.

The notion of similarity also plays a pivotal role in Dual-Process Theory. For example, Wisniewski and others have claimed that increased similarity between constituents leads to increased property interpretations – “in order to apply a property of one concept to another, the concepts must be similar at least to some degree” (Wisniewski, 2000, p.36). Apart, from the general problems with such appeals to similarity (see Goodman, 1972), there is empirical evidence which contradicts Dual-Process’s prediction. Bock and Clifton (2000) found that the semantic similarity of a compound’s constituents does not impact on the number of property-based interpretations that people produce. Using a variety of concepts varying in type (e.g., artefact, natural kind) and also in the degree of similarity between constituents, they found no correlation between constituents’ similarity ratings and the number of property-based interpretations given by participants. This result contradicts the findings of Wisniewski and Love (1998, Experiment 1). Estes and Glucksberg (2000b) also observed similar results to Bock and Clifton, finding that the similarity of a compound’s constituents in no way predicted whether people produced interpretations involving property transfer. This has lead Estes and Glucksberg (2000a, 2000b) to question the role of the Alignment process in conceptual combination. They found that Alignment’s prediction (i.e., that the similarity of the constituent concepts would aid people in finding alignable differences between the head and modifier) was simply not borne out.
Costello and Keane (2001) have also found fault with the predictions of the structural alignment view. In their experiments, Costello and Keane systematically varied the alignability and diagnosticity of properties within compounds and asked people to rate the resulting interpretations. For example, participants received different possible interpretations for the compound *bumblebee moth* that incorporated features that were either aligned (e.g., a bumblebee’s colour) or non-aligned (e.g., a bumblebee’s sting) and diagnostic (e.g., black and yellow colouring) or non-diagnostic (e.g., that bumblebees fertilise plants). When participants rated these possible interpretations, it was found that they preferred non-alignable properties if these properties were diagnostic of the modifier concept, going against the Dual-Process prediction. Costello and Keane (2001) have also shown that this preference exists when people produce their own interpretations for novel compounds. From these studies, Costello and Keane argue that the predominance of alignable differences in the work of Wisniewski and Markman is merely an artefact of their material selection, as they did not control for diagnosticity. In other words, features that could be viewed as alignable differences were actually highly diagnostic properties of the concepts. For example, an alignable difference between “bluebird” and “robin” was the colour blue, but this is also the most diagnostic aspect of bluebirds compared to other birds. In response to this criticism, Wisniewski and Middleton (2002) accept that such interpretations may not emerge from seeking alignable differences. They found that in many cases participants exhibited a preference for interpretations that used non-alignable differences, which they say is not predicted by previous theories that use alignment – “results showed that subjects prefer and produce referents based on differences that researchers would classify as non-alignable”. They propose that property-based interpretations may also arise from
a spatial (rather than structural) alignment of concepts e.g., a *bucket bowl* may be a bowl with a handle attached to the top, while a *coffee-cup bowl* could be a bowl with a handle attached to the side.

Finally, Costello and Keane (2000) have also raised concerns over the lack of parsimony in the Dual-Process account. Dual-Process Theory requires multiple processes in order to account for different types of interpretation – Alignment (structural and spatial) gives rise to property-based interpretations, Scenario Creation produces relational interpretations, while Construal accounts for cases where the referent of the compound is not a member of the head concept’s category. Even with its multiple processes, Costello and Keane argue that Dual-Process Theory cannot explain other types of interpretation such as known-concept interpretations (i.e., where the referent is a concept that already has another name e.g., *pencil bed* meaning pencil case) and what they term reversals (e.g., *slipper bed* meaning a slipper that is used as a bed by a cartoon animal).

In summary, while much of people’s behaviour in interpreting novel compounds appears to be describable in terms of the Dual-Process theory, there is mounting evidence against the need for alignment and against the importance of alignable differences.

### 2.2.3. Constraint Theory

The Constraint Theory of conceptual combination is probably the most completely specified account of conceptual combination. Each aspect of the Constraint Theory has been clearly detailed and has been implemented as a computational model, C3. Constraint Theory is also the first theory to address the
possibility of multiple interpretations for novel compounds and how different types of interpretation are produced, and it does this based on a unitary interpretation mechanism. The Constraint Theory describes conceptual combination as a process of constructing representations of interpretations that satisfy the three central constraints of diagnosticity, plausibility and informativeness. These constraints arise from pragmatic assumptions about compound use and interpretation (Costello & Keane, 2000; Grice, 1975).

Briefly stated, the diagnosticity constraint requires that interpretations should contain diagnostic properties from both the head and modifier concepts. The diagnostic properties of a concept are those properties that best identify that concept. The plausibility constraint ensures that interpretations contain properties that are consistent with our prior experience. Finally, the informativeness constraint makes sure that interpretations provide a certain amount of new information. We describe the constraints in more detail, using examples to demonstrate the effect of each in the construction of interpretations.

**D**iagnosticity

D**iagnosticity requires that interpretations contain diagnostic properties of both concepts, and is similar both to Tversky’s (1977) notion of diagnosticity and to Rosch’s (1978; Rosch & Mervis, 1975) “cue validity”. In each case, the diagnosticity of a feature is determined by how frequently it occurs in a concept compared to other concepts. For example, the most diagnostic feature of the concept “skunk” might be “has a noxious smell”, since people could easily identify this feature as referring to a skunk as opposed to any other concept. The diagnosticity constraint means that, all else being equal, one interpretation is better than another if it incorporates a more
diagnostic feature. In the example below, two possible interpretations for *cactus fish* are given (Costello & Keane, 1997).

1a. A cactus fish is a prickly fish

1b. A cactus fish is a green fish

The diagnosticity constraint states that the first example (1a) would be a more acceptable interpretation than the second (1b), since “prickly” is more diagnostic of “cactus” than the colour green.

This example shows clearly how diagnostic information is evident in property-based interpretations. However, Costello and Keane argue that diagnosticity is equally important to other types of interpretation, including hybrids and relational ones. For example, the hybrid interpretation “a *pet fish* is a goldfish”, implicitly possesses diagnostic predicates of both concepts – “kept by people” is a diagnostic feature of pets and any representation of goldfish contains a number of features diagnostic of fish (e.g., has fins, swims in water, has gills etc.). Similarly, relational interpretations will also implicitly contain diagnostic features of both constituents. For example, “an *oyster hammer* is a hammer for opening oysters” mentions both concepts explicitly in the interpretation, and so means that each concept’s diagnostic features are implicitly present (Costello & Keane, 2000). Obviously, some features of a concept are more diagnostic than others. For example, if only elephants have trunks, then that feature is highly diagnostic of elephants. On the other hand, many birds might have long legs (e.g., ostrich, flamingo, emu) and so this feature would be less diagnostic of these concepts, but this would still be more diagnostic than “has feathers”. In general, the more concepts that posses a particular feature, the less diagnostic that feature is.
Plausibility

The plausibility constraint requires that interpretations are consistent with our prior knowledge. In other words, an interpretation should describe some object that could plausibly exist. For example, an interpretation that refers to an entity that already exists will be completely plausible, since all of the information in the interpretation overlaps completely with a previously encountered instance. In this way, the plausibility of interpretations is determined by the degree to which they overlap with prior knowledge. Here are two candidate interpretations for the compound *angel pig*:

2a. An angel pig is a pig with wings on its torso.

2b. An angel pig is a pig with wings on its tail.

Interpretation (2a) is considered to be more plausible because we have previously encountered entities with wings on their torsos, whereas we have not encountered entities with wings on their tails (2b). In this way, Costello and Keane show that the plausibility constraint allows additional features to be incorporated into interpretations that have co-occurred with each other in the past (e.g., wings on torsos, as opposed to wings on tails).

Informativeness

By Grice's (1975) maxims, the informativeness constraint requires that interpretations contribute some new information. Because people are generally cooperative in discourse, the words of a novel compound can be viewed as necessary and sufficient in describing an intended referent (Costello & Keane, 2000). If this were not the case, the compound producer would have opted for a different phrasing (e.g., a longer phrase or a single lexical item). This means that feasible interpretations
will be culled if they do not communicate any new information relative to either constituent concept. For example, a *bed pencil* meaning “a pencil made of wood” would not be considered a viable interpretation, since “made-of-wood” gives no new information about either concept. In other words, there must be some transfer of information. Downing (1977) argued for a similar idea, stating that compounds with “redundant” modifiers are unacceptable (e.g., *pig pork*).

**The C3 Model of Conceptual Combination**

The C3 model is a computational implementation of the Constraint Theory (Costello & Keane, 2000, 2001; Keane & Costello, 2001). It implements fully the primary constraints of the theory (diagnosticity, plausibility and informativeness), while making certain assumptions about knowledge representation.

There are two principal knowledge assumptions in the Constraint Theory. First, there is a distinction made between artefacts (gun, cup, building) and natural kinds (lemon, bird, tree, snail). This assumption is based around the idea that artefacts have functional roles (e.g., agent, object, recipient, instrument), whereas natural kinds do not. Furthermore, these functional roles have a wider scope in artefacts (e.g., a gun can shoot anything, but a snail can only eat vegetative matter). Second, concepts are assumed to be of a complex predicate structure, containing attributes, objects and relations. In addition, individual concepts in C3 are represented as schemas, containing attributes, roles and relations. This gives C3 access to a wide variety of knowledge to aid the interpretation process, including instances of concepts, prototypes, instances of related concepts, event representations involving these concepts and general domain theories. The extent of available knowledge specified in the C3 model is far greater than that of other theories, which have limited other types
of knowledge to summary, prototype information (see Dual-Process theory and Murphy, 1988).

The input to the model is a noun-noun combination and the output is a set of interpretations, each of which is represented by a set of predicates. Each interpretation has an overall acceptability score that is calculated on the basis of the diagnosticity of the predicates and the plausibility of the interpretation. The steps involved in constructing these interpretations are detailed below.

*Diagnosticity.* Diagnostic predicates of a concept are those predicates which occur frequently in instances of that concept and rarely for other concepts. Before interpretations can be constructed, the diagnosticity of each predicate (or set of predicates) of the constituents is calculated over the entire knowledge base. For example, the predicate [shape(tubular)] gets a high diagnosticity score for the concept “finger” as it does not occur with other concepts in C3’s knowledge base.
Once diagnosticity scores have been calculated, partial interpretations are constructed by combining sets of diagnostic predicates from both concepts. The diagnostic predicates can be unified in many different ways, depending on the roles in each predicate. For example, a predicate “contains” could be instantiated as [contains(x, y)] or [contains(y, x)] to give two different partial interpretations.

Figure 2.1 Diagnostic features of cactus and beetle are highlighted in grey. These features are combined to form different partial interpretations (denoted by X and XY). X refers to some entity that has 6 legs, is small and is prickly. XY refers to a relationship between two entities, X (is prickly) and Y (is small and has 6 legs).

In Figure 2.1 we show how the diagnostic features of cactus and beetle can be combined in different ways to produce alternative partial interpretations. For example, one partial interpretation might refer to an entity that was small, prickly and had six legs. Another partial interpretation could specify a relationship between something that is prickly and something that is small and has six legs, as illustrated. These partial interpretations are then fleshed out in the plausibility stage of the model.
**Plausibility.** Full interpretations are constructed around the partial interpretations from the first stage of the interpretation-construction process. To calculate the plausibility of an interpretation, the overlap between it and all instances in the knowledge base is determined. The most plausible interpretation will have the greatest degree of overlap with the stored instances. Additional predicates can be added to partial interpretations to increase their plausibility if they have co-occurred with predicates that already form part of the interpretation. Figure 2.2 shows how the partial interpretations from Figure 2.1 can be elaborated with co-occurring predicates from prior knowledge.

![Diagram](image)

*Figure 2.2 The plausibility component completes partial interpretations (X and XY) by adding co-occurring features from prior knowledge.*

In Figure 2.2, two full interpretations are constructed by adding features from one (“a prickly beetle”) or both concepts (“a beetle that eats cacti”). Once a full interpretation has been produced its plausibility score is calculated (as the average overlap with stored instances) and it can then be examined for informativeness and overall acceptability.
**Informativeness.** By the informativeness constraint, interpretations must contain new information relative to the constituents of the compound. In order to test this, full interpretations are compared to concept prototypes in the knowledge base. An informative interpretation will contain one or more predicates that are not contained in a prototype. In this way, informativeness is considered an all or nothing affair. Uninformative interpretations are automatically considered unacceptable and so receive no further attention. A candidate interpretation for the compound *cactus hedgehog* might be “a prickly hedgehog”. However, when this is compared to a stored instance of hedgehog, it would emerge that hedgehogs are already prickly and so this interpretation adds no new information. This interpretation would then be rejected as being uninformative.

If an interpretation is deemed informative, then its overall acceptability score is calculated. This score is based on both the diagnosticity of the predicates that make up the interpretation, and its plausibility score. Because C3 generates such a large number of candidate interpretations (around 4,000), Costello and Keane generally imposed a threshold on the acceptability score when testing the model. By using an objective threshold, C3 can return the top ten interpretations for any given compound.

**Evidence for the Constraint Theory**

Through its constraint satisfaction approach the Constraint theory explains a wide variety of phenomena, including the diversity of interpretations provided by people. While other theories generally refer to property- and relation-based interpretations, the Constraint theory also explains interpretations that refer to known concepts (also known as “direct reference”), and those that they define as reversals (e.g., *slipper bed* meaning “a slipper that is used as a bed (by a cartoon animal)”).

30
The Constraint theory also accounts for people’s capacity to produce multiple interpretations for single compounds, shown by Costello and Keane (1997).

Apart from these general aspects of conceptual combination, the Constraint theory makes several specific predictions which are borne out. Costello & Keane (2001) demonstrated people’s preference for diagnostic features of concepts when judging and generating interpretations. For each compound, they had people rate four different interpretations, based on a combination of diagnostic / non-diagnostic and alignable / non-alignable features. People preferred diagnostic interpretations, irrespective of whether they were alignable or not. Also predicted by the Constraint theory is people’s generation of known-concept interpretations. For the compound *stilt bird* people will sometimes give the interpretation as “flamingo”. Costello and Keane (1997) found that these interpretations accounted for approximately 10% of all interpretations. Another type of interpretation predicted by Constraint theory is *focus reversals*, or simply reversals. These are interpretations where the focal concept is the modifier and not the head concept, as is usually the case (e.g., *slipper bed* referring to a slipper used as a bed by two chipmunks). What’s more, the theory predicts that these kinds of interpretations are more likely to be property-based than interpretations where the head is the focal concept. Normally, property-based interpretations account for around 30% of interpretations, but in focus reversals property-mappings account for 50% of responses. This was found to be the case in a series of experiments performed by Costello and Keane (1997; Keane & Costello, 2001).

A basic tenet of the Constraint theory is that there are multiple interpretations for any novel compound. The extent of this polysemy also hinges on the types of concepts being interpreted (e.g., artefacts versus natural kinds). Costello and Keane argue that artefacts can fulfil a greater number of functional roles, and thus Constraint
theory predicts that compounds with an artefact as the head concept will give rise to
greater polysemy. For example, Costello and Keane (2000) compare the examples
\textit{wasp gun} and \textit{wasp cow}. The former could be “a gun for shooting wasps”, “a gun that
uses wasps as bullets”, “a gun used by wasps” and so on. On the other hand, the
possible interpretations for \textit{wasp cow} seem more limited – “a cow that stings” or “a
striped cow”.

Apart from the empirical support for the diagnosticity constraint in Costello
and Keane’s work, the plausibility and informativeness constraints have more general
support stemming from the pragmatics literature (Grice, 1975; Sperber & Wilson,
1986). People rarely, if ever, produce uninformative compounds, which offers a clear
need for such a constraint. Also, Downing (1977) has shown that people find it
difficult to interpret uninformative compounds such as \textit{head hat}. The importance of
the plausibility constraint is evident in people’s use of prior knowledge when
interpreting novel combinations. When a person is interpreting novel compounds it is
always their goal to arrive at a plausible interpretation. However, support for these
ideas is more anecdotal than from empirical evidence.

\textbf{Evidence against the Constraint Theory}

To date there is little empirical work that specifically contradicts the central
ideas of Constraint theory. However, Costello and Keane’s work has not provided
specific empirical testing of the plausibility and informativeness constraints. Estes
and Glucksberg (2000a, 2000b) though, have examined the role of diagnosticity in
conceptual combination. They argue for a form of salience which supersedes
diagnosticity, by incorporating relevance as a factor in producing interpretations.
They found that a head concept required a relevant dimension to match with a salient
feature of the modifier for people to give property-based interpretations. For
example, a zebra bag could be “a bag with black and white stripes” as bag has the relevant dimension of “pattern”. On the other hand, zebra trap would not be interpreted with a property, as trap does not have the relevant “pattern” dimension. For the moment, however, Estes and Glucksberg have relied on people’s intuitions of what a relevant dimension is, without a specific formalisation. It is difficult, therefore, to make direct comparisons with the role of diagnosticity in Constraint Theory.

Markman (in personal communication with Keane, 2001) has argued that the plausibility constraint is really equivalent to the structural alignment aspect of the Dual-Process theory. Markman’s argument is that the plausibility of an interpretation is established through structurally aligning concepts in memory. Keane and Costello reject this criticism, saying that concepts are never structurally aligned since the plausibility stage is an incremental process. Indeed, if comparisons were to be drawn with the Dual-Process Theory, it might be fairer to compare plausibility to the extensional feedback aspect of the Dual-Process account, since both use prior knowledge to elaborate interpretations. However, Wisniewski (personal communication with Keane, 2001) maintains that any relational link between two concepts is through a process of alignment.

In interpreting novel compounds, Constraint Theory requires that compounds be subject to the three constraints, meaning that all compounds require the same degree of processing in order to be understood. However, Constraint Theory currently makes no allowance for the comprehension of familiar or analogically related compounds, which tend to be understood more quickly than completely novel compounds (Tagalakis & Keane, 2003). For example, the compound spam artist (referring to someone who attempts to deceive others via claims made in spam email) can easily be
interpreted with reference to the existing compound *con artist*. Constraint Theory has no facility for allowing such shortcuts in comprehension.

In terms of the C3 implementation of Constraint Theory, Costello and Keane (2000) admit that the generative mechanism is computationally expensive and does not fully reflect people’s processing of novel compounds. C3 produces on average 4,000 interpretations for a given compound and can take upwards of 2 hours to complete its task. In their own empirical studies, Costello and Keane (1997) found that people produce on average two different interpretations, with six interpretations being the most that any individual produced.

Finally, Costello and Keane’s position on reversals (i.e., where the interpretation focuses on the modifier concept, and not the head concept) is open to challenge. For example, the compound *slipper bed*, referred to by two chipmunks in a Disney cartoon, is considered to be a reversal when it means “a slipper that is being used as a bed”, because the item in question is still a slipper. However, this interpretation could also be defined as simply property-based, where the item in question is a bed (with all the properties of a bed) that has the form/shape of a slipper. Recategorising Costello and Keane’s reversals as property-based interpretations would lead to a more parsimonious account of interpretation types. To the person using a *slipper bed*, whether it is in fact a slipper or a bed is an issue of psychological essentialism (e.g., Medin & Ortony, 1989) rather than conceptual combination.

So, while Constraint theory and the C3 model admit the full creativity of conceptual combination, it does not address how people can arrive at interpretations quickly, using compounds already familiar to them. Furthermore, aspects of the Constraint theory are open to further empirical testing (e.g., the role of plausibility). The nature of the plausibility constraint suggests that the more consistent an
interpretation is with prior knowledge the more acceptable it should be. However, this issue has not yet been examined empirically. That said, Constraint Theory and C3 do provide a fully-specified account of conceptual combination that details the exact role of prior knowledge in constructing interpretations, something that has been lacking in other theories (e.g., the Dual-Process theory).

2.2.4. CARIN Theory

The Competition Among Relations In Nominals (CARIN) theory (Gagné, 2000, 2001; Gagné & Shoben, 1997, 2002) is inspired in part by the early linguistic work of Levi (1978) and Downing (1977), who examined the use of taxonomies of relations in order to describe the meanings of nominal compounds. This view of comprehension in conceptual combination sees the process as being less creative than other theories have suggested (see Dual-Process theory and Constraint Theory). The CARIN approach sees the goal of conceptual combination as finding a linking relation (also called a thematic relation) between the constituents of the compound. Philosophically this approach is quite distant from the previous theories we have discussed, with conceptual combination being reduced to fitting all compounds into existing relational templates.

The overarching concept of the CARIN theory is that people possess distributional knowledge about how often particular relations have been used with particular concepts in the past. Distributional knowledge of words is gleaned from statistical analysis of how those words are used in the language (see Redington & Chater, 1998; Saffran, Newport & Aslin, 1996). The idea is that relations (e.g., made of, located in, for, has, etc. See Table 2.3 for a complete list of possible relations used in the CARIN theory to interpret novel combinations) compete for the interpretation
of a compound and that the difficulty of interpreting a compound is a function of the relative strength of the selected relation.

Table 2.3 List of thematic relations used by the CARIN theory and model.

<table>
<thead>
<tr>
<th></th>
<th>Relation Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Noun causes Modifier</td>
</tr>
<tr>
<td>2.</td>
<td>Modifier causes Noun</td>
</tr>
<tr>
<td>3.</td>
<td>Noun has Modifier</td>
</tr>
<tr>
<td>4.</td>
<td>Modifier has Noun</td>
</tr>
<tr>
<td>5.</td>
<td>Noun makes Modifier</td>
</tr>
<tr>
<td>6.</td>
<td>Noun made of Modifier</td>
</tr>
<tr>
<td>7.</td>
<td>Noun for Modifier</td>
</tr>
<tr>
<td>8.</td>
<td>Modifier is Noun</td>
</tr>
<tr>
<td>9.</td>
<td>Noun uses Modifier</td>
</tr>
<tr>
<td>10.</td>
<td>Noun located in Modifier</td>
</tr>
<tr>
<td>11.</td>
<td>Noun about Modifier</td>
</tr>
<tr>
<td>12.</td>
<td>Noun during Modifier</td>
</tr>
<tr>
<td>13.</td>
<td>Noun used by Modifier</td>
</tr>
<tr>
<td>14.</td>
<td>Modifier located in Noun</td>
</tr>
<tr>
<td>15.</td>
<td>Noun derived from Modifier</td>
</tr>
<tr>
<td>16.</td>
<td>Noun by Modifier</td>
</tr>
</tbody>
</table>

Therefore, every relation associated with a noun will have a particular strength that affects how likely it is for it to be used in an interpretation. For example, the locative relation (*in*) for “mountain” is viewed as a high-strength relation as it is used more frequently (e.g., a *mountain stream* is a stream in a mountain location) than the about relation (e.g., a *mountain magazine* is a magazine about mountains) and so will be understood more quickly. Additionally, the CARIN theory holds that the availability of more than one high-strength thematic relation will also slow comprehension. For example, the top three relations for the concept “headache” have frequencies of 33%, 33% and 21%, while the top three for “juvenile” are 34%, 20% and 15%. Although the top ranked relations for each of these concepts have almost identical percentages, CARIN predicts that compounds involving “juvenile” will be easier to understand than those involving “mountain”, since the second and third ranked relations are of much lower frequencies. Mathematically, the strength of a relation for a concept is expressed as an exponential decay function, as shown in Figure 2.3.
Figure 2.3 Formula for calculating the strength of a relation for a noun. PSR is the Probability of the Selected Relation occurring, and PRR\textsubscript{n} is the overall Probability of a Relation Ranked \textit{n} occurring.

From her empirical studies, Gagné makes two important distinctions between CARIN and previous theories of conceptual combination. Firstly, she advocates the primacy of the modifier concept in the interpretation process. Gagné has found that the strength ratios associated with modifier concepts have a significant effect on response times, whereas the strength of head concept’s relations did not. In other words, information about the modifier is more important than information about the head concept. This does not mean that CARIN does not provide a role for the head noun in compounds, simply that there is an asymmetry in the degree of influence of the head and modifier concepts (Shoben & Gagné, 1997). The theory argues that it is the modifying concept that guides the selection of the thematic relation. Secondly, because the CARIN theory is based on the notion of modifier-relation frequency, the production of property-based interpretations is seen as a last resort strategy. This means that people will only produce a property-based interpretation if they have failed to find an appropriate relation between the compound’s constituents. Shoben and Gagné say:

[Property-based] interpretations are secondary to the strategy of relation assignment. That is, people always try to assign a thematic relation to a combination; however, if they are unable to assign a relation, they seek a metaphorical or property-mapping solution. (p.47)
More recently (Gagné, 2000) has proposed that property-based interpretations may be constructed via a new “Resembles” relation. For the example *magazine newspaper*, rather than mapping a property directly from the modifier to the head concept, people first link the concepts (e.g., newspaper RESEMBLES magazine) and then transfer some property (e.g., a glossy newspaper). However, Gagné still holds that these kinds of interpretations are extremely uncommon and remain a last resort.

Using modifier-relation frequencies and the primacy of the modifier, the CARIN model makes strong predictions concerning the ease of comprehension of compounds. Below, we look at the empirical evidence which supports these ideas, and also some conflicting evidence.

**Evidence for CARIN**

Gagné (2000, 2001; Gagné & Shoben, 1997, 2002; Shoben & Gagné, 1997) have provided much empirical support for the central ideas of the CARIN theory. Three of these central claims; namely the fact that ease of comprehension can be predicted using modifier relation frequency, that the role of the modifier is more important than the head in predicting comprehension times, and that constructing property-based interpretations is only a last-resort strategy.

The CARIN theory makes strong predictions about people’s response times for noun-noun compound judgements. Gagné argues that the ease of comprehension of compounds can be predicted based on the modifier-relation frequency. Using this metric, Gagné (2000) created a set of compounds whose constituents used either high-strength (H) or low-strength (L) relations in their interpretation. For example, in a HH compound the relation was high-strength for both the head and modifier concepts (e.g., *plastic toy* – both concepts use the made of relation), while in an LH compound the relation was high-strength for the head concept but low-strength for the modifier.
concept (e.g., chocolate plant – the made of relation is used frequently with the chocolate, but not with plant). Gagné found that when a relation is used frequently with a modifier (e.g., in HH and HL compounds) people were faster to respond than if a relation occurred infrequently with the modifier (e.g., LH compounds). Overall, it was found that the two best predictors of response times were the rank of a relation for the modifier (i.e., higher ranked relations are responded to more quickly) and the number of high-strength relations for the modifier (i.e., fewer high-strength relations means faster response times).

In other response time experiments, Gagné & Shoben (1997) found support for the primacy of the modifier. They found that while there was an effect of relation frequency for the modifier concepts, there was no equivalent effect for head nouns. In other words, varying the modifier relation frequency affected people’s response times, but changing the head relation frequency did not. Thus, compounds that were categorised as HH (High-strength relation for the head compound) and HL (Low-strength relation for the head compound) were responded to equally quickly, but there was a difference between HH and LH compounds. Gagné and Shoben also analysed both the rank of relations and the number of dominant relations associated with head and modifier concepts. Again, the results corresponded to the primacy of the modifier view. For modifier nouns, both variables correlated significantly with comprehension time. For head nouns, there was no such correlation, suggesting that the role of the modifier is indeed more critical than that of the head.

Finally, Shoben and Gagné (1997) have argued that “property matches are the interpretation of last resort” (p.35). From their corpus analyses (see also Gagné, 2000; Gagné & Shoben, 1997), they concluded that property-based interpretations rarely occur and so relational interpretations are made by preference. Analysing data
from Warren (1978) revealed that only 1.6% of interpretations were property-based compared to 86% relational interpretations. In her analysis of her own corpora, Gagné (2000) found only 0.6% of meanings involved property transfer. Additionally, Downing (1977) argued that property-based interpretations were a kind of fallback mechanism. Gagné (2000) tested empirically whether property-based interpretations were indeed more difficult to process than relational ones. In a series of experiments she showed that property interpretations were not easily interpreted and were not easily judged as being acceptable. The CARIN theory does not preclude the generation of property interpretations, but they are arrived at through an elaboration process which derives features of the combined concept, once a linking relation has been found (Gagné, 2000, p370). For example, a combination such as *coat shirt* could first be interpreted using a *resembles* relation (e.g., a shirt that resembles a coat). This interpretation could then be elaborated with features or properties of the modifier (e.g., is thick, is long). In this way, property interpretations arise from a relation linking process. Gagné maintains that these findings provide converging evidence for the prioritising the role of relation linking in interpreting novel compounds.

The CARIN account offers an account of relational conceptual combination and provides a good fit to response time data. However, as we discuss below, there are many areas where CARIN is found wanting.

**Problems for CARIN**

Although CARIN provides a clear theoretical description of how people arrive at relational interpretations, there are problems with the generality of the processes involved, especially when dealing with more creative aspects of conceptual
combination. Additionally, evidence exists to suggest that property-based interpretations should not be considered a last-resort strategy.

Like all taxonomic approaches, which propose a finite set of relations to capture the full range of meanings in conceptual combination, CARIN suffers from a lack of generality. Firstly, if a noun is used for the first time in a compound in the modifier position, CARIN offers no opinion on how this will affect the interpretation process. Secondly, by restricting the taxonomy to 16 relations CARIN restricts itself from the outset. What if new relations are needed? Can the taxonomy be expanded indefinitely? It is easy to think of compounds whose interpretations involve a thematic relation that is not covered by CARIN’s sixteen relations (see Table 2.3). For example, compounds such as party girl (e.g., “a girl that attends a lot of parties”) or metaphorical compounds (e.g., tarantula teacher meaning “a dangerous teacher”) do not fit easily into her taxonomy. A related concern is that while Gagné’s relations might provide adequate coverage, they are oftentimes too abstract to provide meaningful interpretations. For example, the for relation is used in interpreting paint spoon and blueberry spoon, but the former is more likely to refer to “a spoon for stirring paint” while the latter refers to “a spoon for serving blueberries”. People can clearly distinguish between the roles of the spoon in both instances (i.e. to stir versus to eat), but the CARIN theory has abstracted the roles to a level from which they cannot be separated (Wisniewski, 1997a).

Gagné’s last-resort hypothesis for property interpretations is inconsistent with other studies. Her corpus analysis showed that property-based interpretations occur less than 2% of the time, and led her to conclude that they have little theoretical import. However, work by both Gentner and Markman (1993) and Costello and Keane (2000) suggests that these interpretations can account for 30-50% of all
interpretations. Indeed, where the constituent concepts are highly similar, this figure can be as high as 70% (see Wisniewski & Love, 1998). In fact, Wisniewski and Love demonstrate that even where a plausible relation could exist between the head and modifier concepts, people often supplied property-based interpretations if the constituents were similar to each other. These findings greatly undermine Gagné’s position on property-based interpretations, and suggest that the CARIN theory may be neglecting half of the conceptual combination framework.

Tagalakis & Keane (2003) have also questioned some of Gagné’s findings at a more fundamental level. They have argued that many of the results are confounded by the familiarity of the compounds, with more familiar compounds being responded to quicker than less familiar items. For example, compounds which were classed by Gagné as having high relation frequencies, such as wood furnace, gas cloud and office plant, have been shown to be more familiar than compounds which were classed as having low relation frequencies, such as servant language, plastic crisis and family utensils. When familiarity is taken into account, Tagalakis and Keane found that there was no advantage of relational interpretations over property interpretations, in terms of ease of comprehension.

While Gagné has provided much empirical data for the importance of distributional knowledge, determined through prior experience, there are many aspects of the CARIN theory that require greater empirical investigation.

### 2.3. Conceptual Combination in Context

Having examined conceptual combination in isolation we now come to the second pillar of this thesis; conceptual combination in context. Despite the proliferation of theories of conceptual combination (see Concept Specialisation, Dual-
Process theory, Constraint theory and CARIN in the previous section), none have specifically addressed the question of how their theories translate to in-context situations. The key question is whether the empirical findings to date, which have resulted from experiments concerning out-of-context conceptual combination, have any bearing on how novel combinations are processed in context. We look at existing context work, both in the areas of lexical ambiguity research (i.e., dealing with lexical ambiguity in and out of context) and conceptual combination research, and consider whether we have they can give us a clearer picture of (i) how novel compounds are processed in context, and (ii) whether out-of-context interpretations and processes affect conceptual combination in context.

One might suppose that there is something to be learned from context research carried out in other areas. However, mainstream language research focuses on somewhat different issues when pursuing the study of contextual influences (e.g., Glucksberg, Kreuz & Rho, 1986; Kintsch & Mross, 1985; Onifer & Swinney, 1981; Tabossi, 1988, 1989; Till, Mross & Kintsch, 1988). For instance, a key issue in lexical ambiguity research centres on the resolution of homographs; that is, how context affects lexical access to a word that has different meanings (e.g., interest meaning “hobby” versus “money accrued on a loan”). However, this is not necessarily useful to conceptual combination research, as novel noun-noun combinations may still give rise to multiple meanings even when there is no ambiguity in the constituent words (i.e., they are not homographs; Clark & Clark, 1977; Costello, 1997; Costello & Keane, 1997, 2000, 2001; Gleitman & Gleitman, 1970). For example, *cactus beetle* can be interpreted to be “a beetle with spikes”, “a beetle that eats cacti”, or “a beetle found in the desert” without recourse to different senses of “cactus” or “beetle”. Unlike context research carried out in other areas, the
research focus in conceptual combination is less on lexical access, *per se*, and more on how meanings get combined and filtered.

Regarding the conceptual combination literature, the bulk has examined how people understand novel compounds in isolation (cf. Gerrig & Bortfeld, 1999; Gerrig & Murphy, 1992). In these situations, compounds are presented to people without any sentential context or referential cues (see e.g., Coolen, Van Jaarsveld & Schreuder 1991; Costello & Keane, 1997, 2000, 2001; Gagné, 2000, 2001; Gagné & Shoben, 1997; Smith, Osherson, Rips & Keane, 1988; Wisniewski, 1996, 1997; Wisniewski & Love, 1998). We call this *out-of-context presentation*. Clearly, this presentation procedure does not correspond to people's experience in everyday life. In real-world situations, a wide variety of contextual influences must allow listeners to disambiguate novel compounds and arrive at a speaker’s intended meaning.

One piece of work by Gerrig & Bortfeld (1999) does approach the issue of in-context comprehension of novel compounds. They specify two possible standpoints for the role of context in conceptual combination – the *interdependence view* and the *independence view*. Put simply, the interdependence view states that interpreting novel conceptual combinations in context depends on out-of-context factors, while the independence view states that out-of-context factors have no impact on in-context processing. When people are given a compound like *skunk squirrel* in isolation, they produce multiple interpretations such as “a squirrel that is black and has a white stripe down its back” or “a squirrel that emits a noxious smell” (see Costello & Keane, 1997, 2000). But, if skunk squirrel is presented in the context of the sentence:
then the squirrel-that-smells interpretation is likely to be preferred. According to the independence view, this in-context meaning is produced without the activation of other possible meanings, whereas the interdependence view would see this meaning as somehow winning out over other activated possible meanings.

Clearly, the fate of a large body of the conceptual combination literature hinges on the resolution of this dependence debate. If the independence view is true, then most of the research on conceptual combination would have to be written off as being of little relevance to the real-world phenomena of in-context comprehension. We present both views of conceptual combination in context and look at the supporting evidence for each.

### 2.3.1. Evidence for the Independence View

Gerrig and Bortfeld (1999) have provided the main empirical support for the independence view. Gerrig and Bortfeld posited that in-context comprehension was not affected by out-of-context factors. This would mean that there should be no evidence of interference in-context, from out-of-context factors. They pitted the independence and interdependence views against one another by measuring reading times for matched pairs of novel compounds (e.g., *baseball smile / doll smile*) presented in the context of brief stories. The stories supported either a similar innovative interpretation for both compounds (e.g., “a smile evoked in a young child on being given a baseball/doll as a present”), or provided a neutral context. The interpretations were considered innovative as participants did not normally produce them in their out-of-context understanding of the compounds. Gerrig and Bortfeld
showed that the matched compounds differed from one another in the speed with which people could think of an interpretation out-of-context (e.g., doll smile was of High Accessibility because people responded to it more quickly than the Low Accessibility baseball smile). They posited that if the compounds were understood equally easily in-context, this would mean that the out-of-context meanings had not interfered with the in-context processing, thus validating the independence view. However, the results Gerrig and Bortfeld found were more in line with an independence view of conceptual combination. In the innovative contexts, compounds were understood equally easily, irrespective of their out-of-context ease of interpretation. This was true both for online reading of the stories and for paraphrase judgements at the end of each story. In the neutral context, the pattern of results reflected people’s out-of-context processing, with High-Accessibility compounds being understood more quickly than Low-Accessibility ones. Gerrig and Bortfeld concluded that readily available, out-of-context meanings do not interfere with the comprehension of a compound when an innovative meaning is provided by the story context. This result is consistent only with the independence view. They argued that only one meaning was activated in-context, namely the interpretation detailed in the story context.

Gerrig and Bortfeld are the first to explicitly advocate the independence view and provide empirical support for their claims. However, with evidence being thin on the ground, further evidence must be provided before the issue can be clarified.

2.3.2. Evidence for the Interdependence View

Although conceptual combination theories have rarely addressed the effects of context, an implicit assumption of many is that the processes and factors that drive
out-of-context comprehension will also impact on in-context processing. In other words, the empirical findings in out-of-context research will translate to in-context situations. Wisniewski (1996, p.450) makes explicit this argument in saying: “It makes sense to first identify how the meanings of the constituents (i.e., prior knowledge) affect interpretation. Then the role of discourse settings may be more meaningfully understood in light of these prior knowledge affects”. This viewpoint carries over to existing models of conceptual combination. Again, this assumption is presented implicitly with researchers viewing context as a means of filtering the interpretations that would normally be produced out-of-context.

The interdependence view holds that when a compound is presented in-context, the context acts as a kind of filter to make one of the possible out-of-context meanings emerge as the correct interpretation. This view suggests that meanings that are more readily available out of context are also, by default, more readily available in context. Contrasting with Gerrig and Bortfeld’s results, the interdependence view predicts that that this out-of-context availability will affect people’s comprehension in context. Thus far, only Gagné and Spalding (in sub) have provided evidence for the interdependence view in tackling Gerrig and Bortfeld’s position. Gagné and Spalding have tried to address this issue using a similar paradigm to the one described above, and raising some concerns they have with Gerrig and Bortfeld’s (1999) interpretation of their results.

Gagné and Spalding (in sub) examined whether modifier-relation frequency influenced the comprehension of novel compounds in context, as well as out of context. They used matched pairs of novel compounds that differed along the dimension of modifier relation frequency, which forms the basis of the CARIN theory. These compounds were first examined out-of-context and then embedded in
short context stories to examine in-context processing. For example, out of context, they show that the compound *office light* is responded to more quickly than the compound *home light*. The CARIN explanation for this difference is that when functioning as a modifier, “office” uses the locative relation (*in*) more often than does “home”, so *office light* is interpreted more quickly even though both *office light* and *home light* could refer to possible locations for lights. Rather than using innovative meanings for compounds (as Gerrig and Bortfeld did), Gagné’s compounds were based on interpretations that people were likely to produce, based on the frequency of the modifier relation. In this way, the underlying relation is kept constant while the modifier relation frequency is varied – high-frequency for *office* and low-frequency for *home*. Gagné and Spalding posited that if the interdependence view were to be borne out, then high-frequency modifier relations should still be processed more quickly than low-frequency modifier relations when the compounds are presented in context. They found this to be the case, and show that in context, high-frequency items were responded to more quickly than low-frequency items. Gagné and Spalding conclude that both conceptual and contextual factors affect in-context processing. Apart from Gagné and Spalding’s work, no other empirical work has focussed explicitly on the question of interdependence versus independence.

**2.3.3. Evidence Against Current Views on Context**

Two views of conceptual combination in context have been described: the independence and interdependence views. Primary support is provided for the independence view by Gerrig and Bortfeld (1999), who believe researchers are getting a distorted view of conceptual combination by examining comprehension in isolation (i.e., with no surrounding context). Gagné and Spalding (in sub), however, have
provided some evidence to suggest that this is not necessarily the case. Furthermore, the interdependence view has been implicitly adopted by many researchers, who have examined conceptual combination out-of-context, in the hope that their findings would be transferable to in-context scenarios (Coolen et al, 1991; Wisniewski, 1996). However, there are questions over the preliminary empirical support for both views (provided by Gerrig and Bortfeld and Gagné and Spalding).

Regarding Gerrig and Bortfeld’s results, there are two central issues of concern. The first pertains to their interpretation of the results, while the second pertains to the suitability of the materials they have used. Gerrig and Bortfeld showed that their out-of-context factors do not interfere with the interpretation of compounds in context, and posited that this lack of interference provided evidence for an independence view. They concluded that different processes are utilised in context than out of context, and that out-of-context meanings simply do not influence in-context comprehension. However, as Gagné and Spalding point out, absence of interference does not necessarily mean that two sets of processes are needed for comprehension in and out of context. It could be the case that, out of context, there is competition between two interpretations and the response time merely indicates the speed with which the fastest interpretation can be selected. In other words, the role of context may be to provide an initial activation to the in-context meaning. This would suggest that interpretations still compete against each other, but the initial contextual boost for the intended interpretation will make it more likely for it to be selected. As Gerrig and Bortfeld’s results stand, this is an equally valid interpretation that does not require separate processes for in-context and out-of-context comprehension. This is discussed in more detail in the next chapter.
The previous issue refers to Gerrig and Bortfeld’s interpretation of their results. However, from the outset, there were possible problems with the materials they used. Firstly, their use of innovative meanings for compounds in the in-context condition is questionable. These innovative meanings are ones that people do not normally produce when asked to interpret the compounds. For example, the more usual interpretation for *doll smile* is “a smile on a doll”, whereas Gerrig and Bortfeld used the innovative meaning of “a smile made when someone is given a doll as a gift”. With innovative meanings, it is unclear how likely people are to produce them in an out-of-context situation (i.e., their *a priori* availability). It may be that any or all of these innovative meanings are similar to some possible out-of-context interpretations. If an innovative meaning were closer to a meaning normally produced out-of-context, then its comprehension in-context would be facilitated, possibly confounding the results that suggested no out-of-context interference. For example, Gerrig and Bortfeld used a matched pair of compounds *ghost doctor* and *artichoke doctor*. The innovative meaning used for these was “a doctor that went to a fancy dress as a ghost / artichoke”. It seems clear that the in-context meaning for *ghost doctor* is actually quite close to a possible out-of-context interpretation. This would facilitate people’s comprehension, thus speeding their response times. It would then appear to that no interference took place, which was not necessarily the case.

In addition, Gagné and Spalding also suggest that the degree of ambiguity in the compounds may have influenced response times. They argue that the Low Accessibility compounds (e.g., *tax lunch, baseball smile, blood sandals*) tended to have greater ambiguity than the High Accessibility ones. In other words, there were more plausible relations that could exist between the two. The High-Accessibility items tended to have a single dominant meaning (e.g., *family lunch, doll smile, water*
sandals), compared to several plausible interpretations for the Low-Accessibility items. Again, increased ambiguity may have slowed people’s responses, again making it seem like there was no interference taking place.

Finally, Gerrig and Bortfeld used both innovative and neutral stories within their context condition. While the innovative context clearly supported an innovative meaning, the neutral context appears to be more incongruous than unbiased. When the reader encounters the relevant compound it appears out of the blue, with no obvious link to the preceding context. In the neutral contexts for the ghost doctor / artichoke doctor compounds, the story described two doctors who were at a staff party and were nominated for “Best Staff Member”. The final, test sentence of the story was “We all agree that the winner is the ghost / artichoke doctor.” It is easy to see how readers could be confused upon reading this line. Thus, as Gagné and Spalding point out, it is not obvious what the response times in the neutral condition are actually measuring.

Regarding Gagné and Spalding’s work, there are also outstanding issues with the empirical work presented in support of the interdependence view. Unlike Gerrig and Bortfeld, Gagné and Spalding did not use innovative interpretations for their compounds. For their out-of-context tests they used a paradigm similar to that of Lynott and Keane (2002; see also Chapter 3, Experiment 2), where people made judgements on compounds paired with possible interpretations. For example, people saw “an office light is a light for an office” or “a home light is a light for a home” and made a judgement of appropriateness. The problem, however, lies more with the in-context conditions used by Gagné and Spalding. They used supportive contexts that were biased towards the intended meaning, by specifying the relation that exists between the head and modifier concepts. The neutral contexts omitted the line where
this relation was made explicit. However, Gagné and Spalding did not perform any additional controls for the context stories that they used. It could easily have been the case that there were unplanned contextual cues in the stories that lead people closer to one compound over another, thus facilitating people’s comprehension. If people were reading contexts that were more biased towards the high-frequency relations, then this would confound the finding that high-frequency relations were understood more quickly than low-frequency ones. Without clear controls, it is not clear what conclusions we can draw from Gagné and Spalding’s results.

In summary, it is clear that there are many questions yet to be answered by both the independence and interdependence camps. In Chapter 3, I have sought to address some of the issues raised above, with important extensions to the previously used experimental paradigms. Where previous work has used innovative meanings or experimenter-derived relations, I use interpretations normally produced by people. Where others have used different compounds with related meanings, I use the same compounds with alternative interpretations. I also incorporate an additional Mismatching context into my paradigm, which has not been used previously. Each of these factors allows for a more fine-grained testing of the independence / interdependence debate.

2.4. Theories of Novel Compound Production

In the previous sections, we detailed several competing theories of comprehension in conceptual combination. By contrast, the compound production aspect of conceptual combination has received scant attention. Despite people’s frequent and natural use of such novel word combinations, the research effort has been somewhat imbalanced. First, conceptual combination research has focussed
largely on the comprehension of novel compounds and not on their production. Second, many researchers have used compound production as a means of examining other phenomena, as opposed to compound production in itself. Third, the research that has specifically examined compound production has concerned itself primarily with developmental aspects of language, and not adult language usage.

This brings us to the third pillar of the thesis; the production of novel compounds. People create novel compounds all of the time – sometimes unwittingly, sometimes consciously. They play a central role in day-to-day communication and in extending the boundaries of language. People use novel compounds in a variety of scenarios; to label new objects (e.g., a genetically-engineered, crescent shaped fruit could be called *moon fruit*), to distinguish one object from other, similar objects (e.g., a *cactus beetle* is different from other beetles because it eats cacti), or as a shortcut to refer to a previously mentioned entity in a discourse (e.g., *appendix boy* referring to a boy who has had an appendectomy). In these ways, novel noun-noun compounds can increase communicative precision by filling lexical gaps (see Clark et al, 1985; Downing, 1977; Windsor, 1993), and making communication more efficient by replacing long phrases with shorter equivalents.

While there has been an abundance of theories of conceptual combination comprehension over the years (see Costello & Keane, 2000, 2001; Gagné, 2000, 2001, 2002; Murphy, 1984, 1988; Smith et al, 1988; Wisniewski, 1996, 1997a, 1997b), no overarching view of compound production has emerged. This oversight, of examining how people understand compounds rather than coin new ones, applies to both the empirical testing and actual modelling of compound use. So while there has been a large empirical literature on the comprehension of lexicalised (Levi, 1978; Marsh, 1984; Quirk, Greenbaum & Svartik, 1985) and novel nominal compounds
(Coolen, Van Jaarsveld & Schreuder, 1991; Costello & Keane, 2000, 2001; Gagne, 2000, 2001, 2002; Gerrig & Bortfeld, 1999; Murphy, 1990; Smith, Osherson, Rips & Keane, 1988, Wisniewski, 1997), the work examining compound production has been patchy and disjointed, with no consensual view emerging from this research (see Clark, Hecht & Mulford, 1986; Levi, 1978, Downing, 1977, Windsor, 1993).

One reason for the lack of consensus in the compound production literature is that researchers have been concerned with issues other than compound production per se. For example, Clark and colleagues were concerned with compound production as an aspect of language development, focussing on children’s errors in production. Both Levi (1978) and Downing (1977) debated the existence of a taxonomy of relations that could be used to describe produced compounds. Others have examined the use of irregular plurals in compounds as a means of questioning the notion of innate constraints governing morphological processing (Buck-Gengler, Menn & Healy, 2001; Gordon, 1985; Ramscar, Pearson & Ali, 2003). Once these ancillary issues have been set aside, the amount of work that looks at compound production itself is greatly reduced.

A final imbalance exists, in that much of this literature is concerned with child language development (Clark, 1981, 1982, 1983; Clark & Baron, 1988; Clark & Berman, 1984, 1987; Clark, Gelman & Lane, 1985; Clark & Hecht, 1982; Clark, Hecht & Mulford, 1986; Elbers, 1988; Windsor, 1983) rather than the production of new phrases by mature, adult language users. However, some of the findings from the developmental literature do have relevance for adult compound production, which we discuss below.
2.4.1. Evidence from Compound Production Literature

In this section, we look at the existing literature on compound production. We first explore the work that has taken a developmental view of compound production, which makes up the vast majority of this literature, followed by studies that have focussed specifically on adult compound production.

A recurring theme in people’s analysis of compound production is the role of prior knowledge in creating novel compounds. Two such examples can be found in the work of Windsor (1993) and Deutsch & Braun (cited in Elbers, 1988). Both looked at compound production with respect to labelling objects that were either typical or atypical with respect to prior knowledge. While both studies concentrated on children’s creation, Windsor also included an adult condition in her study.

Windsor used a picture-naming task that consisted of the experimenter eliciting descriptions for atypical and typical picture stimuli (what she calls inherent and non-inherent stimuli, respectively). Both children and adults were tested in one of two conditions; referential and non-referential. In the referential condition, the participant (speaker) and experimenter (listener) were separated by a barrier, and so the participant had to label pictures while the experimenter tried to guess which one they were referring to. In the non-referential condition, the participant and experimenter sat side by side, and the experimenter simply pointed to particular pictures and the participant provided a label. This meant either that both parties had visual access to the intended referents (non-referential), or they did not (referential). Windsor found that adults used approximately the same number of compounds for the referential and non-referential conditions, with the same level of compound variability in each (i.e., the same number of unique compounds). It was also found that adults
produced more compounds for atypical stimuli. Deutsch and Braun, who used a similar paradigm, also found that children produced more compounds for less typical items, but that this only occurred in the non-referential condition (i.e., when a referent was visible to both speaker and listener).

Even within these two studies, there are points of contention. Deutsch and Braun found that children produce more compounds for atypical stimuli, but only when both speaker and listener can see the intended referent. However, Windsor found that both adults and children consistently produced more compounds for atypical items, regardless of whether speaker and listener could both see the referent. Furthermore, both of these studies (and many others) use pictorial (i.e., perceptual) stimuli. While this medium is clearly more suited for experiments involving children, it ignores the kind of compound production that adults employ in text-only scenarios (e.g., in newspapers, books, magazines etc.), where there is no explicit perceptual information provided. Although many compounds do make reference to perceptual aspects of their referent, a visual stimulus is not required to do so. This methodological imbalance is something that needs to be addressed, which we consider in more detail in Chapter 4. It may be the case that the situations where perceptual information is readily available can be considered a subset of all cases where compound production occurs, with perceptual information simplifying people’s creation and comprehension novel compounds.

While Windsor has provided some adult-related data, other researchers (Clark and colleagues; Deutsch & Braun) have only examined conceptual combination from a developmental viewpoint. Thus, their conclusions may not be directly transferable to adult compound production. For example, it has been shown that there are differences in children’s abilities to create and to comprehend novel compounds.
Clark and Berman (1987) found that children (aged 3-7) were much better at comprehension than production, with young children having no difficulty in giving meanings to novel compounds with no surrounding context. Also, children can judge well-formed compounds before they can produce or make repairs to their own (Clark & Barron, 1988). Such asymmetries have not been evident in the adult language literature, which means that we should not assume that these developmental conclusions automatically apply to adult language usage. It may be the case that there are greater parallels between adults’ comprehension and creation of novel compounds than is the case with children.

From the developmental research outlined above, several factors emerge that influence children’s creation of compounds, including object typicality, whether speaker and listener have visual access to the referent, and also the age of the child. It remains an open issue as to whether these factors will also affect adult compound production. Furthermore, the preference for methodologies employing perceptually oriented stimuli does not give due consideration to people’s creation of novel compounds in everyday text-based situations. Costello (2002) did perform some preliminary investigation of the role of diagnostic features in compound production, using written descriptions. He found that, given a description of an entity that mentions one feature, people are much more likely to use diagnostic features to create compound labels. For example, for the description “a pig that has a trunk” and “a pig that is grey”, people are more likely to say *elephant pig* for the first description, as having a trunk is diagnostic of elephants, whereas being grey is not. Costello and Keane (2000, 2001) have shown that the three constraints of diagnosticity, plausibility and informativeness impact directly on the comprehension of novel compounds. It
may be that these and other constraints can be used to construct a more complete theory of adult compound production.

2.5. Review Summary

In this review, we have discussed the conceptual combination literature from the point of view of the three pillars of comprehension in isolation, comprehension in context and compound production. We have examined existing theories of comprehension, and considered how conceptual combination in context is related to conceptual combination out-of-context. Additionally, we described the existing research into the production of novel compounds. In each of these areas of conceptual combination research, several issues have emerged.

1. There are several competing theories that address the comprehension of novel compounds, yet each of them still has a number of outstanding empirical and theoretical issues to be resolved. Given that conceptual combination is so common in everyday cognitive life, a strong theory must be prepared to deal with the wide and creative variety of interpretations that people clearly arrive at with ease. In Chapter 3, we look at the first pillar of the thesis in people’s interpretation of novel compounds in isolation.

2. The research into the comprehension of conceptual combination has primarily looked at conceptual combination out-of-context, and none of the existing theories fully addresses the issue of in-context comprehension. Given that concept combinations in the real world are often encountered and comprehended in the presence of discourse context, the effects of context must be fully investigated and integrated into a theory before we can truly explain how concept combinations are
understood. Chapter 3 considers how out-of-context interpretations impact on people’s processing of novel compounds in context; the second pillar.

3. The small body of literature that has looked at the creation of novel compounds has been firmly developmental in its outlook, with only the very occasional nod to adult’s behaviour in this domain. Given that compounds must be produced if they are to be comprehended, the reasons for the lack of adult production literature are not clear, and this oversight must be addressed. In Chapter 4 we focus on the third and final pillar of the thesis; the production of novel compounds.

So while there are several prominent theories of conceptual combination, not only do they currently fail to address the inevitability of context effects, but the entire field of compound production has yet to be incorporated into these theories. In Chapter 5, we take a more holistic view of conceptual combination, drawing on a set of core high-level constraints to explain how people produce and comprehend conceptual combinations both in and out of context.
Chapter 3  Compound Comprehension

3. COMPREHENSION IN CONCEPTUAL COMBINATION

3.1. Introduction

People produce and understand novel compounds all the time. While novel compounds can occur with little or no contextual information (e.g., signs, newspaper headlines), many occur embedded in a surrounding context or discourse. Despite this, researchers have generally focussed on the understanding of compounds in isolation, that is, without the influence of pragmatic factors such as surrounding semantic contexts. Because of this, it is not clear how people actually understand novel compounds in-context, and what factors influence their ease of comprehension.

This chapter describes a series of experimental studies that investigates the relationship between of out-of-context interpretations of novel compounds and the in-context processing of these compounds; the first and second pillars of the thesis. It is well known that novel compounds give rise to multiple meanings (Costello & Keane, 1997, 2000; Kay & Zimmer, 1976; See also Experiment 1) and we examine whether the relative availability of these meanings impacts in-context processing. Stated briefly, we examine whether the patterns observed for out-of-context processing are carried over to in-context processing. This is an important issue for a number of reasons. Firstly, it will demonstrate whether there is a relationship between out-of-context comprehension in conceptual combination and in-context comprehension. Secondly, it will give us a clearer picture of the relevance of existing out-of-context
research and for comprehension in context. Thirdly, the role of meaning availability in comprehension may give us some insight into the production of novel compounds, which we return to in Chapter 4.

3.2. Context

The vast majority of work in conceptual combination has focussed on comprehension out-of-context (see Chapter 2). More recently Gerrig & Bortfeld (1999) have posited two possible standpoints for conceptual combination in context; the independence view and the interdependence view. Briefly stated, the interdependence view sees conceptual combination in context as being influenced by the processes normally engaged out-of-context, while the independence view suggests there are no such influences. In other words, the interdependence view would predict that the pattern of results for out-of-context conditions would still be evident in-context. By contrast, the independence view would predict that a different pattern would emerge.

Gerrig and Bortfeld have provided the only evidence in favour of an independence view of conceptual combination. They examined this hypothesis by looking for a pattern of interference from out-of-context meanings in-context. They postulated that certain compounds are easier to understand than others out-of-context. They constructed pairs of compounds, with one compound being responded to quickly out-of-context, while the other was responded to more slowly. If the same compounds were understood equally easily in-context, even when innovative meanings were ascribed to them (i.e., meanings that people would not normally arrive at), this would lend support to an independence view of conceptual combination. In other words, if the out-of-context ease of interpretation did not interfere with the
processing of innovative meanings in-context, then the in-context processing could be seen to be independent of out-of-context factors. Indeed, they found no pattern of interference and hence concluded that conceptual combination in-context was independent of conceptual combination out-of-context.

However, there are problems with Gerrig and Bortfeld’s approach, and with their interpretation of the results. Gerrig and Bortfeld found evidence of interference in context, and from this concluded that there must different process at play, and therefore comprehension in context was independent of comprehension out of context. This conclusion, however, seems premature. There are many possible explanations for the interference effect that would not require a specialised set of processes for in-context comprehension (see Chapter 2; Gagné & Spalding, in sub).

Additionally, there are issues with the methodology that Gerrig and Bortfeld have used, resulting in possible confounds they have not considered (e.g., use of innovative meanings not normally produced by people). Although Gagné and Spalding (in sub) have attempted to address some of these concerns (e.g., by using relational interpretations that people do produce) there are also concerns that have yet to be tackled. For example, Gagné and Spalding do not eliminate the possibility that surrounding contexts were biased towards particular interpretations.

We aim to address some of these oversights by incorporating three significant extensions to previous paradigms. Specifically, we use people’s actual out-of-context interpretations, use of the same target compound, and a new mismatching context condition.

Using People’s Actual Interpretations. The interpretations employed in the following experiments are based on the actual interpretations produced by
participants. The interpretations Gerrig and Bortfeld used were experimenter-derived, innovative meanings. Thus, these interpretations are not necessarily the same as those normally produced by people. As such, it is unclear how likely they are to be produced in the set of possible meanings for a given compound (i.e., their \textit{a priori} availability). Furthermore, Gerrig and Bortfeld did not look explicitly at the interpretations people produced out of context. Instead they examined only how easy / difficult a given compound was to understand. In this work, in contrast, by using the actual interpretations people produce for the compounds, it is easy to determine their availability relative to one another. Furthermore, in using the interpretations that people produce naturally there is no \textit{a priori} exclusion of any particular class of interpretation. For example, Gagné and Spalding use only relation-based interpretations whereas people often produce ones that are property-based.

\textit{Using the Same Compounds.} In previous studies, tests were made using \textit{different compounds with similar meanings} (e.g., in Gagné and Spalding’s work, \textit{office light} and \textit{home light} meaning “light-in-location-X”, or \textit{doll smile} and \textit{baseball smile} meaning “a smile in response to a particular gift” in Gerrig and Bortfeld’s work). In contrast, we used the \textit{same compound with different meanings} (e.g., \textit{skunk squirrel} meaning “a squirrel that smells bad” and “a squirrel with a black and white stripe on its back”). Most novel compounds have several interpretations that differ in their level of availability; for example, in \textit{skunk squirrel}, the squirrel-that-smells interpretation tends to be highly available relative to the striped-squirrel one (i.e., it is produced first and most often). Our approach allows us to manipulate the level of availability without having to alter the compound itself. This gives us a more sensitive test of comprehension differences, and avoids possible confounds due to the
use of different modifiers, such as syntactic or semantic differences, their word frequency or word length.

**Using a Mismatching Context.** The above changes permitted the introduction of a new context condition absent from previous studies, giving three context conditions in total. In this paradigm, a participant first reads the story context and then performs a comprehension test on the paraphrased interpretation of the compound. We have a *Matching* condition (where the meaning provided by the context matches the tested interpretation), an *Omitted condition* (where the context omits any reference to a particular meaning), and a new *Mismatching condition* (where the context provides one meaning of the target compound, but then tests comprehension for another possible meaning). For example, in the Mismatching condition, the story about the *skunk squirrel* might mention the malodorous-squirrel interpretation but then test comprehension on the striped-squirrel one. As we discuss at the beginning of Experiment 3, this extension of the paradigm usefully differentiates the two theoretical views.

In the following experiments, we test whether the differentially available, out-of-context meanings produced by people for a compound have some effect on understanding that compound in contexts that offer varying degrees of support. As we shall see, the interdependence view predicts a different pattern of results to those of the independence view (see Experiment 3 for details).

We first carried out two out-of-context experiments in order to show differences between specific interpretations for novel noun-noun compounds. In Experiment 1, we asked participants to generate interpretations for target compounds and these interpretations were then categorised as High- or Low-Availability based on the number of times they were produced by the group. In Experiment 2, out-of-
context comprehension response times for these interpretations were measured to determine whether production frequency was reflected in ease of understanding. In Experiment 3 we took each compound and varied both the availability of interpretations (High- or Low-Availability) and the context in which it was presented (Matching, Mismatching and Omitted conditions). We then examined the results from the perspective of both the independence and interdependence views.

3.3. Experiment 1 – Out-Of-Context Interpretations

Before we look at conceptual combination in-context we need to consider people’s behaviour in interpreting novel compounds out-of-context. This experiment collected participants’ out-of-context interpretations for novel noun-noun compounds and analysed the frequency of production of the interpretations. We then classify pairs of interpretations for each concept as either High- or Low-Availability, based on their frequency of production. We could then use this classification to examine whether these interpretations could also be differentiated using an out-of-context response time paradigm in Experiment 2.

3.3.1. Method

Materials and Design

A total of 40 compounds from two different sources (Set A and B) were analysed in this experiment. Set A consisted of 24 novel noun-noun compounds (e.g., carrot bomb, father card) selected from Gerrig and Murphy (1992). The compounds were randomly split into two groups of twelve compounds, and each participant received the compounds in a random order of presentation.
Set B consisted of 16 novel compounds (e.g., giraffe antelope, helicopter airplane) from Costello and Keane (2001). Participants in this experiment were not presented with these compounds, as interpretation data had been collected by Costello and Keane using the same procedure described here. Participants in both cases were of similar age and background. Materials for Experiment are in Appendix A.

**Participants**

Forty-five students at University College Dublin took part in this experiment. All participants were native speakers of English and were unpaid for their participation. Each participant was randomly assigned to a materials group from Set A (see below).

**Procedure**

For all 40 compounds, participants were told that they would be presented with a two-word phrase that they had never seen before and that they should write down “what they thought the phrase could plausibly mean”. If more than one meaning came to mind they were asked to write them down in the order in which they occurred. Participants were asked to write clear and specific meanings. One sample compound was given (not used elsewhere) along with some possible interpretations. Possible interpretations were a mix of both property- and relation-based interpretations.

**Results and Discussion**

Participants produced a total of 1350 interpretations for the 40 compounds. The minimum number of unique interpretations provided for any one compound was 14 (tulip sled), whereas the maximum provided was 25 (foot chip). We defined the *availability* of an interpretation as a percentage equal to the number of participants
that produced that unique interpretation over the total number of participants receiving its associated compound. For example, out of 24 participants that received *tulip sled* to interpret, 13 produced the interpretation “a sled shaped like a tulip”, giving this interpretation an availability of 62.5%. High- and Low-Availability interpretations were selected for each compound. A *High-Availability interpretation* for a given compound was the one most frequently produced for that compound, provided it was produced by more than 40% of participants. Its matching *Low-Availability interpretation* had to have a frequency that was less than or equal to half the frequency of the High-Availability interpretation (with a lower bound of 8% to avoid unique interpretations being used). Of the 40 compounds examined only 14 met these criteria. From this set, 12 compounds were randomly selected for use in subsequent experiments (See Appendix B for the High- and Low-Availability interpretations used). For these compounds, there were equal numbers of property- and relation-based interpretations. The mean frequencies of production were 52% for High-Availability interpretations (Maximum = 62.5%, Minimum = 42%) and 12.5% for Low-Availability interpretations (Maximum = 25%, Minimum = 8%). Thus, there was no overlap in the frequencies of production between any High- and Low-Availability items.

This experiment shows that there are significant problems inherent in finding a suitable material set to test the dependence views. Having analysed over 1300 interpretations to our 40 selected compounds, we found that only 35% (14) of them met the criterion of two interpretations that could be clearly separated by their availability.
3.3.2. Observations on Interpretations Produced

In analysing the interpretations produced by people, we also considered whether the conceptual class (i.e., artefact or natural kind) of the concepts themselves impacted on the variety of interpretations produced. Compounds were divided into 4 categories of combination – Artefact-Artefact (AA), Artefact-Natural Kind (AN), Natural Kind – Artefact (NA) and Natural Kind – Natural Kind (NN). In fact, we found no difference in the number of different interpretations produced for these compounds ($F < 1$). The mean number of interpretations for each type was 16.2 (AA), 18.14 (AN), 18.4 (NA) and 17.5 (NN). Comparing only AA and NN compounds revealed no difference either ($F < 1$). Overall, the conceptual class did not appear to influence the number of interpretations generated.

While every compound used in this experiment gave rise to a wide variety of interpretations, there was also much regularity evident. Even though people can be highly creative while interpreting novel combinations (e.g., a carrot bomb meaning “a good looking redhead”, or a vodka face meaning “a very pale Russian”), there is also a great tendency for people to arrive at the same interpretations. From the 40 compounds in this experiment approximately 50% of people produced one particular interpretation for each compound. For example, over 60% of people produced the interpretations “a carrot bomb is a bomb shaped like a carrot” and “an hour hammer is a hammer that strikes every hour”. This is surprising, since these are novel compounds which the participants have never seen before. Many people used diagnostic features of the modifying concepts in their interpretations, which contributed to people’s consensus in interpretations. For example, over 40% of people gave the interpretations “a giraffe antelope is an antelope with a long neck” and “an arrow window is a window shaped like an arrow”, both of which use
diagnostic features of the modifiers. These findings suggest that despite the huge number of possible interpretations, there is core knowledge that people generally utilise when interpreting these compounds, which allows them to converge on the most acceptable interpretations.

Another aspect of these results is that people displayed a tendency to produce quite concise interpretations, rather than detailed, elaborate ones. For example, a detailed description provided for ballet mother was “a distinguished Russian woman that teaches ballet”. This was produced by only one participant. On the other hand, the less detailed interpretation “a ballet mother is a ballet instructor” was produced by almost 40% of participants. In fact, people specified more than one feature or relation in less than 1% of cases.

Taken together, these findings suggest that, while conceptual combination has the capacity to be highly creative and detailed, there is a distinct tendency for people to arrive at the same, concise interpretations. Obviously, these factors need to be considered when theorising about conceptual combination, and indeed when developing computational models. We consider these issues in more detail when we discuss a model of conceptual combination in Chapter 5.

3.4. Experiment 2 – Online Out-Of-Context Interpretations

Experiment 1 yielded a set of compounds with their associated High- and Low-Availability interpretations. As the dependent measure for in-context experiments is the response time for understanding/verifying the meanings of compounds, we also needed to know whether these High- and Low-Availability interpretations can be differentiated by their out-of-context comprehension times. Experiment 2 sought to confirm that High-Availability interpretations were responded
to reliably more quickly than the Low-Availability ones. In other words, this experiment investigated whether the frequency of production of interpretations reflects their ease of comprehension in a paraphrase test.

3.4.1. Method

Materials, Apparatus & Procedure

Twelve compounds, each with 2 associated interpretations were used in this experiment (all from Experiment 1, see Appendix B). The interpretations were written to control for word frequency and syllable count, while retaining the core meanings originally proposed by participants in Experiment 1. There was no difference between the High- and Low-Availability interpretations for word frequency ($t < 0.1$; using Kucera-Francis, (1967)), syllable count ($t < 0.1$) and word count ($t < 0.1$).

These materials were randomly split into two groups, each of which contained 6 compounds with their High-Availability interpretations and 6 compounds with their Low-Availability interpretations. The two groups were designed so that no participant ever saw both the High- and Low-Availability interpretations of a given compound. Following a method used by Gerrig and Bortfeld (Experiment 3), eight filler-items were also used. These materials were novel compounds (rejected compounds from Experiment 1) paired with implausible interpretations (e.g., “a trumpet olive is a tall building”). These fillers were included to allow participants to answer True to all test items, but False to all filler items. The presentation of all items was randomised for each participant.
The experiment was run on a laptop computer that recorded responses and response times. Participants were seated in front of the monitor with their hands resting on the keyboard and used their dominant hand for the duration of the experiment. The J, K and spacebar keys were used to respond and the stimuli were displayed in the centre of the screen in standard upper- and lower-case type. Participants were presented with novel compounds on-screen and instructed to press the spacebar once they had thought of an interpretation. After pressing the spacebar, a paraphrased interpretation appeared to which the participant had to answer either True or False, using the J and K keys respectively. For test items, the paraphrases related to either the High or Low-Availability interpretation of the compound that had just appeared. After each True/False response, “Ready” appeared on screen for two seconds before the next compound was presented.

Participants

Twenty-four unpaid, native speakers of English studying at University College Dublin were randomly assigned to conditions of the experiment. Three participants were eliminated, prior to data analysis, for giving incorrect answers to more than 25% of the fillers; for the overall average only 5% of fillers were responded to incorrectly.

Measures

The measure used was the time taken for participants to verify a paraphrase of the interpretation presented after its associated compound. Trials that were more than 3 standard deviations from a participant’s mean response-time were removed as outliers, resulting in 1.1% loss of data. Analyses were carried out on the correct responses made in the experiment.
Chapter 3  Compound Comprehension

Results

Comprehension times were analysed using a one-way ANOVA, with Availability as a within-subjects factor. There was a main effect of availability, with the High-Availability interpretations being verified faster (M = 2643 ms) than the Low-Availability interpretations (M = 2943 ms), F1(1, 20) = 6.19, p = 0.007, MSe = 164805; F2(1, 11) = 3.83, p = 0.014, MSe = 197491. The possibility existed that syntactic differences between High- and Low-Availability interpretations might make the former easier to understand than the latter, thus explaining the response time difference. If the High-Availability items were in some way easier to understand, then we would expect lower error rates for these items. However, this was not the case. A one-tailed t-test analysis found that there were no differences in the error rates for High- and Low-Availability interpretations (p > 0.1), with High-Availability interpretations having an error rate of 32% while Low-Availability interpretations had an error rate of 30%. We also analysed the paraphrased interpretations for possible differences in the number of clauses. If High-Availability interpretations had fewer syntactic clauses, then this would obviously allow them to be processed more quickly. However, a t-test revealed that there was no difference in the number of clauses in High- and Low-Availability interpretations (p > 0.1).

Discussion of Experiments 1 and 2

These two experiments provide converging evidence that people consider the High- and Low-Availability interpretations of the compounds to be quite different. Experiment 1 shows that High-Availability interpretations are produced more often than the Low-Availability ones. Experiment 2 shows that, when presented in a paraphrase comprehension test, the High-Availability interpretations are understood more quickly than the Low-Availability ones. Undoubtedly, other tests could be done
to distinguish these two interpretation types. For instance, one could present the compound paired with one or other interpretation immediately, rather than using our two-step method of thinking of an interpretation, then judging the paraphrase. However, the two tests reported here provide sufficient convergent support to be confident that we have captured a key difference between these two interpretation types. Having shown that differences exist between High-Availability and Low-Availability interpretations, both in terms of their frequencies of production and people’s judgment times, we can now examine the effects of context on these \textit{a priori} availability differences.

### 3.5. Experiment 3 – Out-of-context Influences on In-Context Processing

Experiments 1 and 2 provide us with suitable materials to test the interdependence and independence views. We now have a set of novel compounds, each of which has an associated High-Availability and Low-Availability interpretation. These interpretations differ reliably in their frequency of production and in the speed with which people respond to them in an online judgment task. Thus, in this experiment, we can cross the variables of Interpretation Availability (High or Low) and Context (Matched, Mismatched or Omitted) to see whether out-of-context availability impacts upon people’s understanding of novel compounds in-context.

In the present experiment, participants first read a story context and then were asked to make a True/False judgement about a paraphrased interpretation for the novel compound (see Table 3.2 for sample materials). As in Experiment 2, the paraphrased interpretation used either the High- or Low-Availability interpretation for the compound. In the three context conditions, these paraphrases bore one of three
relationships to the preceding story context. In the Matched condition, the paraphrase reflected an interpretation explicitly mentioned in the story context. In the Omitted condition, the paraphrase involved an interpretation that was not explicitly mentioned in the story context – rather, the possible interpretation is first flagged in the paraphrase. In the Mismatched condition, the paraphrase used an interpretation with a different meaning to that presented in the story context; so, if the paraphrase used the High-Availability meaning then the story mentioned the Low-Availability meaning, and vice versa (see Table 3.1 for a schematic of the experiment).

The two proposals about the influence of context make different predictions for these conditions, which distinguish between them. We emphasise the predictions made for the availability manipulation, as this best distinguishes the two views. Briefly stated, the interdependence view expects that availability will matter, whereas the independence view does not.

<table>
<thead>
<tr>
<th>Context Condition</th>
<th>Paraphrase Judgement Task</th>
<th>High-Availability Judgement</th>
<th>Low-Availability Judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matching</strong></td>
<td></td>
<td>High-Availability Story</td>
<td>Low-Availability Story</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High-Availability Judgement</td>
<td>Low-Availability Judgement</td>
</tr>
<tr>
<td><strong>Mismatching</strong></td>
<td></td>
<td>Low-Availability Story</td>
<td>High-Availability Story</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High-Availability Judgement</td>
<td>Low-Availability Judgement</td>
</tr>
<tr>
<td><strong>Omitted</strong></td>
<td></td>
<td>Omitted Relation Story</td>
<td>Omitted Relation Story</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High-Availability Judgement</td>
<td>Low-Availability Judgement</td>
</tr>
</tbody>
</table>

Table 3.1 Schematic of the conditions used in Experiment 3.
### Table 3.2 Example Stories with High-Availability paraphrase judgements for Matched, Mismatched and Omitted Conditions

*" Indicates the key sentence modified in each condition. This was not marked for participants.

**Matched Condition**

John and James were out on the farm.
James had shown John most of the farm animals.
They went to the field where new types of horses were being bred.

* John noticed one horse had a large horn like a rhinoceros.
James was pleased that John liked the animals on the farm.
John asked “where did you get that rhinoceros horse?”
James gave John the full story.

Paraphrase judgement
A “rhinoceros horse” has a big horn.

**Mismatched Condition**

John and James were out on the farm.
James had shown John most of the farm animals.
Then they went walking in the fields with the horses.

* John thought one horse had very thick skin like a rhinoceros.
James was pleased that John liked the animals on the farm.
John asked “where did you get that rhinoceros horse?”
James gave John the full story.

Paraphrase judgement
A “rhinoceros horse” has a big horn.

**Omitted Condition**

John and James were out on the farm.
James had shown John most of the farm animals.
They went to the field where new types of horses were being bred.

* They both enjoyed getting out in the fresh air for a while.
James was pleased that John liked the animals on the farm.
John asked “where did you get that rhinoceros horse?”
James gave John the full story.

Paraphrase judgement
A “rhinoceros horse” has a big horn.
Figure 3.1 Predictions made by the (a) interdependence and (b) independence views for the conditions of Experiment 3
Predictions for the Matched Condition

In the Matched condition, both views agree in expecting the Low-Availability interpretations to be understood as quickly as the High-Availability ones (see Figures 3.1a and 3.1b), albeit for different reasons. For example, the striped-squirrel meaning for skunk squirrel (i.e., the Low-availability meaning) would be judged as quickly as the squirrel-that-smells meaning for skunk squirrel (i.e., the High-availability one) when the story context mentions their respective interpretations. For the interdependence view, this prediction follows because the Low-Availability interpretation is supported by the context and becomes as highly activated as the High-Availability one (see Figure 3.1a). For the independence view, this prediction follows for a different reason; the context supplies the meaning for the High- and Low-Availability interpretations in the same direct way without recourse to out-of-context meanings (see Figure 1b for predictions).

Predictions for the Omitted Condition

In the Omitted condition, the views diverge in their predictions of High- and Low-Availability differences because of the respective importance they give to either availability or context (see Figures 3.1a and 3.1b).

For the interdependence view, the High-Availability interpretation should be understood faster than the Low-Availability one (see Figure 3.1a). The High-Availability squirrel-that-smells meaning should be judged more quickly than the Low-Availability striped-squirrel meaning for skunk squirrel (i.e., the Low-availability meaning) given a non-specific story context. This prediction follows because, in the absence of specific contextual support for one interpretation or the other, the out-of-context availability differences should be persist in assessing the paraphrased interpretation.
For the independence view, the Omitted condition parallels that of the Matched condition; that is, the Low-Availability interpretation should be again understood as quickly as the High-Availability one (see Figure 3.1b). In this case, according to this view, the interpretation for the compound is encountered for the first time in the paraphrased sentence and is understood, as if, it were an additional sentence of the story. In this sense, it is understood in exactly the same way as the story sentences presented in the Matched condition. Hence, as context is king, the High-Availability interpretation should be understood as easily as the Low-Availability one, based on the information given in the paraphrase context. This process is neither helped nor hindered by the out-of-context availability of the interpretations.

Predictions for the Mismatched Context Condition

In the Mismatched condition, the two views again diverge in their predictions of High- and Low-Availability differences.

The interdependence view continues to predict that the High-Availability interpretation should be understood faster than the Low-Availability one. However, one would expect responses in this context condition to be a good deal slower than in the other context conditions, because people are being “sent the wrong way”. For example, if the story mentions the Low-Availability striped-squirrel meaning for *skunk squirrel* and then the paraphrase presents the High-Availability squirrel-that-smells meaning, people should be able to quickly verify the paraphrase (as the High-Availability meaning is also highly activated while understanding the story). In contrast, if the story mentions the High-Availability squirrel-that-smells meaning and then the paraphrase presents the Low-Availability striped-squirrel meaning, people
should verify the paraphrase less quickly (as the Low-Availability meaning is less highly activated while understanding the story).

For the independence view, again no difference would be predicted for the High/Low-Availability manipulation; though this condition would be expected to be slower than the other contexts due to people’s confusion (see Figure 3.1b). In this account, a given interpretation will be constructed from the story context and then a different interpretation is constructed for the paraphrase. As both of these interpretation-building events are made without recourse to out-of-context meanings, there is no reason to expect the High-Availability meaning to be any easier to understand than the Low-Availability one. Again, the process is neither helped nor hindered by the out-of-context availability of the interpretations.

3.5.1. Method

Materials and Design

The design was a 2 (Interpretation Availability – High or Low) X 3 (Context – Matching, Omitted or Mismatching) mixed design with Availability being a within-participants factor and Context being varied between participants. Three short stories (one for each context condition) were created for each of the 12 noun-noun compounds selected from Experiment 2, giving a total of 36 story contexts. Each story was 7-8 lines long, mentioning the compound at the second-last line, and contained one sentence (line 4) that was modified in the different conditions. For the Matched and Mismatched conditions, the sentence mentioned explicitly the interpretation for the head and modifier of the novel compound (using the appropriate High- or Low-Availability meaning). For the Omitted condition, the sentence continued the story without mentioning the head or modifier, or any aspect of the
interpretation. Each story was followed by either a High- or Low-Availability paraphrased interpretation, making a total of 72 material items (see Table 3.1 for sample stories and paraphrases. All materials for Experiment 3 are given in Appendix C).

The materials were then divided into 6 groups; 2 for each context condition (Matched, Mismatched and Omitted). Each group contained 12 items (6 with High-Availability paraphrased interpretations and 6 with Low-Availability paraphrased interpretations), chosen as per Experiment 2 to ensure that each participant did not see both the High- and Low-Availability interpretations for the same compound. Eight filler items were added, making 20 items total. All test items had plausibly true statements while the filler items had all implausible (i.e. false) statements (as in Experiment 3, Gerrig & Bortfeld, 1999; Glucksberg & Estes, 2000a). The presentation of items was randomised for each participant.

Participants

Thirty students from University College Dublin, all of whom were native speakers of English, were paid a nominal fee for their participation in the experiment. Three participants were removed prior to the analyses for answering more than 25% of the filler questions incorrectly.

Procedure

Participants were given instructions that asked them to read the stories and answer the True/False statements at the end of each story (i.e., the paraphrased interpretation). If they thought the statement could plausibly be true (i.e., not just with reference to the story they had just read) then they should answer “True”, otherwise they should answer “False”. Five practice stories with feedback were
presented before the experiment proper. All stories were presented one line at a time. When participants had read and understood each line, they pressed the spacebar to continue. At the end of each story, the phrase “Question to follow. Get ready.” appeared for two seconds, followed by the True/False statement. Participants used the J and K keys to agree or disagree with the statement. Once they had done this, “Press ‘Space’ when you're ready to begin the next story.” appeared. When they did so, “New Story” was displayed for two seconds before the first line of the next story appeared.

Measures

As with Experiment 2, the measure we used was the time taken for participants to verify the paraphrased interpretations at the end of each story. We also analysed the time taken for people to read the sentence containing the novel compound. Responses that were more than 3 standard deviations beyond each individual’s mean response time were eliminated as outliers; 0.6% of the data was eliminated on this criterion.

Results & Discussion

Overall, the results are more supportive of the interdependence view than the independence view. There was a main effect of interpretation availability, and also a main effect of context.
Response Times to True/False Judgements of Paraphrased Interpretation

A 2 x 3 ANOVA, with the within factor of availability (High- or Low-Availability) and between factor of context (Matched, Mismatched or Omitted), was carried out on the times taken by participants to judge the paraphrase stating the compound’s interpretation (See Figure 3.2). This analysis revealed a main effect of availability with High-Availability interpretations being verified faster (M = 3718 ms) than Low-Availability interpretations (M = 4415 ms) - F1(1, 24) = 12.597, p = 0.002, MSe = 1570047; F2 (1, 33) = 4.136, p = 0.05, MSe = 5221222. There was a main effect of context (F1(2, 24) = 4.739, p = 0.018, MSe = 5272370, F2(2, 33) = 5.651, p
= 0.008, MSe = 3030404), with the Matched context being fastest (M = 3795 ms),
followed by the Omitted (M = 4273 ms), and Mismatched contexts (M = 5437 ms).
There was no reliable interaction between context and interpretation availability
(F1(2, 24) = 1.985, p = 0.159, MSe = 1570047; F2(2, 33) = 0.649, p = 0.529, MSe = 5221222).

Although the interaction was not reliable, to check the predictions of the two
views we performed pairwise comparisons of the conditions. Breaking these results
down by Context condition revealed that the High-Availability items were responded
to faster than the Low-Availability ones in the both the Omitted (F1(1, 10) = 9.962, p
= 0.01, MSe = 986501; F2(1, 11) = 8.391, p = 0.015, MSe = 718616) and Mismatched
conditions (F1(1, 8) = 4.018, p = 0.0799, MSe = 3412998; F2(1, 11) = 2.681, p =
0.13, MSe = 7983313), but not in the Matched condition (F1(1, 6) = 0.841, p = .395,
MSe =435768); F2(1, 11) = 0.215, p = 0.652, MSe = 4341873) (see Figure 3.2 for
mean response times). This pattern is also visible when we examine responses by
materials in each condition. In the Matched, Omitted and Mismatched the percentage
of times the High-Availability version was responded to most quickly was 58%, 75%
and 83% respectively. In addition, when one examines responses for High-
Availability items only, pairwise comparisons using Bonferroni adjustments reveal
that none of the contexts are reliably different (ps>0.1). However, examining
responses for Low-Availability items only shows that Matched and Mismatched
conditions are different (Matched = 3992 ms, Mismatched = 6381 ms, p = 0.023).

Error rates for all conditions are shown in Table 3.3. A one-tailed t-test
analysis found that there were no differences in the error rates for High- and Low-
Availability interpretations, either as a main effect (p > 0.1) or in individual
conditions (all ps > 0.1). However, error rates did increase from Matched to Omitted to Mismatched, which was reliable using Page’s Trend test – L(2) = 28, p = 0.028.

Overall, this pattern of results supports the interdependence view and fails to support the independence view. There is a pronounced effect of Interpretation Availability, an effect that is not predicted by the independence view.

*Reading Times for Target Compound Sentences*

In the present experiment, we measured the response time to make a True/False judgement of a paraphrase containing the compound after reading the story. For the sake of completeness, we also report the results of the reading time of the sentence containing the novel compound. Overall, we find no effect of availability, with compound sentences from High-Availability stories (M = 2959 ms) being read no more quickly than compound sentences from Low-Availability stories (M = 2931 ms), or when these responses are separated by context condition (F’s < 1). These results reflect those of Gerrig and Bortfeld, which show the reading times of compounds within stories that support a particular meaning do not exhibit any effects of out-of-context availability. This shows that previously-used paradigms may not be sensitive enough to fully test for the influences of out-of-context understanding. Participants may not be adequately examining the meaning of the compound at the point of reading it in a story, but when they are forced to respond to paraphrased interpretations, they appear to explicitly address the meaning of the compound they have read.
<table>
<thead>
<tr>
<th>Context Condition</th>
<th>High-Availability</th>
<th>Low-Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching</td>
<td>2.9</td>
<td>6.7</td>
</tr>
<tr>
<td>Omitted</td>
<td>29.9</td>
<td>31.9</td>
</tr>
<tr>
<td>Mismatching</td>
<td>32.4</td>
<td>38.1</td>
</tr>
</tbody>
</table>

Table 3.3 Error rates for High- and Low-Availability interpretations in Matched, Omitted and Mismatched conditions

### 3.5. Experiments 4a and 4b – Bias Post-tests

An important consideration is that key results may have arisen due to some bias in the story materials that favours the high-availability interpretations. Firstly, it could be the case that the High-Availability interpretation was better supported by the story than the Low-Availability one; that some unspecified aspects of the story (such as contextual cues) predisposed people to accept the High-Availability meaning more readily than the Low-Availability one. Secondly, it could also have been the case that the paraphrases for the High-Availability were in some way better descriptions for the novel compounds than the paraphrases for Low-Availability interpretations. Both of these possibilities were examined in two post-test studies.

#### 3.5.1. Biases in Context Stories?

Following a procedure proposed by Gerrig and Bortfeld, we tested if there were possible biases in the context stories we used in Experiment 3, leading people to respond more quickly to one set of interpretations over another.
Method

Materials and Design

The materials used were the design were two sets of context stories used in Experiment 3; one set biased towards High-Availability interpretations, and the other biased towards Low-Availability interpretations. Thus there were 24 stories in all (12 High-Availability and 12 Low-Availability). Participants saw 12 stories in all, but never saw High- and Low-Availability versions of the same story.

Participants

Twenty-one undergraduate students from University College Dublin took part in this experiment. Participants had not taken part in any previous experiments, and were unpaid for their participation.

Procedure

The sentence in each story that contained the novel compound was highlighted. Participants read one version of each story and were asked to rate “how well the highlighted sentence fit in with the preceding context?” Participants marked on a seven-point scale - with 1 being “Fits poorly in context” and 7 being “Fits perfectly in context” - what their rating was.

Results

There was no reliable difference in the rated goodness-of-fit between the stories involving the High-Availability meanings (M = 4.69) and Low-Availability meanings (M = 4.47) in either by-participants or by-items analyses; F1(1, 20) = 1.155, p = 0.295, MSe = 2.694, F2 (1, 11) = 0.742, p = 0.408, MSe = 3.546). This finding suggests that there was no a priori additional support present in the story for the High-Availability meaning.
3.5.2. Biases in Paraphrased Interpretations?

To test the possibility that one set of paraphrases was, in some way, better suited to the associated novel compound, participants were presented with compounds and one of the possible paraphrased interpretations (High-Availability or Low-Availability) as descriptions. Participants were asked to judge how well the descriptions (interpretations) captured the relevant novel compound.

**Method**

*Materials and Design*

The materials used were the High- and Low-Availability paraphrased interpretations used in Experiment 3. There were 24 judgements used in total; 12 of High-Availability and 12 of Low-Availability. Participants saw twelve judgements, but never saw both High- and Low-Availability versions of the same compound.

*Participants*

Nineteen undergraduate students from University College Dublin took part in this experiment. Participants had not taken part in any previous experiments, and were unpaid for their participation.

*Procedure*

Participants read one interpretation of each compound. Each interpretation was phrased as a description of the item (e.g., A wall made using pieces of furniture is a sofa wall), with the novel compound highlighted in each case. They were asked to rate “how well the description describes the highlighted item?” Participants marked on a seven-point scale - with 1 being “Describes the item poorly” and 7 being “Describes the item perfectly” – what their rating was.
Chapter 3  Compound Comprehension

Results

There were no reliable differences between the High-Availability (M = 4.96) and Low-Availability (M = 4.53) paraphrases, either by-participants (F(1, 18) = 2.339, p > 0.1, MSe = 4.503) or by-items (F(1, 11) = 2.595, p > 0.1, MSe = 3.685) analyses. These post-tests confirm that the findings of Experiment 3 were not confounded by some a priori bias in the stories or paraphrases.

3.6. Overview of Results

Based on the key predictions of the independence and interdependence views, our findings show that a strong view of independence is not tenable. While we found no differences in responses for sentences containing the novel compound, we found that the paraphrase judgements of those interpretations that were responded to quickest out-of-context (High-Availability interpretations) were still processed more quickly in context. However, a priori interpretation availability was not the only factor that influenced in-context processing, with the type of context people were reading also impacting on response times. People responded more quickly in Matched conditions, with Omitted and Mismatched conditions being responded to more slowly. While an independence view might predict these effects of context, the availability effect conflicts with the central tenets of this view.

A further examination of the availability results showed that High-Availability response times were not reliably separable across conditions, while Low-Availability ones were. This suggests that interpretations that are normally highly available out-of-context are less susceptible to the effects of context than interpretations that are less available out-of-context. This too strengthens the case for interdependence, with
a clear gradation of context effects dependant on the *a priori* availability of interpretations.

The question must be asked as to how these findings sit with previous results. Obviously, we have found evidence for an interdependence view of conceptual combination, while Gerrig and Bortfeld were pro-independence. However, Gerrig and Bortfeld’s results could also be considered with a more interdependence-oriented interpretation. Their conclusion was based on the fact that compounds that were understood quickly out-of-context were not understood slowly in-context, when an innovative meaning was attached to it. In other words, no evidence of interference means no evidence of interdependence. What their results actually showed was that those compounds that were processed quickest out-of-context were still processed quickest in-context. Such a finding could easily be construed as evidence for an interdependence between out-of-context and in-context processing – compounds that are understood quicker out-of-context are understood quicker in-context, even if an innovative meaning is attached to that compound. Thus, it is easy to see how our findings can sit neatly alongside those of Gerrig and Bortfeld. Similarly, the work of Gagné and Spalding (in sub) seems to suggest an interdependence viewpoint is more reasonable, although their materials require a more rigorous examination to eliminate possible biases due to context, or other factors.

### 3.7. Conclusions

These results demonstrate that a strong view of independence cannot hold sway. Even if context does influence the understanding of novel compounds, each compound also brings its own conceptual baggage to each scenario, which we have shown impacts processing in different ways. The availability of particular knowledge
out-of-context can still influence people’s processing in-context. Indeed, it is perhaps not useful to consider the question of contextual influences as a case of independence or interdependence, rather as a complex interaction of out-of-context availability of interpretations and the degree of contextual support provided. Where contextual support is strong, out-of-context availability effects will be reduced, as in our Matched condition. Where contextual support is weak, as in the Omitted and Mismatched conditions, out-of-context effects will be most evident.

Furthermore, the issue of interpretation availability may provide some important insights into the issue of novel compound production. It is likely that the availability of particular meanings could affect people’s choice of compounds, depending on how available their intended meaning is in the set of possible interpretations of that compound. This issue is returned to in Chapter 5, where we outline a theoretical account of compound comprehension and production, and consider how to incorporate context in the modelling of conceptual combination.
4. COMPOUND PRODUCTION IN CONCEPTUAL COMBINATION

4.1. Introduction

In Chapter 3, we focussed on the first two pillars of conceptual combination: comprehension in isolation and in context. In this chapter we turn our attention to the third and final pillar; compound production. People create novel compounds all the time – to describe new technologies, to fill lexical gaps, to contrast between existing concepts; there are an infinite number of opportunities where people exercise their creativity by coining neologisms and novel compounds. Despite the ubiquity of this phenomenon, little is known about it. As we have discussed in the review section (Chapter 2), much energy has been expended on comprehension research in conceptual combination, but far less on production.

In this chapter, we focus on the role of plausibility in compound production. The importance of plausibility in the comprehension of novel compounds has been advocated by Costello and Keane (2000) in their Constraint Theory. The central role of plausibility in comprehension raises the issue of how plausibility might affect the creation of novel compounds. If someone encounters some object or scenario that differs in some way to what they are used to (e.g., they might see a beetle that has yellow and black stripes), they will often produce a novel compound to convey this information (e.g., by describing it as a bee beetle). Can the plausibility of information
being conveyed impact people’s word choice? Or on how well a compound expresses its intended meaning? Answers to these questions have not been forthcoming in the literature, and form the core of the issues presented in this chapter. We first outline how the plausibility might impact compound production with some specific examples. This is followed by a series of experiments in which we examine how varying levels of plausibility impact differentially on people’s production of novel compounds.

4.2. Plausibility in Compound Production

Different views of novel compound usage have been proposed, with some researchers taking the viewpoint that novel compounds always serve a communicative function, such as filling lexical gaps (see Eve Clark and colleagues). Others have stated that this is not necessarily the case, with compounds sometimes being used when they do not serve to increase communicative precision (see Elbers, 1988; Windsor, 1993). However, none of these accounts considers the potential role of plausibility in people’s compound production.

As plausibility is an important aspect of understanding novel compounds, it is a distinct possibility that it impacts in some way, on the creation of novel compounds. Clark and Clark (1979, p786) maintain that a speaker must have good grounds for believing that a listener can, on the basis of mutual knowledge, identify a conversational referent uniquely. In this case, a speaker and listener have access to mutual knowledge, which should be plausible to both parties for the speaker to successfully communicate their intended meaning using a novel compound. Obviously, if the information the speaker wishes to convey is highly plausible they should have no difficulty in producing a label, and be happy that it adequately conveys the correct meaning. But, what happens when the information that a person
wishes to convey is not as plausible as it might be? How will this affect the compound labels they create?

In the next section we introduce a paradigm for investigating compound production which we use in a series of experiments focussing on the affects of plausibility. These experiments provide a clearer picture of the role of plausibility in creating novel compounds.

4.3. A Paradigm for Compound Production

Vendler (1967) describes the process of compound production as "packing a sentence into a bundle that fits into other sentences". One can characterise this process as finding some minimal set of words, usually selected from the previous discourse, that will refer to some complex event/entity described in the discourse. For example, in Philip K. Dick's passage (see below), the words “Runciter” and “money” are selected to refer to “money on which the head of someone called Runciter is suddenly appearing”.

‘I have a five-poscred note,’ John Ild said, ‘with a beautiful steel-engraving portrait of Mr. Runciter.’… Al said, ‘I’ve got two of them. Already. Who else?’ He looked around the table. Six hands had gone up. ‘Eight of us,’ he said, ‘have what I guess we should call Runciter money.’


Computationally, we could imagine a process that finds candidate noun pairs from the preceding discourse, and assesses each candidate in some way. Any such selection mechanism needs to be guided by pragmatic constraints, as the words selected need to convey the intended meaning unambiguously. The use of words from a shared, previous discourse provides some common ground for understanding
the compound, but this in itself is not enough. Another key aspect is the existence of the knowledge about what the combined words mean. For example, the shared knowledge that dollar bills can be referred to by the person depicted on them (e.g., “I have an Abe Lincoln here that wants to meet you”; “clutching a wad of Benjamins”) must partly support the choice of “Runciter” in the compound *Runciter money*. Thus, the phrase *Runciter money* works because it describes a plausible entity (i.e., money with a picture of someone’s face on it). Consider an illustration of this point.

Imagine being asked to provide a “shortened phrase” to refer to the following description:

a) A wine that is made from grapes and contains alcohol.

Intuitively, most people would be happy using the single word, *wine*. The description in a) is highly plausible to us, since our prior knowledge tells us that wine is typically made from grapes and is alcoholic. So, longer phrases, like *grape wine* or *alcoholic wine*, convey no additional information. However, if the plausibility of what is being described is altered, or is violated in some way, longer phrases are needed. For example, b) is most likely to be described as *apricot wine* and c) as *non-alcoholic wine*.

b) A wine made from apricots that contains alcohol.

c) A wine made from grapes that contains no alcohol.

These examples show how plausible, prior knowledge constrains the words selected for the compound. However, there may be times when it operates against a given word choice. For example, using *potato wine* for d)

d) A wine made from potatoes that contains alcohol.
does not seem as good as *apricot wine* is for b). Why? Our prior knowledge suggests that wine is made from soft fruits/berries, whereas strong liquors tend to come from vegetables / cereals. Thus, *potato wine* may inadequately convey the intended meaning, and so may be misinterpreted by others. Another compound, like *potato alcohol*, would appear to be a better choice because it avoids the misleading interpretations that *potato wine* might invite. Several predictions follow from this account of compound production, what we shall call the Confidence and Convergence Predictions.

**The Confidence Prediction**

The plausibility of a description should affect people’s confidence in their created compounds. If the description is highly plausible then people should be more confident that the created compound fulfils its communicative intent. For example, they will be confident that *apricot wine* is a good label for b). The confidence issue relates back to the notion of meaning availability which we discussed in Chapter 3. If a description is plausible, then it is more likely to be a candidate interpretation for the newly created compound. In other words, a plausible description will be more available as an interpretation, compared to a less plausible description. By virtue of this, the compound producer will exhibit greater confidence in the label they have produced. In contrast, if a description is considered less plausible, people should have less confidence in their chosen compound. In our previous example d), people should have less confidence that *potato alcohol* or *potato wine* are good compounds for that description. In other words, because the description is likely to be less available as an interpretation for those compounds, the compound producer will be less confident that their compound will convey the correct meaning.
Chapter 4

Compound Production

*The Convergence Prediction*

If plausibility constrains the selection of a compound’s terms, and affects people’s confidence in their compound choice, then it is likely that it will impact on the variability of compounds produced; we call this *convergence*. If there is a high degree of plausibility then the overall convergence in created compounds should be high (i.e., the number of unique compounds produced by people should be low). So, a description like b) will exhibit increased convergence, with possibly only a single compound being produced by different people to label it (e.g., *apricot wine*). In other words, increased plausibility will aid in people converging on particular labels for such descriptions. In contrast, reduced plausibility in the description will result in reduced convergence. For instance, a description like d) will give rise to a variety of compounds (e.g., *potato alcohol, vegetable wine, tuber wine*). But why should this occur? If people are faced with a less plausible description, we have suggested that this will affect their perceived confidence in the compounds they construct. If their confidence is reduced, it is likely that people will attempt to come up with other candidate labels to increase the chances of communicative success. Thus, in considering more labelling options people will produce a greater variety of compounds across a population. In this way, the plausibility of a description links a compound producer’s level of confidence, the availability of particular meanings and the degree of convergence (or divergence) on particular compounds.

Some of the existing literature indirectly supports these predictions. First, they are consistent with Costello & Keane’s (2000, 2001) plausibility constraint, used to explain compound comprehension. Briefly, this constraint asserts that interpretations of novel compounds supported by prior knowledge will be considered more plausible, and therefore more acceptable. For example, a *shovel bird* could be a bird with a beak...
shaped like a shovel, as it is known that birds can be have beaks shaped like various objects. The availability of this prior knowledge means that this would be viewed as a good interpretation. By contrast, “a bird that uses a shovel” would be a poorer interpretation for the same compound. There is no available knowledge to suggest that birds use tools in such a way, making this less plausible. Second, several developmental studies have found that children are more likely to produce novel compounds in response to stimuli that are less typical (see Windsor, 1993; Deutsch and Braun, cited in Elbers, 1988). Additionally, Windsor found that there was no difference in the variability of compounds produced (i.e., convergence) in two different tasks (i.e., referential and non-referential tasks. See Chapter 2), but she reports no findings on differences in convergence due to varying degrees of plausibility.

4.3.1. Outline of Experiments

We first establish a set of materials (i.e., descriptions) that differ in terms of their plausibility. In two experiments, we used these descriptions to elicit novel compounds from participants. In a final compound production experiment, participants were asked to create labels for a different set of descriptions, which were derived from interpretations given for novel compounds in an earlier comprehension experiment. Our aim is to show that plausibility is an important consideration for in the production and assessment of novel compounds.

4.4. Experiment 5 – Establishing Different Levels of Plausibility

Before testing the affects of plausibility on people’s creation of noun-noun compounds, we need to establish a suitable set of materials. Initially, pairs of
descriptions were constructed where one of each pair was considered highly plausible (e.g., A box that contains files belonging to lawyers) and the other less plausible (e.g., A box that contains sandwiches belonging to lawyers). This original classification was based only on intuitive judgments of the materials, so a more empirical basis for this distinction was needed. Experiment 5 confirms a difference between these pairs of descriptions by asking people to rate the plausibility of each description. If our intuitions are correct about the materials, we should see a clear separation in people’s plausibility ratings of the High-Plausibility and Low-Plausibility descriptions.

Method

Participants

Twenty-two native English-speaking undergraduates at University College Dublin took part in the experiment. One participant's data was not used as he/she failed to complete the questionnaire.

Materials

The materials consisted of 44 pairs of descriptions designed to have different levels of plausibility. Each description consisted of a subject followed by two objects that were linked using a variety of relations (e.g., made from, used for, causes). All descriptions followed this general form, though they did not have identical syntactic structures (see Appendix D for all materials). Examples in Table 4.1 show how the High and Low-Plausibility items differed through the alteration of just one word.
Example Description Plausibility

1  A bruise on a finger caused by a pencil.   High
    A bruise on a toe caused by a pencil.   Low

2  A game played by children on a street.   High
    A game played by children on a roof.   Low

Table 4.1 Example descriptions used in Experiments 5, 6 and 7.

Design and Procedure

Each description in a pair was randomly assigned to one of two questionnaires, giving 44 test items for each questionnaire. These were interspersed with 15 filler items that were of similar form but were tautological in nature (e.g., a bicycle that has two wheels and a saddle). Only one item appeared on each page, and each booklet was randomised for each participant. The experiment had a single factor design with Participant Ratings as the dependent variable, and Plausibility as an independent within-subjects variable. Participants were asked to read each description and rate how plausible they considered each one. The rating scale ranged from 1 (“not plausible at all”) to 7 (“completely plausible”).

Results and Discussion

Participants’ ratings of the materials supported the prior classification for all item pairs. High-Plausibility items (M = 5.451) had higher plausibility ratings compared to the Low-Plausibility items (M = 3.563). An ANOVA of the participants’ ratings revealed a main effect of Plausibility, treating participants and items as random factors, F1(1, 20) = 312.640, MSe = 2.591, p < 0.0001; F2(1, 43) = 126.4, MSe = 6.259, p < 0.0001. The mean difference between the pairs was 1.88, with the greatest difference being 4.1 and the smallest difference being 0.13. As an additional precaution, the descriptions were analysed for differences in word frequency between
the key nouns that were altered. Kucera and Francis’s (1967) frequency counts were used for this calculation. A two-tailed t-test revealed that there was no difference between the High- and Low-Plausibility items – \( t < 1, p = 0.34 \). From these results we now have a suitable material set to examine the effects of differential levels of plausibility in eliciting nominal compounds from people.

4.5. Experiment 6 – Producing Novel Compounds 1

In this experiment, we tested the convergence and confidence predictions (outlined previously) by presenting people with descriptions that were either of High or Low-Plausibility. People were asked to provide a shortened phrase to capture each description they were given. In this way, Experiments 6 and 7 reflect the production of compounds observed in Dick’s *Runciter money* example, and in the examples outlined earlier in this Chapter (e.g., *apricot wine*). Our measures were based on classifications of the coined phrase and people’s confidence ratings of the labels they produced.

**Method**

*Participants*

Twenty-four native English-speaking undergraduates at University College Dublin participated. Two participants’ responses were removed prior to analysis as they failed to complete the questionnaires.

*Materials*

Forty-four pairs of descriptions were used (from Experiment 5). Each pair consisted of one High-Plausibility description and one that was Low-Plausibility. The full set of materials is given in Appendix D.
Design

The description pairs were randomly split to create two lists of materials containing 22 High-Plausibility and 22 Low-Plausibility items. No list contained both versions of a description-pair, but all items occurred equally often. Also included were 15 filler items that were tautological in nature. These were descriptions that people could adequately describe using only one word, whereas test items could not easily be described by a single word. The experiment had a single factor design with Plausibility as an independent within-subjects variable, and Participants’ Ratings and number of unique compounds per description as dependent variables.

Procedure

Participants were given instructions to “provide a shortened phrase that conveys the same information as each description”, making sure that they wrote down what they thought was the “best” phrase for each. They were also asked to rate how well they thought their label conveyed the information in the description. The rating scale ranged from 1 to 7, with 1 being "very good" and 7 being "very poor" (i.e., the lower the score, the better they considered their compound). The instructions provided an example description (not used elsewhere) and possible labels, along with an example of how to use the rating scale. Although people were not asked specifically to produce noun-noun compounds, the examples used were noun-noun constructions.

Scoring

The labels produced were categorised either as "Noun-Noun Compounds" or as "Other". The "Other" category included adjective-noun labels (e.g., “strange monkey”), verb-noun labels (e.g., “pecking bird”), single-word labels (e.g., “beach”), blends of two or more words and non-words (e.g., “binmenitis”). Of the 340 unique
noun-noun compounds created by participants less than 1% were lexicalised. This demonstrates that people do not feel restricted when they are required to produce novel compounds. Compounds were considered lexicalised if they appeared in Collins 21st Century Dictionary. Both the experimenter and supervisor (Prof. Keane) agreed 100% in their independent classifications of responses.

<table>
<thead>
<tr>
<th>Description: A pet owned by sailors found on a ship</th>
<th>Compound</th>
<th>Frequency of Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ship Pet</td>
<td>33%</td>
<td></td>
</tr>
<tr>
<td>Sailor Pet</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>Ship Cat</td>
<td>16%</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2 Compounds produced for one description and percentage of participants that produced them

Results

There were a total of 953 responses made, of which a high percentage (77%+) were noun-noun compounds (see Tables 4.3 and 4.4). The convergence prediction was tested by noting the total number of unique compounds given to the items in the High and Low-Plausibility conditions. The prediction was confirmed by a one-tailed, pairwise t-test showing that fewer unique compounds were produced in response to High-Plausibility items (M = 3.093) compared to Low-Plausibility items (M = 4.023), t(43) = 2.924, p = 0.003. Obviously, convergence is affected by the number of participants that saw each description; however, as this was equal for High- and Low-Plausibility items we were able to use the raw frequency counts. Table 4.2 gives an example of a description and the compounds participants produced, along with the frequency of production for each compound, while Table 4.5 gives sample compounds for High- and Low-Plausibility versions of a description.
Chapter 4  Compound Production

Experiment 6  Experiment 7  

<table>
<thead>
<tr>
<th>Total Responses</th>
<th>953</th>
<th>840</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Noun-Noun Compounds</td>
<td>741 (77.75%)</td>
<td>683 (81.31%)</td>
</tr>
<tr>
<td>Total Lexicalised Compounds</td>
<td>2 (0.6%)</td>
<td>4 (1.65%)</td>
</tr>
<tr>
<td>Total Unique Compounds</td>
<td>307</td>
<td>230</td>
</tr>
</tbody>
</table>

Table 4.3 Summary of Responses from Experiments 6 and 7

The confidence prediction was tested using a one-way repeated measures ANOVA, to analyse people’s confidence ratings, with Plausibility (High / Low) as a within-subjects factor. This analysis revealed a main effect of Plausibility on participants' confidence ratings, with participants being more confident about their coined compounds for High-Plausibility descriptions (M = 3.014) than for Low-Plausibility ones (M = 3.648). This finding was reliable, treating participants and items as random factors, F1(1, 21) = 33.599, MSe = 2.133, p < 0.0001, F2(1, 43) = 23.248, MSe = 2.911, p < 0.0001.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Plausibility</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Mean Confidence Ratings</td>
<td>3.014</td>
<td>3.648</td>
</tr>
<tr>
<td>Mean No. of Unique Compounds</td>
<td>3.093</td>
<td>4.023</td>
</tr>
<tr>
<td></td>
<td>2.736</td>
<td>3.316</td>
</tr>
<tr>
<td>% Noun-Noun Compounds</td>
<td>72.48</td>
<td>82.5</td>
</tr>
<tr>
<td></td>
<td>75.48</td>
<td>86.67</td>
</tr>
</tbody>
</table>

Table 4.4 Summary of responses for Experiments 6 and 7

There was a distinct tendency for people to coin their phrases using the exact words in the sentence description. Over 80% of compounds used the subject of the description as the head of the compound, with 62% of all noun-noun compounds coined drawing both words from the description. Overall, 78.6% of the nouns used in all compounds were from the descriptions. When people did not use the exact words in the description they often used near-synonyms (e.g., "educator" for "teacher"),

103
semantically related terms (e.g., "army" in place of "war", "sun" in place of "desert") or nominalised verbs from the description (e.g., "jumper" from "jumps", "pecker" from "pecks"). Interestingly, people were more confident about the goodness of phrases that contained these new words (M = 3.15), than they were about ones that contained only words from the description (M = 3.36), F1(1, 21) = 13.292, MSe = 1.594, p < 0.001; F2(1, 43) = 0.595, MSe = 2.039, p = 0.44. Though this result is not reliable in the by-items analysis, it does suggest that when people selected new words for their compounds, they were driven by the need to find terms that more accurately conveyed the meaning of the description. For example, for the description “a pencil carved from wood and used for art”, both art pencil (Existing words) and artist pencil (New Words) were produced four times. The mean confidence rating for the New Words was 2.5, compared to 2.75 for the Existing Words.

These results provide support for the confidence and convergence predictions. However, in Experiment 5 (where test items were originally classified as being of High or Low Plausibility), some item pairs were not reliably different from each other, which may have confounded our results. We repeated our analysis, removing items whose mean difference in plausibility ratings was less than 1 for the High- and Low-Plausibility versions, which resulted in a loss of 9 items. Removing these items did not affect the results, with convergence and confidence predictions still being borne out (All p’s < 0.01).
Table 4.5 Compounds created in response to the High-Plausibility and Low-Plausibility versions of a description.

4.6. Experiment 7 – Producing Novel Compounds 2

Experiment 6 introduced a paradigm for eliciting novel compounds from participants. We used simple descriptions and asked people to produce new labels for each description. We found that differences in plausibility for the descriptions gave rise to different behaviour over the population of participants. Firstly, there was increased convergence in the compounds coined for High-Plausibility descriptions (i.e., fewer unique compounds were produced). Secondly, people exhibited greater confidence for compounds that were produced in response to High-Plausibility items. However, with the materials that we used there was always the possibility that people were not considering the choice of terms very deeply, and were simply reading off the words as they occurred in the descriptions. It may be that the word / clause order of the descriptions was the driving force behind people’s word selection, and not the knowledge associated with the relevant concepts. To remove this possible explanation, we re-ran the previous experiment, but with the nouns in the sentence descriptions re-ordered, while retaining the meaning of the original descriptions. For
example, an item of the form i) in Experiment 6 was re-ordered to be item of the form ii) in this experiment:

i) An X done by a Y on a Z (e.g., A game played on a street by children)

ii) An X on a Z done by a Y (e.g., A game played by children on a street)

So, although the word order has been changed, the constituents of the description and its meaning remain constant across the two experiments. If people are constructing phrases based on not on the order of the words, but on the knowledge underlying these descriptions, then the compounds created here should be identical to those in Experiment 6.

Method

Participants, Materials and Design

Twenty native English-speaking undergraduates at University College Dublin participated in the experiment. Thirty-eight pairs of descriptions based on those from Experiment 6 were used. In altering the word order of the descriptions, 6 items were judged to be nonsensical and omitted from the test materials in Experiment 7. Both the experimenter and supervisor judged five items for omission; four were common to both, with one unique item each. Apart from the change in the number of materials, the design and procedure were the same as in Experiment 6.

Scoring

As in Experiment 6, the phrases coined were independently categorised by the authors as either "Noun-Noun Compounds" or as "Other". There was 100% agreement in this classification. Of 230 unique noun-noun compounds produced, 1.6% were lexicalised (see Tables 4.3 and 4.4).
Results

The main findings of Experiment 6 were replicated, re-confirming the convergence and confidence predictions. People produced a total of 840 compounds to the presented descriptions, of which a high percentage (81%+) were noun-noun compounds (see Tables 4.3 and 4.4). The convergence prediction was confirmed by a one-tailed, pairwise t-test which showed that fewer unique compounds were produced to High-Plausibility items (M = 2.736) compared to Low-Plausibility items (M = 3.316), t(37) = 2.136, p = 0.0197.

The confidence prediction was tested using a one-way repeated measures ANOVA, to analyse the people’s confidence ratings, with Plausibility (High / Low) as a within-subjects factor. This revealed a main effect of Plausibility on participants' confidence ratings, with people being more confident about their coined compounds for High-Plausibility items (M = 3.265) than for Low-Plausibility ones (M = 3.633) – with lower scores representing greater confidence. This finding was reliable, treating participants and items as random factors, F1(1, 19) = 7.076, MSe = 1.710, p = 0.013; F2(1, 37) = 6.333, MSe = 2.888, p = 0.016.

Finally, we looked at whether people’s responses in both experiments were essentially the same – if responses were very different compared to those of Experiment 6, the word order of the descriptions could be seen as the a driving force in people’s compound production. In comparing both sets of compound, we found that in only 1.7% of cases were novel compounds produced in Experiment 7 that had not been produced in Experiment 6. We looked more specifically at the most produced compounds for each description (the top two responses account for almost 70% of compounds created). We found that in 72.4% of cases the top two rankings were the same for Experiments 6 and 7. In 94.7% of cases, one of the top two phrases
from Experiment 7 was ranked equivalently in Experiment 6. Finally, when the responses of Experiments 6 and 7 were collapsed, both the convergence and confidence predictions were reliable (ps < 0.05). The high overlap of compounds produced in both experiments suggests that people were not simply reading off compounds, but were giving consideration to the relevant concepts in the descriptions. Moreover, the replication of the main findings of Experiment 6 offer further endorsement for the consideration of Plausibility as a factor in creating novel compounds.

4.7. Observations on Compound Production in Experiments 6 and 7

We found in both Experiments 6 and 7 that increased Plausibility gave rise to greater convergence in the compounds produced and increased confidence for compound producers. Aside from results relating directly to the convergence and confidence predictions, Experiments 6 and 7 proved to be a rich source of data. Below, we detail additional findings, which provided a broader picture of people’s responses in this compound production task.

4.7.1. Compound Convergence and Confidence Ratings in Filler Items

For descriptions of high plausibility, people exhibited high confidence, and the overall convergence in compounds produced was also high. Following this line of reasoning, we might also expect a difference between people’s responses for the test items and the filler items. As the fillers were tautological, they could be viewed as being the most plausible descriptions of all. Thus we would expect people to produce the lowest number of unique labels for these descriptions (i.e., greatest convergence),
and also exhibit the highest levels of confidence from coiners. We found that convergence was reliably higher for filler items compared to the test items - $F(2, 100) = 17.863, MSe = 2.18, p < 0.001$. Using Planned Bonferroni comparisons, Fillers were differentiated from High- and Low-Plausibility items (both $p$s < 0.001). Furthermore, responses to filler items had reliably higher confidence ratings than the High- and Low-Plausibility conditions, $p < 0.001$ for both using Planned Bonferroni comparisons - $F1(2, 21) = 47.772, MSe = 5.042, p < 0.001$; $F2(2, 57) = 41.993, MSe = 2.593, p < 0.001$. This pattern provides further support for the impact of Plausibility in people’s creation of novel compounds.

Additionally, we found that people displayed a greater tendency to use noun-noun compounds in response to Low-Plausibility items compared to High-Plausibility ones – approximately 74.1% compared to 84.4%. This finding was also reliable both by participants and by items - $F1(1, 21) = 4.71, MSe = 0.236, p = 0.042$; $F2(1, 43) = 4.352, MSe = 0.207, p = 0.043$. This suggests that when Plausibility is decreased, people are more likely to use noun-noun compounds. Logically, if this were the case then we should see people creating even fewer noun-noun labels for the filler items in these experiments. In fact, people only produced noun-noun labels in 30.45% of cases in response to filler items - $F1(2, 21) = 97.265, MSe = 0.337, p < 0.001$; $F2(2, 43) = 4.352, MSe = 0.207, p = 0.043$. Using Bonferroni comparisons, the use of noun-noun compounds for Filler items was reliably different from High- and Low-Plausibility items (both $p$'s < 0.001). It is worth noting that the preference for novel compound usage for Low-Plausibility descriptions is possibly greater in reality than has been reported here. It has been shown that people's tendency to produce novel compounds increases with their exposure to compounds (Berman & Clark, 1989; Clark et al, 1985; Windsor, 1993). Thus, with people responding to many
descriptions that lend themselves to noun-noun labels, the increase in overall compound usage may dampen the size of the actual effect. However, even with this possibility the difference in noun-noun compound usage for High- and Low-Plausibility items was reliable.

It is possible that the tendency to produce noun-noun labels for less plausible descriptions arises due to the polysemous nature of noun-noun compounds. In fact, for compound producers, there could be a trade off between the capacity of noun-noun compounds to convey a variety of meanings and the potential of ambiguity for prospective comprehenders. So, a person may select a noun-noun compound in response to a Low-Plausibility description because they feel it has a better chance of conveying the intended meaning, but this is offset by the potential for ambiguity, which ultimately reduces their confidence in the newly produced compound.

### 4.7.2. Incorporation of Redundant Knowledge

Compound coining is primarily used as a communicative tool; to express some concept that is new or different from existing concepts. As such, we would expect informativeness to play a key role in people’s word selection. If participants were following the pragmatic constraint of informativeness then the coined compounds should not be tautological, with new information being provided relative to the head concept. In almost all instances this was found to be the case. Where people had an option of encoding “superfluous” information in their coined compound, that route was seldom taken. For example, in response to the description “A pencil carved from wood and used for art”, no participant produced the compound *wood pencil*. The knowledge we have about pencils tells us that pencils are generally made from wood, so there is nothing to be gained by using *wood pencil* over “pencil” on its own. On
the other hand, the compound *stone pencil* was produced several times in response to “a pencil carved from stone and used for art”. This is because “made of stone” is not usually a feature of pencils, and so is informative relative to the knowledge we have about pencils.

Overall, people exhibited a clear predisposition towards creating informative novel compounds, following a core assumption of many views of language production (Grice, 1975; Levelt, 1989; Sperber & Wilson, 1986). Where, Costello and Keane (2000) have demonstrated the importance of informativeness as a constraint in comprehending novel compounds, it would also seem that it is an essential factor in producing novel compounds.

### 4.8. Experiment 8 – Compound Production from Interpretations

Experiments 6 and 7 have demonstrated that the plausibility of descriptions affects the production of novel compounds. Additionally, we found that it is the knowledge of the concepts themselves, and not necessarily the word order of the descriptions that was the primary force behind people’s word selection (see Experiment 7). In other words, despite the fact that the word order was different in the two sets of descriptions (while the descriptions’ meanings were retained), people produced the same compounds in both experiments.

In this experiment, we focus on the link between the compounds people produce and the interpretations that people arrive at for novel compounds. There are two principal motivations for this. First, in the two previous experiments we used experimenter-derived descriptions to elicit compounds from people. This may have lead to more artificial examination of compound production. In this experiment the descriptions are not experimenter-derived, but are taken directly from interpretations
that people have previously given for novel compounds. Specifically, we use pairs of descriptions (i.e., interpretations) that differ in their levels of Availability, and examine how this might affect the compounds people produce. Second, this approach also allows us to consider compound production as the inverse of compound comprehension. If people produce target compounds (i.e. the compounds that people provided the interpretations for) more often for High-Availability descriptions, it will suggest as strong symmetry between the production and comprehension of novel compounds. In other words, there is a continuity between the interpretation that is most available for a given compound and the compound that is most available for that description. On the other hand, if people do not produce target compounds more often for High-Availability descriptions, it points to a greater asymmetry between production and comprehension.

In this experiment, we use interpretations that were either of High- or Low-Availability as descriptions. These interpretations were taken from Experiments 1 and 2. Interpretations that were of High-Availability were produced more often in an interpretation task, and also they were responded to more quickly in a response time task. Thus, for each description (i.e., the original interpretation) there is a target compound. If people produce target compounds more often for the High-Availability descriptions, then this provides support for the link between production and comprehension (cf. Costello, 2002). For the target compound skunk squirrel, the two descriptions used were:

a) A squirrel that smells bad (High-Availability)

b) A squirrel with a black and white stripe on its back (Low-Availability)
As well as looking at the question of whether compound production can in some way be construed as the inverse of compound comprehension, we look again at the predictions of convergence and confidence. Unlike Experiments 6 and 7, however, the descriptions are differentiated by their *a priori* availability for particular compounds, as opposed to the actual plausibility of the descriptions. Thus, it is unclear whether interpretation availability will lead to differences in convergence and confidence for the two classes of description.

**Method**

*Participants*

Twenty four native English-speaking undergraduates at University College Dublin participated in the experiment.

*Materials*

Materials consisted of 24 descriptions that were derived from interpretations of novel compounds in an earlier comprehension experiment. The interpretations were originally classified as either High- or Low-Availability, based on their frequency of production, and on differences in an online paraphrase judgement task. Therefore, there were 12 descriptions of High-Availability and 12 of Low-Availability.

*Design*

The design was single factorial with, Interpretation Availability as a within-subjects factor. Dependent variables were participants’ confidence ratings and the actual compounds people created in response to the descriptions.
**Procedure**

The procedure was the same as that of Experiments 6 and 7. Participants were asked to create a shortened phrase of no more than two words to capture the description, and to rate their confidence in their produced label.

**Scoring**

As in Experiment 6 and 7, the phrases coined were independently categorised by the both the author and supervisor as either "Noun-Noun Compounds" or as "Other". There was 100% agreement in this classification. Of 51 unique noun-noun compounds produced, 3.9% were lexicalised.

**Results**

There was a main affect of Interpretation Availability on the production of target compounds. Overall, 50% of responses for High-Availability descriptions resulted in the production of the target compound, compared to only 28% of responses for Low-Availability descriptions. For example, in response to the High-Availability description “a film that is about lakes”, 50% of participants produced the target lake *film*. This compares to 33% of participants who produced the target for “a film that is made beside a lake”. The effect of Interpretation Availability was reliable, treating both subjects and items as random factors - F1 (1, 21) = 8.471, p < 0.008, MSe = 0.205; F2 (1, 11) = 6.39, p < 0.028, MSe = .128. There was also a significant correlation between the original frequency of production of interpretations and the ranking of the produced compounds - r = 0.678, p < 0.001.

There was a main effect of Availability on confidence ratings, but compounds produced for Low-Availability descriptions had better confidence ratings than those produced for High-Availability items. This was reliable treating both participants and items as random factors in a one-way ANOVA - F1 (1, 21) = 8.644, p < 0.001, MSe =
3.464, F2 (1, 22) = 4.665, p < 0.054, MSe = 2.12. The mean confidence ratings were 2.92 and 2.46, for High- and Low-Availability items respectively, remembering that lower ratings indicate greater confidence.

There was no difference in convergence in the compounds produced for High- and Low-Availability items (t < 1), with 2.08 and 2.36 unique compounds produced for High-Availability and Low-Availability descriptions respectively.

**Discussion**

The finding that more people produced target compounds for High-Availability descriptions supports the notion of a symmetry between the comprehension and production of novel compounds; that there is a link between the compounds that are most available for a particular description, and the interpretations that are most available for a particular compound. While this result is in line with some views in the literature (Clark & Berman, 1984), it conflicts with findings of other researchers (Costello, 2002). Costello found that interpretations that were judged as “good” by one set of participants did not reflect the likelihood of a target compound being produced. However, our results contradict this finding, suggesting that there is a link between the availability of an interpretation and the likelihood of a target compound being produced.

The finding that people had greater confidence for compounds produced for Low-Availability descriptions implies that the availability of interpretations can not be characterised by the plausibility of the interpretation alone. If availability were essentially the same as plausibility, the pattern of results for confidence ratings from Experiments 6 and 7 should have been replicated. In a similar vein, the convergence prediction was not borne out for this experiment. People did not exhibit greater convergence for the High-Availability descriptions, despite the fact that more people
produce the target compound. Again, this goes against the alignment of interpretation availability and plausibility. Additionally, there was no difference in the number of noun-noun compounds produced in the High- and Low-Availability conditions, which was also evident in the previous experiments.

To summarise, the findings of this experiment first suggest that there is a link between the interpretation availability and the likelihood of particular compounds being produced. In other words, there is a symmetry between the availability of an interpretations I for a compound C, and the availability of the compound C for the Interpretation I. Second, the availability of interpretations cannot simply be characterised by their plausibility. There are many factors that contribute to the availability of an interpretation (e.g., diagnosticity, informativeness), as well as plausibility. In a post-test, we asked a separate set of 20 participants to rate the plausibility of the descriptions used in this experiment (using the same procedure as Experiment 5). We found that there was no difference in the levels of plausibility between the two sets of interpretation, with means of 4.44 and 4.77 for High- and Low-Availability items, respectively. Thus, the fact that the convergence and confidence predictions were not borne out in this experiment remains consistent with our previous result, since interpretation availability does not carve up the space of possible interpretations in the same way as plausibility by itself (i.e., an interpretation may be considered plausible, but not be highly available).

4.9. Conclusion

In this chapter we have shown that differing levels of plausibility affect people’s behaviour in creating compounds. In producing compounds for descriptions that are considered plausible, people converge on a smaller number of labels, while
also exhibiting greater confidence in their compounds. In contrast, in response to
descriptions that are considered less plausible, people produce a greater variety of
possible compounds while also exhibiting reduced confidence in their compounds.
We also observed that, at least in some instances, there is a relationship between the
availability of an interpretation and the availability of the compounds used to label
that interpretation. This was evidence by the fact that more people produced target
compounds for interpretations that were of higher availability for people. Finally, we
saw a distinction between the availability of interpretations (as descriptions) and the
plausibility of descriptions.

With this empirical work, we have not only demonstrated factors that impact
on people’s creation of novel compounds, but we have also outlined possible
symmetries that exist between the production and comprehension processes. Observing the link between the availability of an interpretation and the compounds
used to label it points to the possibility that people recruit aspects of the
comprehension system in order to achieve their communicative goals (i.e., when
producing compounds). Although most researchers treat comprehension and
production as distinct areas, recent research has suggested that, in dialogue, people
align their comprehension and production representations, thus smoothing the
communication process (Pickering & Garrod, in press). In Chapter 5, we outline a
theory of conceptual combination that takes into account factors that impact both
comprehension and production of novel compounds.
5. THEORY AND MODEL

5.1. Outline

The empirical work carried out in Chapters 3 and 4 has focused on the three pillars of conceptual combination; comprehension in isolation, comprehension in context and compound production. This work has given us a better understanding of how people comprehend novel compounds (in and out of context), and how they go about producing such compounds in the first place. Comprehension was tested in Chapter 3 by having people interpret novel compounds presented in isolation, presented with specific interpretations, and also presented embedded in different contexts. These experiments produced three principal findings. First, the frequency of production of an interpretation affects how quickly people can process that interpretation (Experiments 1 and 2). Second, the out-of-context availability of an interpretation affects how quickly people can process that interpretation in context (Experiment 3). Third, the degree to which the context is biased towards a particular meaning for a compound affects how quickly people can process a particular interpretation (Experiment 3). Thus, both the availability of interpretations and the surrounding context were identified as important factors in understanding novel compounds.

Production was tested in Chapter 4, where we looked at people’s creation of novel compounds given short descriptions. From these experiments, we have three
main findings. First, we found that the plausibility of a description affects the extent to which people converge on particular compounds. Second, plausibility affects people’s confidence in these compounds. Third, we found that when people are presented with an interpretation that is commonly given to a particular compound, they are more likely to reproduce that compound than if they were presented with a less common interpretation.

In this chapter, we outline our Integrated Production and Comprehension Theory (IPAC theory) of conceptual combination. This theory has modified elements of the Constraint Theory, developed by Costello and Keane (2000), but also has several novel extensions. The most significant aspect of this theory is that we characterise the comprehension and production of novel compounds within a single theoretical framework. The theory also describes how various factors interact to dictate the acceptability of both interpretations and newly created compounds. We also discuss how the theory addresses the phenomena and findings from the wider literature as well from our own experiments. The IPAC theory is realised in a computational model called PUNC (Producing and Understanding Novel Combinations). We test a number of facets of PUNC’s comprehension and production systems, and show how the model mirrors human performance in conceptual combination tasks. Finally, we compare PUNC to existing models of conceptual combination, and discuss how our model differs.

5.2. IPAC Theory of Conceptual Combination

Traditionally, research of language comprehension and language production has tended to be kept quite separate. Researchers perceive the issues of comprehension and production as connecting only in very general of ways. It has
been suggested, however, that constraints from the comprehension system may impact on the production system and vice versa (MacDonald, 1999). In this section, we describe the Integrated Production and Comprehension Theory that takes a holistic view of conceptual combination, using a set of core constraints to describe both the comprehension and production of novel compounds. We describe how the processes involved in interpreting and creating novel compounds are affected by many of the same factors (see Figure 5.1). In the subsequent sections, we look at the commonalities shared by comprehension and production and consider the various factors that have been shown to influence them. We also outline how these factors interact to impact on the perceived acceptability of interpretations and compounds.

5.2.1. Commonalities in Comprehension and Production

In creating a novel noun-noun compound, the goal of the coiner is to produce an acceptable label that conveys an intended meaning. Specifically, coiners are using the minimum set of words (and no fewer) needed to convey this meaning. The acceptability of a compound is shaped by a combination of factors, including the prior knowledge of the constituent concepts and the intended meaning of the compound. In understanding a novel compound, the goal of the comprehender is to arrive at an acceptable interpretation that represents, or literally, re-presents the meaning intended by the coiner. The comprehender must assume that the coiner of a compound is following cooperative principles (Grice, 1975), and has chosen the best words to convey the intended meaning. The acceptability of an interpretation is also affected by a combination of factors, including the context in which the compound has occurred, and our prior knowledge of the constituent concepts themselves.
At this general level, it appears that the production and comprehension systems can and do feed into each other, not least in that they must both respect basic Gricean principles of cooperation in communication, and are both affected by our prior knowledge of the concepts involved. However, to appreciate the depth of commonality between production and comprehension, we must go well beyond this level of abstraction. Several empirical factors have been shown to influence conceptual combination, and we now turn to their examination.

### 5.2.2. Factors Affecting Conceptual Combination

From the empirical work presented in Chapters 3 and 4 (and work by other researchers), we know that several factors influence the processes of comprehension and production. Figure 5.1 illustrates the factors that have been shown empirically to impact on people’s comprehension, production or both, including diagnosticity, informativeness, plausibility, familiarity, availability, and context. Of these factors, some have been shown to arise from individual concepts (e.g., diagnosticity), and have direct effects on people’s understanding and creation of novel compounds. Others appear to emerge less directly, from the compounds themselves or through interactions between other factors (e.g., familiarity, availability), to affect people’s comprehension and production. In the remainder of this section, we outline the evidence for these factors in conceptual combination. The subsequent sections then separately examine compound comprehension and production, and discuss in more detail how these factors influence processing.
Figure 5.1 The many factors that have been shown to influence comprehension and/or production of novel compounds

**Diagnosticity**

Diagnosticity has been implicated in Comprehension and Production. People are more likely to construct interpretations using diagnostic features than non-diagnostic features. In comprehension, Costello and Keane (2000, 2001) have shown that, with competing interpretations, those with diagnostic features are rated as better interpretations (e.g., “prickly beetle” is better as an interpretation for *cactus beetle* than a “green beetle”). In production, descriptions that use diagnostic features of concepts are more likely to elicit target compounds than descriptions that use non-diagnostic features. For example, “a pig that has a trunk” is more likely to be an *elephant pig* than “a pig that is grey”, even though both properties are equally valid (Costello, 2002).

**Informativeness**

In Comprehension and Production the evidence for informativeness is thin, though it appears to play a necessary role. In comprehending novel compounds,
people do not generate uninformative interpretations (Costello & Keane, 1997; Downing, 1977). In producing compounds, people always aim to be informative (Downing, 1977; Chapter 4, Experiments 6 and 7). For example, in producing compounds for the description “a pencil that is made of wood and used for art”, no participant produced the compound *wood pencil*, as using the word *wood* is no more informative than using *pencil* on its own (see Experiment 6 and 7, Chapter 4).

**Plausibility**

In production, plausibility has been shown to impact on the variability of compounds people produce (i.e., the convergence) and also on people’s confidence in their newly created labels (see Experiments 6 and 7, Chapter 4; see also Lynott & Keane, 2003). For example, if the description of an object is highly plausible, people tend to converge on a smaller set of candidate labels than if the description is less plausible. Similarly, people are more confident about the compound they have produced if it labels a highly plausible description. In comprehension, Constraint Theory holds that increased plausibility boosts the overall acceptability of an interpretation (Costello & Keane, 2000).

**Familiarity**

Compounds that are familiar to people are understood more quickly than compounds that are not familiar. For example, people would be quicker to understand *night club*, a familiar compound, than *night court*, a less familiar compound. As such, familiarity of compounds has been shown to be a confounding factor in response time experiments (Tagalakis & Keane, 2003), where experimenters have not considered it a factor. In addition, familiarity influences the interpretations that people generate for a particular compound, even if people are only familiar with a related compound. For
example, bullet car could be interpreted as “a high-speed car” with direct reference to the familiar compound bullet train (Tagalakis, Keane & Lynott, 2003).

Availability

People respond more quickly to highly available interpretations compared to interpretations that are less available (see Experiments 1 and 2, Chapter 3; see also Lynott & Keane, 2002). Furthermore, interpretations that are more available out-of-context are also responded to more quickly when placed in context (see Experiment 3, Chapter 3). In addition, when people are presented with an interpretation that is commonly given to a particular compound, they are more likely to reproduce that compound than if they were presented with a less common interpretation.

Context

Different types of preceding context have been found to affect people’s judgements of novel compounds and their interpretations (see Experiment 3, Chapter 3; Gerrig & Murphy, 1992). For example, a context that is biased towards a particular interpretation of an ambiguous compound will facilitate people’s comprehension. On the other hand, in an unbiased context, people’s responses reflect the out-of-context availability of interpretations (see Experiment 3, Chapter 3). Additionally, the history of how concepts have been combined in previous contexts also impacts on how easily people arrive at a target interpretation (Gagné & Shoben, 2002). If a concept is used in a compound with a particular relation in a preceding context, then that relation will be primed in any subsequent compound that contains the same concept. For example, having previously read the compound adolescent magazine as “a magazine for adolescents”, it is easier to interpret the compounds adolescent doctor or animal doctor using the same relation. Thus, specifying a relation in advance in a context facilitates comprehension.
5.2.3. Compound Comprehension

The Integrated Production And Comprehension (IPAC) Theory of conceptual combination views the goal of compound comprehension as one of constructing an interpretation that best captures the intended meaning of the compound. In essence, compound comprehension acts as the recipient in the communication of conceptual combination.

There are a number of different stages to the comprehension process. Firstly, comprehenders hear or read the compound (e.g., *cactus beetle*) and identify the head noun (e.g., “beetle”). Secondly, they must retrieve the constituent concepts of the compound (e.g., *cactus* and *beetle*) from their long-term memory. Thirdly, they must have the most diagnostic features of the concepts made available (e.g., “has spikes”, “can conserve water”, “is found in the desert” for *cactus*). Lastly, comprehenders must generate an interpretation by using the diagnostic features of both concepts to modify the head concept. This is a summary view of the comprehension process that glosses each step, because any and all of these stages are subject to more specific influences from a number of different sources.

Compound comprehension is guided by pragmatic constraints, such as those described by the Constraint Theory (Costello & Keane, 2000); diagnosticity, informativeness and plausibility. However, there are many other factors that have been shown to influence conceptual combination. Some of these factors arise from individual concepts (e.g., diagnosticity), while others emerge from the combination itself or through interactions between other factors (e.g., familiarity, availability). In the following sections, we consider each factor in turn, and discuss how they
influence, either directly or indirectly, the comprehension and acceptability of novel compounds.

**Diagnosticity**

Diagnosticity is a central factor in the comprehension of novel compounds. Following from Costello and Keane (2000; see also Chapter 2, Section 2.2.3.), an interpretation for a novel compound should contain diagnostic features of both head and modifier concepts, with diagnostic features being those features that best distinguish a concept from other concepts. For example, “has a trunk” is a diagnostic feature of elephants, so interpreting *elephant pig* as “a pig that has a trunk” incorporates a diagnostic feature of the modifying concept.

As in Constraint Theory, the overall acceptability of an interpretation is tied to the diagnosticity of the features in the interpretation; the more diagnostic the features used, the greater the acceptability of the interpretation. Thus, our previous interpretation of “a pig with a trunk” for *elephant pig* is considered a better interpretation than “a pig that is grey”, since “has a trunk” is more diagnostic than being grey. In this way, making use of more diagnostic features of the constituent concepts increases the overall acceptability of an interpretation.

However, diagnosticity is not a static phenomenon. Regardless of a feature’s original diagnosticity in a constituent concept, once it is used in an interpretation that feature is elevated to being the most diagnostic feature of the combined concept. The reason for this is that the new feature distinguishes the combined concept from other instances of the head concept. Let us return to our earlier example of *elephant pig* as “a pig that is grey”. In this case, the colour feature of both constituents (i.e., “is grey” for *elephant*, “is pink” for *pig*) is not particularly diagnostic. However, in combining these concepts, the modifier feature “is grey” replaces the head feature “is pink”, and
is thus elevated to being the most diagnostic feature of the combined concept. In this way, *elephant pig* is interpreted as “a pig that is grey” as the colour feature “is grey” is more diagnostic than any other feature of *pig*.

**Informativeness**

We follow Costello and Keane’s view of informativeness in comprehension. Creating a novel compound is a means of conveying some new information such that both words in the compound are necessary and sufficient. If this were not the case, the speaker would have used fewer or more words. Therefore, an acceptable interpretation will provide some new information relative to the original constituent concepts of the compound. As we have mentioned (see Chapter 2), *cactus hedgehog* as “a prickly hedgehog” would not be an informative interpretation as our prior knowledge of hedgehogs tells us that they are already prickly. Alternatively, “a hedgehog that eats cacti” would be an informative interpretation as there is new information being provided relative to both concepts – hedgehogs do not usually eat cacti, and cacti are not usually eaten by hedgehogs. Accordingly, informativeness is implicitly linked to acceptability, as uninformative interpretations are not acceptable.

**Plausibility**

When a person encounters a novel compound, they can assume that there must be some plausible link between the concepts; otherwise the compound producer would have used a more verbose description, and not a two-word compound. Therefore, an acceptable interpretation for a novel compound must specify some link between the constituent concepts that is consistent with our prior knowledge of those concepts. Connell and Keane (2003a, 2003b, in press), describe such links as contributing to the *concept-coherence* between concepts. The concept-coherence of an interpretation is established with reference to the prior knowledge of the
compound’s constituents. In other words, if an interpretation fits well with the knowledge we have about the individual concepts, it is said to have high concept-coherence and therefore high plausibility.

The concept-coherence of an interpretation (and therefore its plausibility), can be deduced by examining the exact manner in which the constituent concepts are linked together. For example, interpreting *cactus beetle* as “a beetle than eats cacti” appears to be more plausible than interpreting *brick beetle* as “a beetle that eats bricks”. Why is this the case? Our prior knowledge tells us that *cactus* is vegetative matter and so it can be eaten, and *beetle* is a living creature and so it has the ability to eat things. Because of this, the two concepts in *cactus beetle* have a reciprocal relationship centred on eating, which gives the “beetle that eats cacti” interpretation a high level of concept-coherence, and therefore plausibility. By contrast, bricks are not usually edible, so this reduces the coherence and plausibility of the “beetle that eats bricks” interpretation. Thus, while the interpretations are similar, they are distinguished by their concept-coherence, with the former appearing as the more acceptable interpretation. Extending this further, interpreting a *brick apple* as an “apple that eats bricks” would be even less plausible, and therefore less acceptable, because although an apple can be eaten, it cannot eat things, and even if it could bricks are still not edible.

In this way, plausibility, as is the case in Constraint Theory, both constrains the interpretations that can be constructed, and contributes to the calculation of the overall acceptability of an interpretation.

**Familiarity**

Intuitively, familiar compounds can be understood more quickly and easily than unfamiliar compounds (Tagalakis & Keane, 2003). Familiarity refers to having
knowledge of an interpretation for a compound. This includes lexicalised compounds such as *finger food* (i.e., where the compound itself has a standard meaning) and previously encountered compounds that have already been interpreted (or where an interpretation has been specified). For example, to someone who has just read the above discussion of plausibility, the compound *cactus beetle* would be familiar. Familiarity also affects the interpretation of compounds that are related to lexicalised or previously encountered compounds, so interpreting *bullet car* as “a high-speed car” is easier for someone if they are already familiar with *bullet train*. Interpreting *bullet car* in this way gives a highly acceptable interpretation for two reasons. First, the interpretation incorporates the diagnostic feature of a bullet (i.e., its speed), and second, the interpretation specifies a plausible link between the constituent concepts. Because we have previously encountered *bullet train*, we know that the high-speed of a bullet can be plausibly linked to other concepts. Tagalakis and Keane (2003) have argued that the semantic similarity between the head concepts will also facilitate such an interpretation (i.e., *car* and *train* are both modes of transport). Similarly, Gagné and Shoben (2002) have shown that if people are made familiar with the compound *adolescent magazine* as meaning “a magazine for adolescents”, they are faster to interpret *adolescent doctor* as “a doctor for adolescents”

Accordingly, familiarity with related compounds can be seen to have two primary effects. The first is to facilitate particular interpretations; the second is to boost the acceptability of these interpretations, by virtue of their prior use.

**Availability**

Availability refers to how easily/quickly a particular interpretation for a compound can be generated/understood. When given a compound to interpret, many people will arrive at the same interpretation. These frequently generated
interpretations are also understood more quickly, and are therefore said to be highly available (see Experiment 2, Chapter 3). Conversely, interpretations that are generated less often will be considered less available.

The availability of an interpretation arises through a combination of other factors. Interpretations using highly diagnostic features of the constituent concepts are more available than those that use non-diagnostic features, with the former being produced much more often by people (Costello & Keane, 2001). Interpretations of familiar compounds are more available than those of unfamiliar compounds, with people consistently responding more quickly to them (Tagalakis & Keane, 2003). Finally, interpretations that have been fully specified in a preceding context tend to be more available than interpretations that have not (see Experiment 3, Chapter 3). In essence, some interpretations are more available than others because of how they utilise the features of the constituent concepts (i.e., diagnostic versus non-diagnostic), how the compound relates to existing combinations (i.e., familiar versus unfamiliar) how they pertain to the current contextual environment (i.e., fully specified versus not specified). Thus, the availability of an interpretation clearly reflects many complex influences that interact to affect how readily people can comprehend/generate that interpretation.

Context

Context is a factor external to a compound, yet it acts on the factors we have described above, to influence the interpretation of that compound. The context surrounding a particular compound can affect the diagnosticity of particular features of the constituent concepts, and the availability, acceptability, and informativeness of particular interpretations.
Context affects diagnosticity by prioritising features of concepts that are particularly relevant, and by demoting features that are less relevant. When people interpret compounds in isolation, the features of the constituent concepts are quite stable and are ordered according to default levels of diagnosticity (e.g., a skunk produces a bad smell, is black with a white stripe, etc.). However, context can act to promote or demote the relevance of particular features, depending on what information is present. For example, when the compound skunk squirrel is presented in isolation, the interpretation “a squirrel with a bad smell” is produced more often than the interpretation “a squirrel with a white stripe” (see Experiment 3, Chapter 3). However, if the compound is embedded within a context that makes reference to the colour of a skunk, then the stripe feature of skunk is promoted above the odour feature because it is more relevant to the present context. In this context, people understand the interpretation “a squirrel with a white stripe” more readily than the interpretation “a squirrel with a bad smell” (i.e. context affects the availability of particular interpretations over others). Also, in this context, “a squirrel with a white stripe” will be considered the more acceptable interpretation. Thus, context acts to reorder the diagnostic features of the constituent concepts according to what is currently relevant, which in turn influences the availability and acceptability of interpretations.

Context can also affect the informativeness of particular interpretations. Normally, the interpretation “bacon made from pig meat” for the compound pig bacon would be uninformative and unacceptable. However, there exists a product known as turkey bacon, which is salted and cured, but is made from turkey. So, in the context of this product, the above interpretation for pig bacon becomes informative and acceptable. Therefore, context may enhance the acceptability of particular interpretations by making them appear more informative than they would in isolation.
5.2.4. Compound Production

The IPAC Theory of conceptual combination views the goal of compound production as constructing a compound that best encompasses the intended information of the referent. In essence, compound production acts as the provider in the communication of conceptual combination.

As with comprehension, there are a number of different stages to the production process. Firstly, producers must identify the entity they wish to label (e.g., “a beetle that has spikes”). Secondly, they recognise the entity as an instantiation of a particular concept (e.g., beetle), and retrieve this concept from their long-term memory. Thirdly, producers must then identify any differences that exist between the entity’s features and the features of the head concept (e.g., “has spikes”). Lastly, producers can use the diagnosticity of these differences to retrieve other concepts, and create a label by allowing another concept to serve as a modifier to the head concept. Again, this is a summary view of the production process that glosses each step, because, as with comprehension, any and all of these stages are subject to influence from a number of different sources.

We have already discussed how compound comprehension is guided by many pragmatic and experiential factors, and it is a central tenet of the IPAC theory that comprehension and production are both influenced by the same core set of factors. However, the exact nature and influence of some of these factors is less well understood in terms of production than in terms of comprehension – for example, the roles of availability and context in the production process have not been specifically addressed in the literature. In the following sections, we consider each factor in turn,
and discuss how they influence, either directly or indirectly, the production and acceptability of novel compounds.

**Diagnosticsity**

In comprehension, we have described how diagnosticity, as outlined by the Constraint Theory, is a central factor. Diagnosticity is also important in the production of compounds (Costello, 2002; see also Section 2.4.1, Chapter 2). Produced compounds frequently incorporate diagnostic features of other concepts. For example, an *elephant seal* suggests the large size of an elephant, a *seahorse* refers to the shape of a horse, and a *crab spider* exhibits the sideways walking manner of a crab; all taken from the Encyclopaedia Britannica (2001). Incorporating diagnostic features in this way exemplifies the kind of creativity often seen in compound production, without the need to sacrifice clarity or accuracy. The diagnosticity of such features incorporated into these compounds impacts on the overall acceptability of the compounds, with highly diagnostic features leading to more acceptable compounds. In contrast, features that are not diagnostic of a concept will not be used to produce a compound using that concept. If we take two descriptions, (a) and (b)

(a) a fish that has spikes

(b) a fish that is green

The compound *cactus fish* can only be considered an acceptable label for (a). Although the colour green is a feature of the concept cactus, it is not highly diagnostic, and so diminishes the acceptability of *cactus beetle* as a label for (b).

**Informativeness**

People create compounds in order to convey some new information that could not be expressed using a single word. As such, an informative compound must be
informative with respect to its constituent concepts. For example, describing “a wine that is made from grapes and contains alcohol” as a grape wine would not be informative, since no new information is being added to the concept of wine (i.e., wine is normally made from grapes). In this case, the word wine would suffice as a label (see also Experiments 6 and 7, Chapter 4).

As with comprehension, informativeness is fundamental to the pragmatics of communication. It allows people to be efficient by preventing superfluous information from being imparted where less will meet communicative requirements. Similarly, informativeness is bound to the acceptability of a produced compound. Compounds that provide no new information relative to the constituent concepts will be considered unacceptable.

**Plausibility**

Plausibility is a central consideration in compound production, as it is in compound comprehension. In production, plausibility refers to the plausibility of the compound with respect to the entity being described. Crucially, this means that plausibility affects not only the overall acceptability of a compound, but also the likelihood that people will use a noun-noun compound as a label in the first place.

When creating labels for novel objects, people are less likely to choose noun-noun labels (than adjective-noun or single-word labels) if they regard the object’s description as quite plausible. The reason for this lies in the greater polysemy of noun-noun compounds compared to other options. If the object being described is completely plausible (i.e., if it completely fits with our prior knowledge), then it is likely that a single word will suffice as a label (e.g., “something that has a handle and can be opened and closed” can be adequately described as a door). However, because noun-noun compounds can give rise to a wide variety of meanings, they are more
suitable when describing less plausible entities. An adjective-noun compound may be adequate if the object being labelled fits reasonably well with our prior knowledge of the concepts involved and has only minor violations (e.g., *purple bird*). However, noun-noun labels become necessary as objects stray further from our perception of what is plausible. For example, “a lobster that kills shrimp with its web” is not considered plausible because it does not fit well with our prior knowledge (i.e., we have not encountered lobsters spinning webs). This entity would be difficult to describe adequately using a single word, whereas a noun-noun compound works better (e.g., *web lobster*, *spider lobster*). Thus, the plausibility of the entity being described allows the producer to gauge the type of label needed to describe it adequately.

Plausibility also affects the confidence with which people use compounds to label objects. If the entity being labelled seems plausible to the producer, then they assume it will seem plausible to a potential comprehender, and that the comprehender will find it relatively easy to understand their compound correctly. Therefore, the producer is likely to be more confident that whatever compound they use to label the object will be easily interpreted. On the other hand, dealing with less plausible objects makes compound production more difficult. In this case, the producer will be less confident, as they know that regardless of the compound they produce, the intended meaning of the compound will seem less plausible to the comprehender and may be in danger of being rejected for a more plausible interpretation. This leads to a second, knock-on effect. If people are less confident about the labels they are producing, by virtue of the plausibility of the original description, then they are more likely to try alternative labelling options in order to increase acceptability of their compound. This leads to greater variability in the compounds that people produce for
the same description (see Experiments 6 and 7, Chapter 4; see also Lynott & Keane, 2003). Thus, reduced plausibility lowers producers’ confidence and diminishes the extent to which people converge on the same labels.

As with interpretations in comprehension, plausibility affects the acceptability of compounds (as distinct from the plausibility of the entity being described); all other factors being equal, a highly plausible compound will be more acceptable than a less plausible compound. In comprehension the coherence of concepts in an interpretation (with respect to the original compound) affected the overall acceptability of that interpretation. In production, the coherence of concepts in the original description will affect the acceptability of the compound used to capture the description. Let us consider our two sample descriptions – “a beetle that eats cactus” and “a beetle that eats bricks”, and two candidate compound labels – cactus beetle and brick beetle. In the first example, our prior knowledge tells us that beetles are living creatures and as such have the capacity to eat things. Since cacti are vegetative matter, then they can be consumed. Thus, cactus beetle serves as a plausible label for the entity being described. In contrast, using brick beetle as a label is considered less plausible, by virtue of the fact that bricks are not edible, and so it is less likely this compound will convey the correct meaning, thus reducing its plausibility as a label.

Therefore, plausibility plays an essential role in firstly constraining the compounds that may be produced, and also in contributing to the overall acceptability of a compound, with respect to the prior knowledge of the constituent concepts (i.e., their concept coherence).

**Familiarity**

When creating a compound, familiarity with other existing compounds is likely to influence people’s choice of words. If a description appears to be close in
meaning to that of a familiar compound, then people may regard that compound as an acceptable label. For example, familiarity with the compound *holiday home* might lead people to use it as a label for “a farmhouse in the country where a family goes on holiday”, even though this compound excludes other relevant information about the object being described. In other words, people may exhibit a preference for using an existing compound because of its comforting familiarity, even if it does not convey the full and correct meaning. It is a distinct possibility that this familiarity will increase compound acceptability by making inexact labels more acceptable.

Additionally, compounds related to existing combinations may sometimes be favoured in place of completely new, untested labels. For example, familiarity with the compound *holiday home* could lead someone to use *winter home* as a label for “a cottage where one holidays during winter”, or *winter house* if familiar with *summer house*, rather than *winter cottage*, which employs the actual words from the original description (see Chapter 4). Analogically created compounds of this type (e.g., *winter home*) may offer a communicative advantage over more original compounds (e.g., *winter cottage*), because the comprehender is likely to interpret the compound with reference to its familiarity, and in doing so smoothing the comprehension process. Similarly, *winter cottage* may only communicate its constituent concepts, while *winter home* may also communicate overtones of *holiday home*. So, while familiarity may increase the perceived acceptability of a compound, it can also allow the producer some communicative shortcuts by packing more information / meaning into a compound.

**Availability and Context**

In comprehension, the availability of different interpretations and the influence of different contexts have been shown to affect how people understand a compound
(Experiments 2 and 3, Chapter 3). By the same token, it seems likely that availability and context can also influence how people create a compound. Certain compounds will be more available than others, depending on a combination of the factors we have outlined previously (e.g., feature diagnosticity, coherence between constituents of a compound). Similarly, surrounding contexts may constrain word selection for compounds, or indeed make normally unacceptable compounds appear acceptable. As we discussed in the section on comprehension, uninformative compounds may suddenly appear informative and acceptable (e.g., \textit{turkey bacon} versus \textit{pork bacon}) makes the normally uninformative compound \textit{pig bacon} informative), but its influence may be more far-reaching than that.

When a compound producer is trying to label a particular object, the preceding discourse context will influence what concepts are most available. For example, let us assume that the feature “has black and yellow stripes” is equally diagnostic of both \textit{tiger} and \textit{bee}. So, the description “a box painted with black and yellow stripes” can be equally well labelled as a \textit{tiger box} or a \textit{bee box}. However, if the preceding context had mentioned insects and honey, then the \textit{bee} concept would be more available than the \textit{tiger} concept, and the producer may be more likely to produce the compound \textit{bee box} rather than \textit{tiger box}.

While it is clear that context and availability are intertwined, their precise relation and the effects they bring remain opaque without further empirical investigation.
5.2.5. Summary of Comprehension and Production

The previous sections have described a theory of compound production and comprehension in conceptual combination, specifying a number of factors that are central to both systems, and discussing how each of these factors affects the processes and outputs of both systems. Figure 5.1 illustrates a general connection between the comprehension and production, and how producing compounds must necessarily recruit aspects of the comprehension system to achieve communicative goals. Thus, our theory provides a holistic view of conceptual combination, drawing together many aspects that have only been alluded to previously. In the remainder of this chapter, we describe our cognitive model of conceptual combination, PUNC (Producing and Understanding Novel Combinations). This model is the computational implementation of the IPAC theory, and we show how the model’s outputs closely correspond to human performance in a range of tasks concerning conceptual combination.

5.3. The PUNC Model

The IPAC theory describes how a core set of factors acts to influence the processes of conceptual combination. In this section, we introduce PUNC (Producing and Understanding Novel Combinations), our computational instantiation of the IPAC theory. We describe how key factors are implemented, and detail how PUNC models their combination and interaction with regard to the comprehension and production of novel noun-noun combinations (see Figure 5.2). We present several simulations of empirical data, testing PUNC in both comprehension and production sides of conceptual combination. In addition, we discuss how PUNC differs from previous models of conceptual combination.
Figure 5.2 The comprehension and production aspects of the PUNC model
5.3.1. Factors Modelled

Our theory of conceptual combination describes how a number of factors contribute to both the comprehension and creation of compounds. In PUNC, our objective is to model how different factors affect both the interpretation and production of novel noun-noun combinations. In brief, the factors of diagnosticity, informativeness, plausibility, availability, and context are implemented in PUNC\(^1\), as described below.

Within the description of Diagnosticity, we also describe the knowledge base over which PUNC operates for both comprehension and production. Finally, the section on context describes how we utilise the Latent Semantic Analysis model (Landauer & Dumais, 1997) to simulate contextual influences in the comprehension side of PUNC.

Modelling Diagnosticity

Diagnosticity values are encoded in the knowledge base for each feature of each concept. Therefore, diagnosticity values are immediately available upon retrieval of a concept. These diagnosticity values are utilised in both the comprehension and production sides of PUNC as the same knowledge base as utilised

\(^1\) As stated, PUNC models the production and comprehension of novel noun-noun compounds. However, the factor of familiarity is more concerned with non-novel (i.e., familiar) compounds, and their analogical use in related compounds. Therefore, PUNC omits this particular factor (but see Lynott, Tagalakis & Keane, in press; Tagalakis, Keane & Lynott, 2003).
by both parts of the model. We describe the structure of the knowledge base and how diagnosticity values were ascribed to individual concept’s features.

PUNC uses a simple ontology to represent knowledge associated with each concept. This knowledge encoded includes diverse information such as a taxonomic concept hierarchy, diagnostic features of concepts, and non-taxonomic relations of concepts to other concepts. For example, the concept *cactus* is represented by diagnostic features such as “has spikes”, “grows in the desert” and “can conserve water”. Within the concept hierarchy *cactus* will also inherit features from *plant*, so features such as “can grow” and “can be eaten” are also part of the concept representation. Similarly, the concept *beetle* will inherit features from *insect*, which in turn inherits from *invertebrate*, *creature* and *entity*. So, a beetle’s features such “has shiny black colour” will be augmented by features from these concepts higher in the taxonomy. Table 5.1 provides some example concepts with their associated diagnosticity values.

The diagnosticity values encoded in the knowledge base were arrived at from the judgements of three raters (the author and two independent raters). Each rater was given a list of concepts and asked to list the features of each one, ranking them in descending order of their diagnosticity for each concept. The raters listed these features independent of any particular compounds or interpretations, and were based purely on their intuitions concerning the individual concepts. When the raters had completed this task all of the features were combined into a single representation for each concept. Where there was disparity over the relative diagnosticity of a feature, a value was assigned based on the consensual view of the raters.
Table 5.1 Example concepts from the PUNC knowledge base. Concepts are represented by their features, in descending order of diagnosticity. Thus, lower diagnosticity scores represent greater diagnosticity.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Features and Diagnosticity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cactus</td>
<td>has spikes = 1, is found in deserts = 2, can conserve water = 3, has green colour = 4, can be eaten = 5, has a cactus shape cactus = 6, can grow = 6, can photosynthesise = 7, can be bought = 8, can be owned = 9, can be delivered = 10, can be talked about = 11</td>
</tr>
<tr>
<td>Beetle</td>
<td>has wings = 2, is shiny = 3, has colour black = 4, can fly = 5, has six legs = 6, can crawl = 7 has antennae = 8, is creepy = 9, has small size = 10, can eat things = 11, can move = 11, can be eaten = 11, can grow = 13, can kill things = 13, can catch things = 13, has a beetle shape = 19, can be taught = 19, can be owned = 19, can be treated = 19, has a mouth = 19, has legs = 19, has a head = 19, has a body = 19</td>
</tr>
<tr>
<td>Magazine</td>
<td>can be bought = 1, has a topic = 1, has a glossy finish = 2, has a free gift = 3, is made from paper = 4, can fold = 5, can be read = 7, is a location = 10, can be used for something = 10, has a magazine shape = 10, can be moved = 12, can be thrown = 12, can be caught = 13, can be owned = 15, can be delivered = 16, can be talked about = 17</td>
</tr>
<tr>
<td>Skunk</td>
<td>has a bad smell = 1, has black colour with white stripe = 4, can eat things = 5, can move = 5, can be eaten = 5, can grow = 7, can kill things = 7, can catch things = 7, has a skunk shape = 13, can be taught = 13, can be owned = 13, can be treated = 13, has a mouth = 13, has legs = 19, has a head = 19, has a body = 19</td>
</tr>
</tbody>
</table>

For the concept \textit{cactus}, the features “is spiky”, “is found in deserts” and “can conserve water” have been assigned diagnosticity values of 1, 2 and 3, respectively, with feature diagnosticity values ranging from 1 to 20 (approx.), depending on the
individual concept. Apart from these features, represented explicitly as being part of cactus, other features are inherited. Any features inherited from a parent concept are automatically assigned lower diagnosticity values than those of the child. While “can photosynthesise” might be diagnostic of plant, its diagnosticity value will be poorer in the representation of cactus. So, if the lowest diagnosticity value in a concept were 3, then any inherited feature would have a score of 4 or more, depending on its own diagnosticity in the parent concept.

Since PUNC merely represents features that are important to each concept, the knowledge represented is quite diverse. For example, the feature “has spikes” (a property) is important for cactus, whereas the concept vet might have the feature “treats animals” (a role) or the concept chess might have the feature “is strategic” or “is tactical” (a more subjective property). Of course the importance of a feature for a particular concept can change from one situation to another or from one person's perspective to another's; this is something we discuss in the section on context (below), where we have incorporated the Latent Semantic Analysis model to affect changes on the diagnosticity values of features of individual concepts. As the same knowledge base is used in both comprehension and production, the same diagnostic features (and values) are utilised in both. They are utilised in constructed interpretations and compounds, and also in calculating the acceptability of both.

Modelling Informativeness

Informativeness is a pivotal factor in the comprehension and production systems. However, informativeness is not implemented as a distinct process, but rather is emergent from the interpretation and production processes – uninformative interpretations and compounds are simply not generated. Since only informative interpretations and compounds are output, informativeness does not contribute
explicitly to the acceptability of a compound or interpretation. Rather, the fact that a particular interpretation or compound is output means that has at least a certain degree of acceptability by virtue of its informativeness.

**Modelling Plausibility**

Plausibility is modelled according to how closely the head and modifier concepts fit together with respect to the original compound or description (i.e., the concept-coherence). If concepts are seen to fit well together in an interpretation or a compound, by virtue of the features in their representations, then their plausibility will be enhanced. Plausibility is encoded in both the comprehension and production systems.

**Modelling Availability**

Availability arises from the interaction of diagnosticity, plausibility and context (i.e. they each contribute to how available an interpretation or compound is). Context could act to increase the diagnosticity of a feature, and therefore increase the resulting availability of an interpretation using that feature. Therefore, the acceptability scores calculated by PUNC are indicative of the availability of a particular interpretation for a given compound, or the availability of a particular compound for a given description.

**Modelling Context**

The contextual component of PUNC is handled using Latent Semantic Analysis (LSA, Landauer & Dumais, 1997). In short, LSA can be used to provide an indication of contextual similarity and so can be used to weight particular features with more or less importance, therefore altering their original diagnosticity values, resulting in changes in the acceptability scores of interpretations using those features.
Currently, Context is used as a factor only in the comprehension system. Below we give a more detailed description of LSA and of how it is used in conjunction with PUNC to simulate the effects of context.

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 LSA, a contextual distribution of a word is formed by moving through the corpus and counting the frequency with which it appears with other words in its surrounding context. In this way, every word may be summarised as a vector – or point in high-dimensional distributional space – showing the frequency with which it is associated with other lexemes in the corpus. In a similar way, a whole sentence, or longer tract of text, may be represented as a single point in distributional space by merging its word points. LSA uses the weighted sum of constituent word vectors to represent tracts of text. This 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. Thus, the similarity between two word (or sentences) is given as a score between -1 and +1. The closer a score is to zero, the further apart the
terms are in the high-dimensional space. For example, comparing “scalpel” to the words “surgeon” and “gardening” gives scores of 0.59 and 0.05 respectively, showing how LSA captures the semantic relatedness of terms.

What evidence is there to suggest that LSA is suitable to act as a model of contextual activation? LSA has been used in isolation to model many different aspects of language comprehension. It has been shown that distances in distributional space can be used to predict semantic priming effects (Landauer & Dumais, 1997; Lund, Burgess & Atchley, 1995). Connell and Ramscar (2001) showed that the typicality of category membership could be predicted by distances in distributional space, and could also predict 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 hunting mice tend to rate owl as most typical. Additionally, LSA has also being used in combination with existing models to contribute to accounts of more complex phenomena. Kintsch (2001) has combined LSA and his Construction-Integration model of text comprehension to develop a model of metaphor interpretation. Ramscar and Yarlett (2003) used LSA to simulate activation of concepts from memory in a model of analogical retrieval. It was found that LSA could simulate the retrieval of structured representations from memory and not just individual terms as had previously been shown (Landauer & Dumais, 1997). Such evidence suggests that LSA may also be suitable for modelling the effects of context in activating specific features of concepts for interpreting novel combinations.

In PUNC, LSA is used to mediate the diagnosticity of a concept’s features depending on the surrounding context. In other words, if a context refers to a feature F1 but not to F2, then we would expect the LSA score between F1 and the preceding context to be greater than the score between F2 and the preceding context, thus giving
a greater weighting to F1’s diagnosticity score. To interpret a compound taking into account LSA’s contextual influences PUNC undergoes 2 stages before generating interpretations:

i. The preceding context is compared to the individual features of the constituent concepts. The LSA score between each feature and the preceding context is used to ascertain the level of activation of every feature within each concept.

ii. Once each feature’s level of activation has been established, PUNC uses the LSA scores to weight the existing diagnosticity scores of the individual features. For example, a feature that normally has a low diagnosticity score but a high LSA score will have its diagnosticity score reduced accordingly.

When these two stages have been complete, PUNC generates a candidate set of interpretations as we describe in the following sections.

We use the example compound *skunk squirrel* and two candidate interpretations (“a squirrel that has a bad smell”, “a squirrel with a white stripe on its back”) to demonstrate how LSA acts to affect the diagnosticity of individual features, and ultimately the availability of the interpretations. In PUNC, the “has a bad smell” (Diagnosticity = 1) feature of *skunk* is represented as being more diagnostic than the “has a white stripe” feature (Diagnosticity = 4). Using LSA, these individual features can be compared to a preceding context in order to compute a contextual weighting on their diagnosticity scores. For example, we first compare these features to an unbiased context (i.e., that does not make reference to any specific interpretation for *skunk squirrel*), which mentions the relevant compound. The “has a bad smell” feature gets a score of 0.59, compared to 0.48 for the “has a white stripe” feature.
features are input directly from PUNC’s knowledge base by simply removing syntactic markers from the representation (e.g., (hasColour_red) would be input as “has colour red”). LSA is not sensitive to word order so “has red colour” would return the same score as “has colour red”. The contextual weighting for each feature is calculated based on the rank of the feature’s LSA score compared to the rest of the concept’s features. Thus, the most highly ranked feature in the LSA comparison will have a contextual weighting of 1. Then, by calculating the mean of the context weighting (from LSA) and the existing diagnosticity of the feature (from the PUNC knowledge base), the in-context diagnosticity of the feature is calculated i.e., (Context Weight + Original Diagnosticity) / 2 = Feature’s New Diagnosticity Score. In this way, LSA affects the relevance of individual features in specific contexts.

5.3.2. Compound Comprehension in PUNC

PUNC’s comprehension system takes a noun-noun compound as input and outputs a set of candidate interpretations. Additionally, PUNC outputs an acceptability score for each interpretation, which it calculates using metrics of diagnosticity and plausibility. Compounds are interpreted by PUNC using the comprehension procedure described in the IPAC theory. When PUNC is presented

2 All comparisons in LSA are performed using the General Reading up to 1st Year College semantic space, with document-to-document comparison at maximum factors. This LSA corpus is used to represent the cumulative lifetime reading of an American first-year university student. LSA is available freely from http://lsa.colorado.edu.
with a compound (e.g., chocolate bee), it first identifies which is the head noun (e.g., “bee”) and which is the modifier noun (e.g., “chocolate”). Second, PUNC retrieves the constituent concepts of the compound (e.g., chocolate and bee) from the knowledge base, which accesses the set of each concept’s diagnostic features (e.g., “has yellow and black stripes”, “produces honey”, “can sting”, etc. for bee) and any features inherited from parent concepts. Third, PUNC assigns the head concept as the base concept for all interpretations (e.g., a chocolate bee is at its base a bee). Lastly, PUNC generates candidate interpretations by using the diagnostic features of both concepts to modify the base interpretation. This interpretation component is the core of the comprehension system. In PUNC, candidate interpretations are generated using four different modification procedures; insertion, transfer, insertion-and-transfer, and interaction.

- **Insertion** involves inserting a modifier noun into a feature of the head; for example, bee has the feature “produces honey”, and so the insertion procedure can interpret chocolate bee as “a bee that produces chocolate”.

- **Transfer** involves transferring a modifier feature directly to the head concept; for example, chocolate has the feature “is brown”, and so the transfer procedure can interpret chocolate bee as “a bee that is brown”.

- **Insertion-and-transfer** involves inserting the head noun into a modifier feature and then transferring it to the head concept; for example, chocolate has the feature “can coat things”, and so the insertion-and-transfer procedure can interpret chocolate bee as “a bee that is coated in chocolate”.

- **Interaction** involves matching a head and modifier feature and allowing them to interact; for example, chocolate has the feature “can be eaten”, and bee has the
Figure 5.3 Pseudocode description of comprehension system in PUNC

```
begin

parse compound
    set head = noun2
    set mod = noun1

retrieve head concept
    retrieve mod concept

    for each feature in head concept
        set diagnosticity = feature diagnosticity

        // INTERACTION
        if feature interactive with mod feature
            set interpretation = feature + head concept
            set plausibility = high
        end if

        // INSERTION
        if feature can take mod as argument
            set interpretation = feature + head concept
            set plausibility = medium
        end if

        acceptability = plausibility*0.1 * (head diagnosticity * 0.55 + modifier diagnosticity * 0.45)
    end for

    for each feature in mod concept
        set diagnosticity = feature diagnosticity

        if feature present in head concept
            go to next feature
        end if

        // INSERTION & TRANSFER
        if feature can take head as argument
            set interpretation = feature + head concept
            set plausibility = medium
        end if

        // TRANSFER
        if feature not present in head concept
            set interpretation = feature + head concept
            set plausibility = low
        end if

        acceptability = plausibility*0.1 * (head diagnosticity * 0.55 + modifier diagnosticity * 0.45)
    end for

output interpretations and acceptability scores
end
```
• (contd.) feature “can eat things”, and so the interaction procedure can interpret chocolate bee as “a bee that eats chocolate”.

This overview of PUNC’s comprehension system is further illustrated by the pseudocode in Figure 5.3. We now describe the system in more detail as we discuss how the influence of each factor is implemented by the model.

**D**iagnosticity in Comprehension

As mentioned previously, diagnostic values are encoded as part of a feature’s representation for each concept. Once a compound is parsed as head and modifier, both concepts are retrieved from the knowledge base. When a concept is retrieved, not only are its features available, but also the diagnosticity values of every feature. PUNC constructs interpretations using the modification procedures outlined above (see also section below on Plausibility), with the feature’s of a compound’s constituents being processed in descending order of diagnosticity. Thus, interpretations constructed first are more likely to use features of high diagnosticity, with later interpretations using features of lower diagnosticity.

The diagnosticity values contribute to the overall acceptability of an interpretation. If features of high diagnosticity are used in an interpretation, this will result in a better acceptability score. Once output, acceptability scores are log transformed so that higher scores indicate more acceptable interpretations (e.g., Acceptability Score = $\log$(Model Output Score), where $j = 0.15$). The formula for calculating the overall acceptability is given in Figure 5.3 along with the pseudocode. As mentioned previously, diagnosticity is not a static phenomenon, and the diagnosticity of a feature can be affected by prior context, via LSA, which we have discussed in the previous section on context.
Informativeness in Comprehension

Previously we have stated that informativeness plays a central role in the comprehension system, acting to prevent uninformative interpretations from being generated. Rather than being implemented as a distinct process, informativeness is emergent from the interpretation generation process. In PUNC, informativeness prevents uninformative, duplicate information being used to construct candidate interpretations (within the four procedures of insertion, transfer, insertion-and-transfer, and interaction). Informativeness is intrinsic to the interpretation process itself, rather than existing as a separate check performed once all interpretations have been generated (c.f. Costello & Keane, 2000).

In generating interpretations for the compound *cactus hedgehog*, PUNC will use the concept *hedgehog* as the base for each possible interpretation. Therefore, each interpretation will begin with the features of *hedgehog* and those features inherited from hedgehog’s superordinates (e.g., *creature, entity*); among these will be the feature “is spiky”. During the generation process, “is spiky” will also be available as a feature of the concept *cactus*. In trying to create a candidate interpretation PUNC considers whether this feature adds new information to the base interpretation (of the concept *hedgehog*). As this feature is already part of the base interpretation, PUNC rejects the possibility of it forming a new interpretation. Clearly, the interpretation “a spiky hedgehog” is not informative with respect to what we already know of hedgehogs.

Essentially, duplicate features are never used to create interpretations in PUNC. This prevents uninformative interpretations from being generated, which in turn reduces the cost of generating possible interpretations. If PUNC were to consider duplicate features as forming valid interpretations, the number of candidate
interpretations would swell and would not reflect the kind of interpretations people produce.

**Plausibility in Comprehension**

We have mentioned that plausibility is a core factor in modelling comprehension in conceptual combination. In comprehension, plausibility is established through how closely the constituent concepts of a compound cohere in an interpretation; in other words, how closely the features of the concepts fit together. In PUNC, the plausibility of an interpretation is calculated based on one of the four modification procedures - insertion, transfer, insertion-and-transfer, and interaction. Through examples we describe how interpretations constructed using the different procedures are assigned varying plausibility scores.

Previously we described how the modification procedures would generate different interpretations for the compounds chocolate bee. For each candidate interpretation we describe how the level of plausibility is calculated, with plausibility scores ranging from 1 to 12 (12 being the least plausible). The four interpretations given previously were –

(a) a bee that produces chocolate (Insertion)

(b) a brown coloured bee (Transfer)

(c) a chocolate-coated bee (Insertion-and-Transfer)

(d) a bee that eats chocolate (Interaction)

The interpretation in (a) is assigned a score of medium plausibility (e.g., 6), where the modifier concept is inserted into a feature of the head concept in the base interpretation. Since bee has the feature “produces honey”, the modifier is inserted into this feature, overwriting the original value, to give the new feature
“produces chocolate”. As bees are known to produce things (i.e., honey), it is at least somewhat plausible that a bee could also produce chocolate. Based on this information, PUNC assigns a moderate plausibility score.

The interpretation in (b) is assigned a score of high plausibility (e.g., 3), where a feature from the modifier concept is transferred into the head concept, overwriting an existing value in the base interpretation. The “brown” colour feature from chocolate is transferred and overwrites the existing colour value in the head concept (i.e., black and yellow stripes). Because both concepts contain explicit colour information, there is greater coherence between them. If there was no colour feature in the head concept, and the modifier feature was simply transferred without overwriting some default value in the head concept, the assigned plausibility score would be lower since the coherence between the concepts is reduced. This interpretation receives a higher plausibility score than the interpretation in (a), because in (a) chocolate does not contain any features indicating how it is produced.

The interpretation in (c) is assigned a score of medium plausibility (e.g., 5), where the head concept is inserted into a feature of the modifier concept and then transferred to the base interpretation. The representation of chocolate has the feature “can coat things”, but the concept bee does not have any explicit information saying that it can be coated by chocolate or anything else. Therefore, PUNC can only give a score of medium plausibility. On the other hand, if the compound were chocolate cake, the cake concept would have the feature that it can be coated, and so the overall plausibility for the interpretation “a chocolate coated cake” would be greater.

The interpretation in (d) is assigned a score of high plausibility (e.g., 1), where a feature of the modifier concept interacts with a feature of the head concept. The interpretation is seen as one of high plausibility, since the representation of bee has
the feature “can eat things” that is reciprocated by the “can be eaten” feature in chocolate. Due to this coherence between the constituent concepts in this interpretation, PUNC assigns this interpretation a high plausibility weighting. If, on the other hand, the interpretation “a bee that eats steel” were generated for the compound steel bee, it would be assigned a lower plausibility weighting (e.g., 9) as the coherence between the constituents is lower (i.e., steel does not have a “can be eaten” feature).

In PUNC, the plausibility of an interpretation is based on how the interpretation is constructed (e.g., interaction or transfer), and to what extent the features of the constituent concepts cohere. If there is high coherence between them (e.g., “eats things” and “can be eaten”) then the plausibility is high, whereas if the coherence is low (e.g., “eats things” and “is not eaten”) then the plausibility is low. In this way, PUNC uses the prior knowledge of the constituents to ascertain how they can interrelate, which in turn dictates the overall plausibility of a candidate interpretation.

**Availability in Comprehension**

The availability of an interpretation refers to the likelihood of someone producing that interpretation for a given compound. Within the comprehension system, the availability of an interpretation arises through a combination of the diagnosticity and the plausibility of the interpretation, as we have outlined previously. An interpretation that is given a high diagnosticity score and a high plausibility score would be seen as an interpretation that would be considered highly available. Conversely, low scores of diagnosticity and plausibility would indicate an interpretation of low availability. Alternatively, interpretations that have either very high diagnosticity scores or very high plausibility scores (but not both) may also be of
high availability. Interpretations are not explicitly categorised as being of High-Availability or Low-Availability, rather, the higher the overall score an interpretation is assigned the more available it is considered.

Two candidate interpretations that PUNC generates for the compound *cactus beetle* are “a beetle that eats cacti” and “a green coloured beetle”. The first interpretation is assigned an acceptability score of 0.118 – given by \[ \text{log} [\text{Plausibility}(1 * 0.1) * \text{Diagnosticity} (11 * 0.55 + 5 * 0.45)] \] - which is better than the score of -0.96 of the second interpretation - \[ \text{log} [\text{Plausibility}(3 * 0.1) * \text{Diagnosticity} (4 * 0.55 + 4 * 0.45)] \] - and so is judged to be more available. The relatively low diagnosticity scores of the “is green coloured” feature in *cactus*, and “is black coloured” in *beetle* reduce the overall acceptability score, despite having relatively high plausibility. Therefore, the latter interpretation is considered less available.

As we have discussed, the availability of interpretations for a particular compound is not necessarily fixed; a biasing context may augment the relevance of specific features of a concept, thus making interpretations using these features more available. In PUNC, contextual activation is calculated using the Latent Semantic Analysis (LSA) model, which we have discussed in detail previously (see also Simulation 2, Chapter 5).

**Context in Comprehension**

PUNC uses LSA (Landauer & Dumais, 1997) comparisons to derive scores that represent the contextual activation of specific features of a concept, which we described in the introduction to the PUNC system. To interpret a compound in PUNC with reference to a specific context, the context must first be compared to all of the features of the compounds constituent concepts. Each comparison between the context and a feature returns an LSA contextual similarity score. The features are
then ranked by these LSA scores and the ranks are then used as weights on the diagnosticity values of the individual features. With the new diagnosticity values, PUNC interprets the compound (as we have described above), with each interpretation’s acceptability scores reflecting the influence of the current contextual environment.

5.3.3. Compound Production in PUNC

PUNC’s production system is essentially the inverse of the comprehension system. It takes a description as input and outputs a set of candidate noun-noun compounds. Additionally, PUNC outputs an acceptability score for each compound, which it calculates using variables of diagnosticity and plausibility. Compounds are created by PUNC using the production procedure described in the IPAC theory.

When PUNC is presented with a description of an object (e.g., “a bee that produces honey and lives in a cactus”), it first parses it and identifies the head noun (e.g., “bee”), the specified features (e.g., “produces honey”, “lives in a cactus”), and any potential modifiers (e.g., “honey”, “cactus”). Second, PUNC retrieves the head and modifier concepts (e.g., bee, honey, cactus) from the knowledge base, which accesses each concept’s set of diagnostic features (e.g., “has yellow and black stripes”, “produces honey”, “can sting”, etc. for bee). Third, PUNC examines the entity’s described features (e.g., “produces honey”, “lives in a cactus”) and eliminates any that are already diagnostic of the head (e.g., a bee normally “produces honey”). Lastly, PUNC produces candidate compounds by using the entity’s remaining diagnostic features to identify concepts that can be appropriately combined with the head.
This combination component represents the core of the production system. In PUNC, candidate compounds are produced by combining with the head those concepts that best represent the most diagnostic features of the entity in the description. This is done using four combination procedures, which are essentially the same as those used in the comprehension component of the system; insertion, insertion-and-transfer, transfer and interaction.

- Insertion involves checking if an entity feature is a variant of a head feature; for example, if the entity is described as “a bee that lives in a cactus” then this is a variant of the *bee* feature “lives in a hive”, and so the insertion procedure can label “a bee that lives in a cactus” as a *cactus bee*.

- Insertion-and-transfer involves checking if an entity feature is a variant of a given modifier’s feature; for example, if the entity is described as “a bee coated in chocolate” then this is a variant of the *chocolate* feature “cat coat things”, and so the Insertion-and-transfer procedure can label “a bee coated in chocolate” as a *chocolate bee*.

- Transfer involves retrieving a concept of which an entity feature is highly diagnostic; for example, if the entity is described as “a bee that has spikes” then “has spikes” is highly diagnostic of *cactus*, and so the transfer procedure can label “a bee that has spikes” as a *cactus bee*.

- Interaction involves finding interacting relations between the constituent concepts in the entity description. For example, the entity description “a beetle that eats cacti” contains an interaction between the two concepts – a beetle can eats things and a cactus can be eaten, thus resulting in the compound *cactus beetle*.
This overview of PUNC’s production system is further illustrated by the pseudocode in Figure 5.4. We now describe the system in more detail as we discuss how the influence of different factors is implemented by the model.

**Diagnosticity in Production**

As we discussed in the description of comprehension component of PUNC, diagnosticity values are encoded directly in the knowledge base. Therefore, when concepts that are mentioned in a description are retrieved, the diagnosticity of the features is automatically available. These diagnostic features play a role in the formation of candidate compounds with their diagnosticity values contributing to the calculation of a compound’s final acceptability (as shown in Figure 5.4). For example, the description “a spiky beetle” would result in the compound *cactus beetle* being constructed, as “is spiky” is a diagnostic feature of *cactus*. Because “is spiky” has a high diagnosticity score the resulting acceptability score will be quite high. Non-diagnostic features of a concept (e.g., features inherited from a parent concept) do not facilitate the production of candidate compounds. Although the feature “can photosynthesise” is part of the representation of *cactus*, the description “a beetle that can photosynthesise” will not activate *cactus* as a potential modifier. Omitting non-diagnostic features in this way prevents the production of many candidate compounds that would ultimately prove to be poor labels.
Figure 5.4 Pseudocode description of production component of PUNC

```
begin
parse description
    set head = noun1
    set features = feature1, feature2, etc.
    set mods = noun2, noun3, etc.
retrieve head concept
    retrieve mod concepts
    for each feature in description
        if feature present in head concept
            go to next feature
        end if
        if feature in head concept with different argument
            set diagnosticity = head feature diagnosticity
            // INTERACTION
            if feature interactive with mod feature
                set compound = mod + head
                set plausibility = high
            end if
            // INSERTION
            if feature can take mod as argument
                set compound = mod + head
                set plausibility = medium
            end if
        end if
    end for
    if feature in mod concept with different argument
        set diagnosticity = mod feature diagnosticity
        // INSERTION & TRANSFER
        if feature can take head as argument
            set compound = mod + head
            set plausibility = medium
        end if
    end if
    // TRANSFER
    if feature diagnostic of other concept in knowledge base
        retrieve concept
        set diagnosticity = concept feature diagnosticity
        set compound = concept name + head
        set plausibility = low
    end if
    acceptability = plausibility*0.25 * diagnosticity*0.5
end for
output compounds and acceptability scores
end
```
Informativeness in Production

As with the modelling of comprehension, informativeness is a core, emergent aspect of the production process. In the production system, uninformative compounds are never constructed. By examining the relations that exist between concepts in a description, and comparing these to features of the stored concepts themselves, PUNC can allow compounds to be formed based on whether the modifying concept provides some new information relative to the head concept. There is not an explicit informativeness module in PUNC. Rather, informativeness is established in concert with the plausibility stage (discussed below) of the model.

If we take the description “a pencil that is made of wood and is used for art”, PUNC would output *art pencil* as an acceptable compound label. However, the compound *wood pencil* would not be constructed, since “made of wood” is already part of the stored representation for *pencil*, and is therefore uninformative. As in the comprehension side of PUNC, informativeness is a binary affair; if a compound is output it is perceived as being at least somewhat informative.

Plausibility in Production System

As with the comprehension component of PUNC, plausibility is modelled based on the concept coherence of the features linking the head and modifier concepts (i.e., how well these features fit together). The plausibility of a compound is calculated using the four combination procedures described below – Interaction, Insertion, Insertion and Transfer and Transfer.

Previously we gave the examples of four compounds produced using the different mechanisms –

a) *cactus beetle* from “a beetle that eats cacti” (Interaction)
b) *cactus bee* from “a bee that lives in a cactus” (Insertion)

c) *chocolate bee* from “a bee coated in chocolate” (Insertion-and-Transfer)

d) *cactus bee* from “a bee that has spikes” (Transfer)

Example a) is assigned a high plausibility score (e.g., 1) because of the reciprocal interaction between the constituents of the compound. The use of both concepts in the description reflects a plausible interaction given the stored knowledge of these concepts (i.e., it is represented that *cactus* is edible and that *beetle* as a creature can eat things). If the compound *brick beetle* were created for the description “a beetle that eats bricks”, it would be assigned a lower plausibility score (e.g., 9) since *brick* would not be represented as being edible.

Example b) is assigned a score of medium plausibility (e.g., 6). In this case, the value “cactus” can be inserted into the existing feature of *bee* “lives in a hive”. If the representation of *cactus* specified that it could also be lived in, the overall plausibility would be higher (e.g., 3). However, without this information, *cactus bee* is only seen as having medium plausibility for this description.

Example c) is assigned a score of medium plausibility (e.g., 6). In this example, the feature in the description is a variant of a modifier feature (i.e., chocolate can coat things). As such, this compound is at least somewhat plausible. If insects were quite often coated in chocolate, this knowledge would elevate PUNC’s assigned plausibility score.

Finally, example d) is assigned a score of low plausibility (e.g., 8). In this case, a feature that is diagnostic of some other concept is being used to create the compound. As the representation of *bee* does not contain explicit knowledge of different textures that it might have (e.g., spiky, smooth etc.), this compound is
assigned only a low plausibility score. Although the resultant compound might be considered reasonably acceptable by PUNC, due to the high diagnosticity of “has spikes” in *cactus*, its plausibility remains low.

Using each of these procedures, PUNC assigns varying plausibility scores to each candidate compound that is produced. These plausibility scores are then combined with the diagnosticity scores of the compounds in order to assign a final acceptability score for each compound.

### 5.4. Testing the Model

Costello and Keane (2000, p326) describe several levels of detail at which a computational model can parallel people’s behaviour. At the most abstract level, a model can solve the problem inherent in the relevant phenomenon by specifying an effective procedure. At a less abstract level, a model can capture the broad shape of phenomena, such as capturing the effects of different classes of input. At the most specific levels of detail, a model can actually mirror people’s responses or reflect people’s response times and error patterns. As Costello and Keane suggest, very few models capture all of these levels of detail, and PUNC is no exception to this. In these simulations, while we do examine the specific interpretations and compounds people produce, we also look at the more general effects of different types of context and the effects differing degrees of plausibility in descriptions. In four simulations we consider PUNC’s performance with respect to the three pillars of conceptual combination: comprehension in isolation, comprehension in context and compound production.

Simulation 1 examines compound comprehension out-of-context, and compares interpretations generated by PUNC for particular compounds to those given
by people in Experiment 1 (Chapter 3). In Simulation 2, we turn to compound comprehension in context, and examine PUNC’s output for particular contextualised compounds and compare it to human data from Experiment 3 (Chapter 3). Simulation 3 then examines compound production, and compares compounds produced by PUNC for particular descriptions to those produced by people in Experiment 6 (Chapter 4). Finally, in Simulation 4 we test the model on itself by having PUNC produce labels for descriptions previously generated as interpretations, and examine how different factors affect this performance (see Experiment 8). These simulations show that PUNC successfully implements the factors that influence compound comprehension and production, and illustrate how PUNC closely models human performance in a series of conceptual combination tasks.

5.4.1. Simulation 1: Comprehension Out of Context

In this simulation, we test PUNC’s comprehension system by interpreting compounds presented out of context. We input noun-noun combinations into PUNC and note the interpretations generated and their respective acceptability scores. We then compare PUNC’s interpretations to those that people generated for the same compounds in Chapter 3’s Experiment 1, and examine the extent to which PUNC successfully mirrors human performance in out-of-context comprehension.

Method

Materials

We use 40 novel noun-noun compounds, taken from two sources in the literature; 24 compounds from Gerrig and Murphy (1992) and 16 from Costello and Keane (1997). These compounds had previously been given interpretations by human
participants (see Experiment 1, Chapter 3), which were then ordered by their frequency of production.

**Setup**

Novel compounds were input into PUNC. PUNC returned multiple interpretations for each compound with an associated acceptability score. We first compared the actual interpretations that PUNC output to those of the human participants, examining to what extent PUNC generated the same interpretations as people. We then compared the acceptability scores output by PUNC to the frequency of production data gathered from the human participants.

**Results & Discussion**

Overall, PUNC’s interpretations account for 46.8% of the interpretations generated by the human participants. However, when we remove those interpretations produced by only one person, PUNC’s coverage rises to 68.4%. We compared the PUNC acceptability scores to the frequency of production of the human interpretations a linear regression, giving an $r^2 = .64$, $p < 0.001$ (see Figure 5.5).

We found that in 77% of cases the most frequently produced interpretation by people was assigned the best acceptability score by PUNC. For example, for the compound *plate paper*, the interpretation that was ranked highest by PUNC was “paper that is used to make plates” (Acceptability Score = 0.84), which matched the most frequently produced interpretation by participants. For some compounds, PUNC’s highest ranking interpretation was not the interpretation produced most often by people, but was still among the set of interpretations people produced. For example, *bee hat* meaning “a hat that is worn by a bee” was considered a very good interpretation by PUNC (score = 0.671), but it was only the third most frequently produced interpretation by participants. At the other end of the scale, interpretations
that were given poorer scores by PUNC were often produced with less frequency by people. For example, a ballet mother as “a woman that is graceful” was given a relatively low acceptability score of -0.122, and was produced by only 5% of participants.

![Graph showing comparison between PUNC acceptability scores and human frequency of production](image)

Figure 5.5 Comparison of PUNC acceptability scores for interpretations and human frequency of production for the same interpretations.

These results demonstrate that PUNC can generate many of the interpretations produced by people in a conceptual combination interpretation task. Additionally, PUNC’s acceptability scores, calculated using the diagnosticity and plausibility scores for each interpretation, correlate well with the frequency of production of people’s interpretations. This finding suggests that PUNC’s acceptability score can be aligned with the likelihood of a particular interpretation being produced for a given compound.
5.4.2. Simulation 2: Comprehension In Context

In this simulation, we test the context factor in PUNC’s comprehension system by interpreting compounds presented in context. We give PUNC context stories and accompanying noun-noun compounds, and note the interpretations generated and their respective acceptability scores. We first look at PUNC’s scores in isolation and then compare scores for each interpretation to people’s responses in the paraphrase judgment task in Experiment 3 (Chapter 3). In this way, we examine the extent to which PUNC successfully models the effect of context on compound comprehension.

Method

Materials

We use 8 novel compounds, each with 2 interpretations; one of High-Availability and one of Low-Availability (calculated in Experiment 2, Chapter 3). These compounds and interpretations are then embedded in 3 different context conditions taken from Experiment 3 (Chapter 3); a Matched condition, an Omitted condition and a Mismatched condition. Each Matched condition is biased towards the particular interpretation that is being tested. For example, in a Matched condition, if the story makes reference to an interpretation $i_1$, then interpretation $i_1$ is the one that is being examined. An Omitted condition omits any reference to possible interpretations. For example, if the interpretation being examined is $i_1$, the preceding context will not be biased towards this or any other interpretation. A Mismatched condition is biased towards one interpretation, but then an alternative interpretation is tested. For example, if the preceding context is biased towards $i_1$ an alternative interpretation $i_2$ will be tested.
Setup

We first establish a baseline “no context” condition where PUNC generates interpretations for the 8 novel compounds. The scores and ranks of the High- and Low-Availability interpretations are noted for each compound. We then examine the effects of the different context conditions using Latent Semantic Analysis. For each compound, each context was compared in LSA to the features of the constituent concepts. The LSA scores that were returned were used to rank the features of both concepts. These rankings were then used as Context Weights on the original diagnosticity scores of the concepts’ features (see Context Section, Chapter 5). Thus, the LSA scores are reflecting the relative contextual activation of each concept’s individual features.

For each compound, we examined the effect of the preceding context on PUNC’s acceptability scores for the High- and Low-Availability interpretations. In the Matched condition we would expect that the biased context would result in an increase in acceptability relative to the baseline No Context condition. In the Omitted condition, we would expect the No Context orderings to be reflected, as the context is not biased towards either of the candidate interpretations. Finally, in the Mismatched condition we would expect each interpretation’s acceptability to be reduced compared to the baseline condition.

In the results section we first look at whether PUNC’s acceptability scores differentiate between the High- and Low-Availability interpretations, and the Matched, Omitted and Mismatched conditions. We then look at the effect of the preceding contexts on acceptability scores for the High- and Low-Availability interpretations in PUNC, and finally compare these to human responses from Experiment 3.
Chapter 5  Theory and Model

Results

There was a difference between the PUNC scores for High- and Low-Availability scores, with means of 0.2596 and -0.146 respectively - $F (1, 46) = 4.077$, $MSe = 1.088$, $p < 0.05$. Mean PUNC scores for the three context conditions were Matched = 0.2418, Omitted = -0.0401 and Mismatched = -0.0417, although the overall effect of context was only marginal – $F (2, 45) = 2.193$, $MSe = 1.103$, $p = 0.123$. However, the Matched condition was differentiated from both the Omitted and Mismatched conditions ($p = 0.05$).

An analysis of the interpretation ranks revealed that in the Matched condition, 75% of interpretations either increased their rank compared to the baseline condition, or were ranked as the top interpretation. In the Omitted condition, 68.7% of interpretations had the same rank as in the baseline condition. In the Mismatched condition only 37.5% of interpretations had their rankings reduced compared to the baseline. For example, in the No Context condition the Low-Availability interpretation for *skunk squirrel* is “a squirrel with a white stripe on its back”, with an acceptability score in PUNC of -0.122. However, when preceded by a biasing Matched context, its acceptability score is improved to 0.027. If, on the other hand, the interpretation is preceded by the unbiased Omitted context, the acceptability score is reduced to -0.122. Although this score is lower than in the Matched condition, it is in fact the same as the No Context baseline and the score in the Mismatched condition. Table 5.2 gives a PUNC scores for each interpretation in each condition, together with mean response times for those items in Experiment 3’s paraphrase judgement. When we compare PUNC’s rankings of the interpretations to the human response times from the paraphrase judgement task in Experiment 3, we find that
PUNC’s orderings reflected people’s responses 62.5% of the time (68.75%, 68.75% and 50% for the Matched, Omitted and Mismatched conditions respectively).

Table 5.2 PUNC scores and Human Response Times from Simulation 2.

Mean Scores and RTs Appear on the Bottom Line of the Table.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Interpretation Availability</th>
<th>No Context</th>
<th>Matched</th>
<th>Omitted</th>
<th>Mismatched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrot Bomb</td>
<td>High-Availability</td>
<td>0.0215</td>
<td>2164</td>
<td>0.0500</td>
<td>4029</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>0.2190</td>
<td>3001</td>
<td>0.2190</td>
<td>3416</td>
</tr>
<tr>
<td>Tulip Seed</td>
<td>High-Availability</td>
<td>-0.0644</td>
<td>2461</td>
<td>-1.659</td>
<td>7500</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>-0.7982</td>
<td>3226</td>
<td>0.6710</td>
<td>2053</td>
</tr>
<tr>
<td>Bee Hat</td>
<td>High-Availability</td>
<td>-0.2889</td>
<td>2794</td>
<td>-2.703</td>
<td>2573</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>-0.3867</td>
<td>2361</td>
<td>0.3151</td>
<td>2718</td>
</tr>
<tr>
<td>Lake Film</td>
<td>High-Availability</td>
<td>0.6710</td>
<td>2162</td>
<td>0.6710</td>
<td>2481</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>0.2435</td>
<td>2651</td>
<td>0.2435</td>
<td>2584</td>
</tr>
<tr>
<td>Father Card</td>
<td>High-Availability</td>
<td>0.6710</td>
<td>1901</td>
<td>0.6710</td>
<td>2771</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>-0.3010</td>
<td>2229</td>
<td>0.6006</td>
<td>2982</td>
</tr>
<tr>
<td>Arrow Window</td>
<td>High-Availability</td>
<td>-0.1342</td>
<td>2794</td>
<td>-0.1218</td>
<td>3778</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>0.0992</td>
<td>2924</td>
<td>0.0992</td>
<td>3623</td>
</tr>
<tr>
<td>Rhinoceros Horse</td>
<td>High-Availability</td>
<td>0.6006</td>
<td>2270</td>
<td>0.6006</td>
<td>1987</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>-0.1218</td>
<td>2935</td>
<td>-0.2411</td>
<td>6781</td>
</tr>
<tr>
<td>Shnuk Squirrel</td>
<td>High-Availability</td>
<td>0.6006</td>
<td>2343</td>
<td>0.6006</td>
<td>2308</td>
</tr>
<tr>
<td></td>
<td>Low-Availability</td>
<td>-0.1218</td>
<td>2507</td>
<td>0.0270</td>
<td>5124</td>
</tr>
</tbody>
</table>

Discussion

In this simulation we used LSA in conjunction with PUNC to simulate the effects of different preceding contexts on interpretation judgements. The rankings of LSA scores were used as context weights to effect changes on the diagnosticity levels of individual concept’s features. The changes to individual features resulted in changes in the final acceptability scores of the interpretations that PUNC output. We found that the model was able to differentiate between High and Low-Availability interpretations in context, and between the Matched condition and the other context conditions. The model appeared to reflect human performance best in the Matched and Omitted conditions, but was comparatively poor in the Mismatched condition. This difficulty may in fact reflect human difficulties in this condition. In the original
Experiment 3, people were considerable slower at making judgements in the Mismatched condition and there were also many more erroneous responses.

Overall, PUNC performs well by using LSA to alter the diagnosticity scores of individual features. However, in some cases the context weightings were insufficient to change the rankings / scores of interpretations output by PUNC. This could be due to the high degree of similarity between the different context stories. Each story differed by only one line, which may have been too fine a distinction to allow LSA to differentiate between relevant and irrelevant features. For example, the two interpretations for *lake film* (“a film that is about lakes”, “a film that is made near a lake”) are quite similar, which means that LSA will assign similar scores, compared to interpretations that are more distinct. However, PUNC’s performance is quite strong considering these interpretations are ones that people actually produce for these compounds, and not simply experimenter derived interpretations that, in reality, are not usually generated.

### 5.4.3. Simulation 3: Compound Production

In this simulation, we test PUNC’s production system by creating compound labels for described objects. We give PUNC simple descriptions of objects, and note the compounds produced. We then compare PUNC’s compounds to those produced by people for the same descriptions in Chapter 4’s Experiments 6 and 7, and examine the extent to which PUNC successfully models human performance in compound production.
Method

Materials

We use 44 pairs of descriptions which had previously been categorised as High- and Low-Plausibility descriptions in Experiment 5.

Setup

The descriptions were input to PUNC and we noted the compounds that were produced and the acceptability score for each compound. We compared the overlap of the compounds produced by PUNC to those produced by people in Experiment 6.

Results & Discussion

Overall, PUNC produced 43.7% of the compounds produced by people in Experiment 6. If we remove unique compounds from people’s responses this rises to 69.3%. The mean number of unique compounds produced for High-Plausibility items was 1.75, compared to 1.98 for Low-Plausibility items – $t(43) = 1.81$, $p = 0.038$. In 81.8% of cases PUNC produced the same number or fewer compounds for High-Plausibility items as Low-Plausibility ones. Table 5.4 shows the number of compounds produced for each description and the mean acceptability score calculated by PUNC. The difference between PUNC’s acceptability scores for compounds output for High- and Low-Plausibility descriptions was not reliable ($p > 0.3$).
## Table 5.3 Number of unique compounds produced by PUNC for each description, together with mean acceptability scores.

<table>
<thead>
<tr>
<th>Item</th>
<th>Unique Compounds</th>
<th>PUNC Score</th>
<th>Unique Compounds</th>
<th>PUNC Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.3151</td>
<td>1</td>
<td>0.3151</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.0982</td>
<td>2</td>
<td>0.0982</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.4700</td>
<td>2</td>
<td>-0.0502</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.3151</td>
<td>3</td>
<td>0.3151</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.4700</td>
<td>2</td>
<td>0.6710</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.0364</td>
<td>2</td>
<td>1.2137</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.0215</td>
<td>3</td>
<td>-0.0828</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>-0.2172</td>
<td>2</td>
<td>-0.2172</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.0440</td>
<td>3</td>
<td>0.0440</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.3151</td>
<td>2</td>
<td>0.1447</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.6710</td>
<td>2</td>
<td>0.6710</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.1956</td>
<td>2</td>
<td>0.1956</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>-0.0961</td>
<td>2</td>
<td>0.5687</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>0.1956</td>
<td>2</td>
<td>0.1956</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.6710</td>
<td>1</td>
<td>0.6710</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>-0.0502</td>
<td>2</td>
<td>-0.0502</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0.6710</td>
<td>2</td>
<td>0.0215</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0.0982</td>
<td>2</td>
<td>0.0982</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>0.1243</td>
<td>2</td>
<td>1.0364</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>-0.0156</td>
<td>2</td>
<td>-0.2172</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>-0.2640</td>
<td>1</td>
<td>-0.2640</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>0.0440</td>
<td>1</td>
<td>-0.0502</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>0.5687</td>
<td>2</td>
<td>0.5687</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>0.5687</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0.0982</td>
<td>3</td>
<td>0.0982</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>-0.1698</td>
<td>1</td>
<td>-0.1698</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>-0.2640</td>
<td>1</td>
<td>1.2137</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>-0.2640</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>1.2137</td>
<td>2</td>
<td>1.2137</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>0.5687</td>
<td>2</td>
<td>0.5687</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>0.0215</td>
<td>2</td>
<td>0.0215</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>0.3869</td>
<td>2</td>
<td>0.1956</td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>0.1447</td>
<td>2</td>
<td>0.0215</td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>1.2137</td>
<td>2</td>
<td>1.4017</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
<td>-0.0156</td>
<td>2</td>
<td>-0.0156</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>0.3151</td>
<td>2</td>
<td>0.3151</td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td>-0.1698</td>
<td>1</td>
<td>-0.0502</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>0.5687</td>
<td>2</td>
<td>0.1956</td>
</tr>
<tr>
<td>39</td>
<td>2</td>
<td>0.2352</td>
<td>3</td>
<td>0.2352</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>0.3151</td>
<td>2</td>
<td>0.3869</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
<td>-0.1848</td>
<td>2</td>
<td>0.0215</td>
</tr>
<tr>
<td>42</td>
<td>2</td>
<td>0.0215</td>
<td>2</td>
<td>0.0215</td>
</tr>
<tr>
<td>43</td>
<td>3</td>
<td>0.0440</td>
<td>2</td>
<td>-0.0828</td>
</tr>
<tr>
<td>44</td>
<td>1</td>
<td>-0.4788</td>
<td>3</td>
<td>-0.0502</td>
</tr>
</tbody>
</table>

Means 1.75 0.2239 1.98 0.2628
It seems that PUNC successfully mirrors the trend for greater convergence in response to High-Plausibility descriptions, but fails to reflect the difference in confidence ratings that people exhibited. So why does PUNC perform well in one regard, but fail in another? It appears that PUNC produces more compounds in response to Low-Plausibility because these descriptions tend to activate more candidate modifiers overall, primarily due to the activation of other concepts whose diagnostic features have been referred to in the descriptions. However, this in itself is not enough to explain the difference. It also appears to be the case that many candidate modifiers for High-Plausibility items are rejected on the basis of informativeness, and so reducing the total number of compounds output. If we take the example descriptions “a lobster that catches shrimp with its claw” (High-Plausibility) and “a lobster that catches shrimp with its web” (Low-Plausibility), we can see how differences arise. For the High-Plausibility description, both *shrimp* and *claw* are activated as possible modifiers. However, because *claw* is part of the concept *lobster*, it is deemed uninformative and rejected, leaving *shrimp lobster* as only compound output by PUNC. For the Low-Plausibility description, *shrimp* and *web* are initially activated as candidate modifiers. Additionally, *spider* is also activated, since “uses a web” is a diagnostic aspect of *spider*. Each of these candidate modifiers are used to form three compounds – *shrimp lobster*, *web lobster* and *spider lobster* – as they are all considered informative. So, through a combination of a greater number of candidate modifiers being activated (using the diagnostic features of other concepts) for Low-Plausibility descriptions, and the elimination of uninformative modifiers for High-Plausibility descriptions, PUNC reflects the different levels of convergence evident in people’s compound production in Experiments 6 and 7.
PUNC’s failure to find a difference between the acceptability scores of High- and Low-Plausibility items stems from the fact that many of the candidate compounds are formed using the same construction methods (e.g., interaction, transfer etc.). This means that these compounds are being assigned identical plausibility scores, with the diagnosticity scores failing to discriminate between different compounds. From our previous example, *shrimp lobster* is produced by PUNC in response to both High- and Low-Plausibility descriptions. However, the compound is assigned the same score in each case because, firstly, it is formed using the interaction process in both cases (e.g., lobster eats things, shrimp can be eaten), and secondly, the diagnosticity scores for the features used in creating the compound are obviously the same, since the same concepts are being used. Therefore it seems that the gradations in plausibility are too coarse-grained to distinguish between the acceptability of compounds in the same way that people do.

**5.4.4. Simulation 4: Turning PUNC on Itself**

In this simulation, we test the integration of PUNC’s comprehension and production systems by feeding one system’s output into the other. We take the interpretations produced by PUNC in Simulation 1 and input them into the production system, and examine the extent to which the compounds generated reproduce the original noun-noun combinations. In this way, we compare how PUNC can label interpretations it has previously generated.

**Method**

**Materials**

We use 12 pairs of interpretations (one of High-Availability and one of Low-Availability), which were generated in Simulation 1, as descriptions for PUNC’s
compound production component. We examine the frequency with which PUNC produces the target compound that generated these interpretations in the first place. In Experiment 8, we found that people were more likely to produce a target compound if the interpretation used for the description was one that was of high availability.

Setup

We input the 24 descriptions to PUNC’s compound production component. We noted the number of times PUNC produced the target compound for the High- and Low-Availability descriptions. We then compare the acceptability scores assigned by PUNC for these compounds.

Results & Discussion

We found that PUNC produced target compounds for High-Availability descriptions in 100% of cases, compared to only 75% of the time for Low-Availability descriptions. Human participants produced the target compounds 50% of the time for High-Availability descriptions and only 28% of the time for Low-Availability descriptions (see Experiment 8, Chapter 4). Although the difference in PUNC is exaggerated compared to the human results, the finding that target compounds are more likely to be produced in response to descriptions from High-Availability interpretations holds true. There was no reliable difference between the acceptability scores assigned by PUNC the two classes of description (p > 0.1). While this is consistent with our findings from Simulation 3 and Experiment 8, this is more likely to be reflecting PUNC’s inability to distinguish the acceptability of compounds due to the coarse-grained nature of the plausibility scale.

It is clear that some interpretations are better than others for novel compounds. For any given compound, people are more likely to produce one compound over another compound. However, it also seems to be the case that there exists a two-way
relationship between these interpretations and the originating compound. We have found that the more often an Interpretation X is produced for Compound Y, the more likely it is that Compound Y will be produced as a suitable label for Description X (see Experiment 8, Chapter 4). The PUNC model reflects this relationship, using comprehension and production systems that have been implemented independently, but share the same knowledge base and utilise the same factors (i.e., diagnosticity, informativeness, plausibility) to interpret and produce novel compounds. This reciprocality, found in human responses, provides a validation of the theoretical principals of the IPAC theory and also of the computational implementation of this theory in PUNC.

5.4.5. Summary of Simulations

In our four simulations we have seen PUNC simulate a variety of empirical data in both comprehension and production task in conceptual combination. In Simulation 1, we saw that PUNC reflected the interpretations that people produced for a set of novel compounds, with PUNC’s acceptability scores also correlating well with the original frequency of production of those interpretations. In Simulation 2, PUNC (in conjunction with LSA) reflected the effects of different types of contextual environment on the availability of specific interpretations. In Simulation 3, PUNC produced compounds for a set of descriptions that mirrored those produced by people in earlier production experiments, but failed to reflect human ratings for compound acceptability. Finally, in Simulation 4, we used PUNC’s own interpretations as descriptions showing that the model reflects the symmetries seen in people’s responses for production and comprehension tasks.
5.5. Comparing PUNC to Other Models

PUNC differs from previous models of conceptual combination in a wide variety of ways. Firstly, PUNC is the first model to deal with compound production as well as compound comprehension, as other models have previously only approached comprehension (e.g., Compound Specialisation: Cohen & Murphy, 1984; Murphy, 1988; CARIN: Gagné, 2000; Gagné & Shoben, 1997; C3: Costello & Keane, 2000). Secondly, PUNC is the first model to deal with the effects of context, as previous systems (i.e., Compound Specialisation, CARIN, and C3) have only modelled the comprehension of compounds in isolation. Additionally, PUNC is fundamentally different to previous models in a number of other ways, and we now discuss these differences in detail with specific reference to each model.

5.5.2. PUNC versus Compound Specialisation

The Concept Specialisation theory was developed by Murphy and Cohen (Cohen & Murphy, 1984; Murphy, 1988). Superficially there are some similarities between PUNC and Concept Specialisation (CS), but there are radical differences between the two approaches. First, Concept Specialisation, while given a detailed treatment has never been implemented as a computational model. Second, Concept Specialisation generates only relational interpretations for compounds, by filling slots in the head concept with the modifier name (e.g., “a hawk that hunts robins” for robin hawk). This means that, unlike PUNC, Concept Specialisation cannot create property-based interpretations (e.g., “a hawk with a red breast”), or use the head concept to fill possible roles in the modifier concept (e.g., chocolate bee meaning “a chocolate coated bee”). Third, Concept Specialisation does not allow for particular features of concepts to be considered more important than others. PUNC, on the other
hand, uses diagnosticity values to indicate the importance of certain features over
others and to contribute to the calculation of the overall acceptability of
interpretations. Similarly, any interpretation that is constructed in CS is considered
plausible; contrasting with PUNC, where there gradations of plausibility for
interpretations. Finally, CS posits a single-interpretation view of conceptual
combination, while PUNC is based on the core idea that any compound can give rise
to multiple interpretations.

5.5.3. PUNC versus CARIN

The CARIN model is perhaps the most distinct of the existing models of
categorization, focusing exclusively on the distributions of relations that
exist between a compound’s constituents. CARIN is a mathematical model of
categorization, in contrast to PUNC’s computational instantiation. CARIN
gives the modifier concept in a compound the most important role, using differences
in modifier relation frequency as a predictor of response time differences between
different compounds. In PUNC, the features of both the head and modifier concepts
work in partnership to construct interpretations and to establish the relative
acceptability of the multiple interpretations for a compound. That is, there is no
predefined set of templates implemented in PUNC, which on the other hand has
proven restrictive for the interpretations accounted for by CARIN (Costello & Keane,
2000, p335).

Like Concept Specialisation, CARIN deals only with relational interpretations,
ignoring the large class of property-based interpretations. Finally, the CARIN model
has largely been tested on non-novel compounds (see Tagalakis & Keane, 2003),
largely ignoring the effect that this familiarity has on interpreting / responding to such compounds.

5.5.4. PUNC versus C3

C3 (Costello & Keane, 2000) resembles PUNC more closely than the other models do because of its use of diagnosticity, informativeness and plausibility. At the computational level, the theories (Constraint Theory and IPAC) are very similar, but at the algorithmic level PUNC implements these factors very differently to C3.

Previously we have described the structure of the PUNC model – detailing both how it interprets novel combinations and how it produces compounds, given a description of an object. As PUNC is an implementation of the IPAC theory, which has been inspired by Constraint Theory of conceptual combination, it is worth outlining the primary differences between PUNC and C3, which was the first implementation of this theory. C3 does not have a compound creation component, so we focus only on the differences between the compound interpretation aspects of the two models. Although each model implements diagnosticity, informativeness and plausibility, both approaches are quite distinct at each stage of processing.

C3 uses a handcoded knowledge base of concepts and their associated features to calculate diagnosticity. Diagnosticity is carried out in a pre-processing stage by counting the number of times each feature occurs in all of the concepts in the knowledge base. So, if a feature occurred only once in the knowledge base, it would have a high diagnosticity score. However, this approach can lead to problems of underspecification of diagnostic features. For example, people judge the feature “large” as being highly diagnostic of whales (Costello & Keane, 2001), but C3 may
not assign an appropriate diagnosticity score to this feature. Because many concepts can have the feature “large” as part of their representations (e.g., cinema, building, boulder, train and other concepts are all encoded in C3 as having a “large” feature), the feature would not receive and appropriately high diagnosticity score. This would have the knock on effect of skewing the acceptability scores applied to interpretations using such features. On the other hand, PUNC’s diagnosticity values for individual features are derived from human raters. This means that features such as “large” are encoded as being highly diagnostic for relevant concepts (e.g., whale), regardless of their relative frequency of occurrence across the knowledge base. Overall, diagnosticity values should be more appropriate. Although the two models establish diagnosticity values in different ways, they both prioritise the role of diagnosticity in creating interpretations and in establishing their overall acceptability.

The Informativeness constraint in C3 is applied to the set of interpretations as a post-processing stage, in order to eliminate non-informative interpretations from the vast set generated by the model. In this respect, informativeness is considered the least important of the three constraints (p. 66, Costello, 1996). Studies have shown that people do not habitually produce uninformative interpretations (Costello & Keane, 1997; Downing 1977), as this is a considerable drain on cognitive processing and a violation of pragmatic principles. So while no participant would produce the interpretation “a bed that is made of wood” for pencil bed (Costello & Keane, 1997), C3 will still generate such interpretations first, and then later eliminate it as being uninformative.

PUNC implements informativeness in a more integrated, cognitively parsimonious manner. Quite simply, PUNC does not generate uninformative interpretations. In the compound cactus hedgehog, for example, both concepts would
have the feature “is spiky”. Since hedgehog already contains this feature, PUNC would not attempt to create the interpretation “a hedgehog that is spiky”. This means that PUNC can quickly generate a set of several interpretations for a single compound, while ensuring all of the interpretations are informative. In this way, PUNC reflects people’s interpretation of compounds under normal circumstances. Integrating informativeness into the generation mechanism makes it a primary constraint in interpreting compounds and not a secondary mechanism as it is in C3.

In C3, Plausibility is calculated using the overlap of an interpretation with stored exemplars of concepts in memory. In this way, the greater the overlap the higher the plausibility. Thus, plausibility in C3 is a measure of how plausible a particular entity could exist given prior knowledge. The primary disadvantage of the plausibility component is that it leads to a ballooning in the number of candidate interpretations. This in itself is not a problem, except that many of these interpretations will ultimately be rejected as being uninformative.

In PUNC, the plausibility of an interpretation is calculated with reference to the concept-coherence between the constituent concepts (see Connell & Keane, in press). This means that interpretations are only constructed on the basis of some plausible interaction between the concepts of the compound. This reduces both the number of candidate interpretations and the computational cost in constructing a set of interpretations, ultimately providing a closer reflection of the number of interpretations that people normally generate for novel compounds. It is worth mentioning that with C3’s incremental plausibility mechanism, the resultant interpretations can often end up containing the features of the original compound’s head concept. By using the head concept as a starting point for each candidate interpretation, PUNC obviates the need for the incremental plausibility component of
C3. This upshot of this is that C3 can generate more elaborate interpretations, compared to PUNC. We have found, however, that people exhibit a preference for concise, specific interpretations (see Experiment 1, Chapter 3). So while people certainly have the capacity to provide elaborate interpretations, perhaps in response to specific task demands, it is not usually the approach people adopt.

5.6. Conclusion

In conclusion, it is clear that both the IPAC theory and PUNC model offer a unique perspective on conceptual combination, providing an important theoretical link between the understanding of novel compounds and their production. Previous theories and models of conceptual combination have failed to address important aspects of conceptual combination, including the effects of context on interpretations and the parallels between the comprehension of novel compounds and their production.
6. CONCLUSION

6.1. Summary of Accomplishments

We began this thesis with two questions – How do we understand things we’ve never heard before? How do we describe things we’ve never encountered? From these two straightforward questions we have delved into many aspects of conceptual combination, addressing several critical imbalances in the existing literature. In this section we provide a summary of the novel empirical findings in both comprehension and production, providing an overview of the IPAC (Integrated Production and Comprehension) theory which encompasses both sides of the conceptual combination coin. Next, we summarise the performance of the PUNC model in simulating a variety of phenomena from earlier empirical investigations. Finally, we offer some possible directions for future research in conceptual combination and related areas.

6.1.1. Empirical Novelties

We have detailed a series of experiments in Chapters 3 and 4, which have looked at the three pillars of conceptual combination, including out-of-context interpretation, in-context interpretation and compound production.

In Chapter 3 we looked at the interpretation of novel compounds both in and out of context. We used an offline task to examine the interpretations people produce,
and in two online tasks to examine how quickly people make judgements for interpretations that differ in terms of their availability. In Chapter 4 we looked at people’s creation of novel compounds, something which has received scant attention in the conceptual combination literature. We used an offline task where people created compounds for short descriptions and also rated their responses. A number of empirical novelties resulted from these experiments:

- Although novel compounds are highly polysemous, with people producing a wide variety of interpretations for any given novel compound, the majority of people tend to converge on a single interpretation (Experiment 1).

- Interpretations that are produced more often out-of-context are responded to more quickly in an online paraphrase judgement task (Experiment 2).

- Comprehension of novel compounds in context is affected by both the out-of-context availability of an interpretation and the type of context in which the compound is embedded (Experiment 3). Interpretations that are responded to most quickly out-of-context are also responded to most quickly in-context. People’s responses to interpretations in context are also affected by the degree to which the presented paraphrase matches with the preceding context. People respond quickest when the paraphrase matches completely, less quickly when the preceding context is not biased towards any specific interpretation and slowest when the preceding context conflicts with the meaning of the paraphrase.

- People have greater confidence in compounds they produce with reference to descriptions of high plausibility (Experiments 6 and 7).
• People exhibit greater convergence when creating novel compounds for highly plausible descriptions (Experiments 6 and 7).

• People produce fewer noun-noun compound labels in response to highly plausible descriptions (Experiments 6 and 7).

• The compounds people produce are not dependent on the specific word order of the descriptions. Rather, it is the concepts that are being referred to that are the driving force behind people’s behaviour (Experiment 7).

• Given a description derived from an interpretation, people are more likely to create the originating compound, if that description was frequently produced as an interpretation out of context (Experiment 8).

6.1.2. Theoretical Novelties

In Chapter 5 we outlined a theory of conceptual combination (IPAC), Integrated Comprehension and Production. This formulation can be viewed as an extension to Costello and Keane’s (2000) Constraint Theory. Theoretically the primary considerations are:

• The comprehension and production of novel compounds is influenced by many of the same factors, including informativeness, diagnosticity, plausibility, familiarity and context.

• People follow pragmatic principles of communication; they do not produce uninformative compounds or interpretations. Conversely, they aim to produce concise compound labels and interpretations.
• Diagnostic features of concepts are more likely to be used in constructing interpretations and compounds, than are non-diagnostic features of concepts.

• The plausibility of a description affects the compound production process, impacting both convergence and confidence.

• Many factors contribute to the overall acceptability of an interpretation or compound, including diagnosticity, informativeness, plausibility and availability.

Finally, in Chapter 5, we described a computational instantiation of the IPAC theory – the PUNC model (Producing and Understanding Novel Combinations) – which implements the core factors of the theory in an attempt to capture important phenomena of comprehension and production in conceptual combination. In a series of simulations we tested PUNC’s performance in tasks concerning the comprehension (in and out of context) and production of novel compounds. PUNC performed favourably, successfully modelling several aspects of conceptual combination, including –

• The range of interpretations that people produce given a novel compound (Simulation 1).

• The relative frequency with which people produce specific interpretations for novel compounds (Simulation 1).

• The effects of different types of context on the availability of interpretations (Simulation 2).
• The compounds people produce when providing labels for descriptions (Simulation 3).

• The different degrees of convergence observed in compounds produced by people, as a result of greater or lesser plausibility in the objects that are being described (Simulation 3).

• The likelihood of target compounds being produced, based on the availability of the input description as an interpretation for that compound (Simulation 4).

We have found that although there are obvious asymmetries between the comprehension and production sides of conceptual combination, there are several high-level principles and factors that bind both together. These common factors, such as Diagnosticity, Informativeness and Plausibility provide people with the mechanisms necessary for both understanding and creating novel combinations.

6.2. Further Research

In the course of this research we have addressed aspects of conceptual combination that have not received adequate attention in the literature. We have drawn together many disparate facets of conceptual combination to provide a unitary theory of production and comprehension. In doing this, we have opened the door to a host of other issues concerning language creativity in general, the relationship between language comprehension and production and the modelling of such phenomena. Below, we outline some related areas of language research that have yet to be fully addressed.
6.2.1. Phonetic Analogies and Humour

When people understand novel compounds, they can often rely on familiar, existing, or analogically related compounds in order to serve their comprehension needs (Gagné & Shoben, 1997; Tagalakis & Keane, 2003; Tagalakis, Keane & Lynott, 2003). People might understand *mobile virus* as “a digital virus that infects mobile phones” with direct reference to the existing compound *computer virus* (“a digital virus that infects computers”). In addition to this kind of semantic analogy, people’s comprehension can be facilitated by additional phonetic information. For example, the compound *punishment meetings* meaning “difficult meetings between nationalist and unionist politicians in Northern Ireland”, refers to the existing phrase *punishment beatings* meaning “retaliation beatings from nationalist paramilitaries to loyalist paramilitaries, or vice versa”. The added phonetic similarity implies much more information than a compound without this information e.g., *difficult meetings*. It would take a much longer phrase to convey the same amount of information if this phonetic information was missing e.g., Northern Irish bipartisan meetings, perhaps. Such metaphoric usage is not only creative, but appears to be highly efficient in communicating a complex idea. Additionally, there is a humorous aspect to such compound usage (see Veale, 2003), with the coiner poking fun at the participants of such meetings, drawing attention to the uncooperative nature of some of the contributors. Another example, *hostile makeover* - referring to where someone has been coerced into having makeup applied and having their hair done, against their wishes – uses the existing compound *hostile takeover* to embellish the intended meaning for humorous effect. While it may be possible to understand these novel combinations by direct analogy, there is certainly an extra level of meaning where there is also a high degree of phonetic similarity between the words used. It remains
an open issue whether it is the simply the use of phonetically similar words that gives rise to the additional humour value, or if it is the fact that the phonetics facilitate the retrieval of relevant information from the base domain.

Humour, in general, is a side of conceptual combination that has been largely ignored (cf. Veale, 2003). In many instances novel compounds are used as humorous devices. Comedy classics such as The Simpsons, Fr. Ted and Blackadder, have made frequent use of novel combinations in order to amplify the humorous content of a scene. Blackadder spoke of *arrow fodder* to Baldrick, when discussing an imminent battle. Fr. Ted discovered that Mrs. Doyle had made a *cake jumper*, by baking a cake with a jumper in the mixture. The Simpsons have eaten *floor pies* (i.e., a pie that has fallen on the floor) and *trash cookies* (i.e., cookies retrieved from the garbage), while the Flanders endured an *imagination Christmas* when they couldn’t afford to buy gifts that year. In a similar vein, tabloid headline writers often use novel compounds as a humorous device. Take the example, *Foot heads arms body*, referring to the Labour politician Michael Foot who was made head of a committee to investigate arms deals. Further investigation of conceptual combination as a device for humour could provide many insights, not only into theories of humour, but also into analogy and creative thought in general.

### 6.2.2. Creativity in Conceptual Combination

The above examples all stem from the wider issue of creativity in conceptual combination. It is obvious that people can be creative when interpreting and creating these compounds, but it is not obvious what factors contribute to this creativity. A view espoused by Johnson-Laird (see Haught & Johnson-Laird, 2003) is that enhanced creativity emerges when people are forced to operate within greater
Chapter 6

Conclusion

constraints. This refers not only to language, but also to music composition and improvisation (Johnson-Laird, 2002). On the other hand, it may be the case that people are more creative when constraints are relaxed, thus giving them greater freedom to come up with more unusual meanings, interpretations or expressions. In our empirical work, we found that people exhibited less convergence when compounds were created in response to less plausible descriptions. This reduction in convergence could be viewed as an increase in people’s creativity. In this sense, increased plausibility could be construed as a constraint on people’s labelling options, limiting the creative potential. When this constraint is relaxed, people can exercise their more creative side.

Again, these issues have yet to be fully explored in conceptual combination research, but there may be something to be drawn from existing theories of creativity, such as those of Sternberg (1988) and Ward et al (see Finke, Ward & Smith, 1992; Ward, Smith & Vaid 1997).

6.2.3. Extending PUNC

In its current incarnation, PUNC can comprehend and produce novel compounds, taking into account contextual influences in comprehension. Possible extensions to the model might include the capacity in create interpretations analogically, both using semantic relatedness (see Tagalakis, Keane & Lynott, 2003) and phonetic information, as we described above. As we pointed out, many such analogical combinations are intended to be humorous; it would be interesting to examine the extent to which a model could predict the level of humour that particular compounds invite. So, rather than ordering compounds and interpretations by how acceptable they are, the model might order them based on how creatively or
humorously they will be perceived. This issue could be closely connected to the notion of tightening or loosening constraints in order to enhance creativity.

On the compound production side, factors such as familiarity and context effects have yet to receive the necessary empirical investigation that would allow for their incorporation in a computational model. The IPAC theory has outlined possible roles of these factors, but they remain open empirical issues.

6.2.4. Linking Comprehension and Production

Traditionally, language researchers have treated the systems and processes that govern comprehension and production as being quite distinct. We have shown how high-level constraints bind both systems, and how both sides of language can function using the same representations, at least with respect to conceptual combination. It is clear that a full understanding of our linguistic processes will require the integration of both systems. In the last few years, some researchers have taken this step, and attempted to describe ways in which comprehension and production influence each other (Branigan, Pickering & Cleland, 2000; MacDonald, 1999; Pickering & Garrod, in press). For example, Pickering and Garrod (in press) have posited that people’s representations for comprehension and production become aligned during dialogue, thus facilitating more efficient communication. As a microcosm of language usage, conceptual combination could prove to be a crucial test bed for grand theories of language.

6.3. Conclusions

Conceptual combination is a fundamental feature of our language. It assists in the growth of language and allows us to encapsulate complex ideas in manageable
linguistic parcels. As a phenomenon it has proven to be a rich source, both in terms of data and theories, in such diverse areas as categorisation, analogy, creativity and similarity. This thesis has provided an inclusive view of production and comprehension, raising many questions for future research. We hope that investigations into conceptual combination will continue to motivate research in related areas in cognitive science that will also benefit from a deeper understanding of how we contend with language we’ve never heard before.
APPENDIX A – EXPERIMENT 1 MATERIALS

The following noun-noun compounds were used as materials in Experiment 1.

1. apple frame
2. arrow window
3. ballet mother
4. bed chair
5. bee hat
6. bottle furnace
7. bumblebee moth
8. cactus oak
9. carrot bomb
10. doctor drum
11. dog smile
12. elephant pig
13. factory preacher
14. father card
15. foot chip
16. fork night
17. frog beetle
18. giraffe antelope
19. helicopter aeroplane
20. hour hammer
21. jade saw
22. kangaroo monkey
23. lake film
24. leg punishment
25. magazine concert
26. octopus lobster
27. parrot robin
28. plate paper
29. rain canary
30. rhinoceros horse
31. skunk squirrel
32. snail spider
33. snake iguana
34. sofa wall
35. sugar line
36. trumpet olive
37. tulip sled
38. viper slug
39. vodka face
40. whale seal
APPENDIX B – EXPERIMENT 2 MATERIALS

The following noun-noun compounds and their accompanying high- and low-availability interpretations were used as materials in Experiment 2:

<table>
<thead>
<tr>
<th>Availability</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>An “arrow window” looks the same as an arrow.</td>
</tr>
<tr>
<td>low</td>
<td>An “arrow window” has an arrow pattern on it.</td>
</tr>
<tr>
<td>high</td>
<td>A “bee hat” has the same shape as a bee.</td>
</tr>
<tr>
<td>low</td>
<td>A “bee hat” has an image of a bee on it.</td>
</tr>
<tr>
<td>high</td>
<td>A “carrot bomb” is an explosive that looks like a vegetable.</td>
</tr>
<tr>
<td>low</td>
<td>A “carrot bomb” is an explosive that is made from vegetables.</td>
</tr>
<tr>
<td>high</td>
<td>A “father card” is for a father.</td>
</tr>
<tr>
<td>low</td>
<td>A “father card” is a large card.</td>
</tr>
<tr>
<td>high</td>
<td>A “foot chip” is an injury to your foot.</td>
</tr>
<tr>
<td>low</td>
<td>A “foot chip” is put into your footwear.</td>
</tr>
<tr>
<td>high</td>
<td>A “giraffe antelope” has a long neck.</td>
</tr>
<tr>
<td>low</td>
<td>A “giraffe antelope” is very tall.</td>
</tr>
</tbody>
</table>
### Availability Interpretation

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>7. lake film</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A “lake film” is on the subject of lakes.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A “lake film” is made beside a lake.</td>
</tr>
<tr>
<td><strong>8. rhinoceros horse</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A “rhinoceros horse” has a big horn.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A “rhinoceros horse” has tough skin.</td>
</tr>
<tr>
<td><strong>9. skunk squirrel</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A “skunk squirrel” has a very unpleasant smell.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A “skunk squirrel” has a stripe on its back.</td>
</tr>
<tr>
<td><strong>10. sugar line</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A &quot;sugar line&quot; is a row of sugar on a canvas.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A &quot;sugar line&quot; is when people have to queue for sugar.</td>
</tr>
<tr>
<td><strong>11. tulip sled</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A “tulip sled” has the same shape as a flower.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A “tulip sled” has an image of a flower on it.</td>
</tr>
<tr>
<td><strong>12. vodka face</strong></td>
<td></td>
</tr>
<tr>
<td><strong>high</strong></td>
<td>A “vodka face” is a drunken look from consuming too much alcohol.</td>
</tr>
<tr>
<td><strong>low</strong></td>
<td>A “vodka face” is the look someone gets when they taste strong alcohol.</td>
</tr>
</tbody>
</table>
APPENDIX C – EXPERIMENT 3 & 4 MATERIALS

The following compounds, interpretations and stories were used as materials in Experiments 3 and 4. In these experiments, the meanings for the compound given in the story and in the paraphrase judgement create the following conditions:

- **Matched Context:** high-availability story and high-availability paraphrase
  or low-availability story and low-availability paraphrase
- **Omitted Context:** unbiased story and high-availability paraphrase
  or unbiased story and low-availability paraphrase
- **Mismatched Context:** high-availability story and low-availability paraphrase
  or low-availability story and high-availability paraphrase
1. Arrow Window

**Story:**

Jimmy was visiting the new house of his friend Mike. Mike was giving John a full tour of the house. There were a lot of interesting features in the house.

<table>
<thead>
<tr>
<th>High</th>
<th>The windows were very narrow and in the shape of arrows.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>Every room was beautifully designed in a particular style.</td>
</tr>
<tr>
<td>Low</td>
<td>The windows had a strange arrow design on the glass.</td>
</tr>
</tbody>
</table>

Jimmy was very impressed with what he saw. Jimmy wondered how much Mike paid for the arrow windows. After the tour they went for a cup of tea.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>An “arrow window” looks just like an arrow.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>An “arrow window” has an arrow pattern on it.</td>
</tr>
</tbody>
</table>

2. Bee Hat

**Story:**

Diana laughed at the old hat she found while cleaning the attic. No one had been up there for years. Chuck looked at the hat and said, “That's Aunt June's old hat.”

<table>
<thead>
<tr>
<th>High</th>
<th>“She made the hat so to look like a bee for the fancy dress party.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>“Years ago she used to wear that hat when she was going out.”</td>
</tr>
<tr>
<td>Low</td>
<td>“She put a picture of a bee on it for a fancy dress party.”</td>
</tr>
</tbody>
</table>

Diana thought about how near Halloween was. Diana said, “I suppose I could use a bee hat for my costume.” Chuck said, “Aunt June always had interesting outfits.”

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>A “bee hat” has a similar shape to a bee.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A “bee hat” has a drawing of a bee on it.</td>
</tr>
</tbody>
</table>
3. Carrot Bomb

**Story:**
Bill and Jim worked in the military research and development centre.
They had worked together for many years.
They designed new types of bombs for the army.

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Unbiased</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill</td>
<td>Bill was making mines that were shaped like common foods such as carrots.</td>
<td>Bill often stayed in late to work hard on making mines.</td>
<td>Bill was making bombs filled with a mixture of chemicals and carrots.</td>
</tr>
<tr>
<td>Jim</td>
<td>Jim was impressed with Bill’s novel design.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>He wished he had thought of building a carrot bomb.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jim went back to working on his own project.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A “carrot bomb” is an explosive that looks like a vegetable.</td>
<td>A “carrot bomb” is an explosive that is made from vegetables.</td>
</tr>
</tbody>
</table>

4. Father Card

**Story:**
Amelia and Mark were browsing in the new card shop.
The shop had every type of card conceivable.
The cards were organised into a variety of categories.

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Unbiased</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amelia</td>
<td>Amelia said, “I'm looking to buy a card for my father.”</td>
<td>Amelia said, “We could spend all day going around this shop.”</td>
<td>Mark said, “These big cards are like the fathers of the smaller ones”.</td>
</tr>
<tr>
<td></td>
<td>Mark decided to search again from the other end of the shop.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amelia said, “So let me know if you find a good father card.”</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mark and Amelia had looked at hundreds of cards.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A “father card” is for a father.</td>
<td>A “father card” is a large card.</td>
</tr>
</tbody>
</table>
5. Foot Chip

**Story:**
John had always been a keen basketball player.
He played at least once a week.
With training he was improving all the time.
However, because he was small, John got injured a lot.

<table>
<thead>
<tr>
<th>High</th>
<th>Once a bone in John’s foot was chipped during a bad tackle.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>John hoped that with enough practice he would be as good as everyone else.</td>
</tr>
<tr>
<td>Low</td>
<td>Once John slipped a chip of wood under each foot to make himself taller.</td>
</tr>
</tbody>
</table>

Most of John’s friends were on the team.
His friends thought that the foot chip would be very painful.
John still loved every game he played.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>A “foot chip” is an injury to your foot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A “foot chip” is put in your footwear.</td>
</tr>
</tbody>
</table>

6. Giraffe Antelope

**Story:**
Kevin and Elaine were on safari in Africa.
The sun was high in the sky and it was very hot.
Yesterday they were lucky enough to see a pride of lions.

<table>
<thead>
<tr>
<th>High</th>
<th>Today they saw an antelope that had a neck as big as a giraffe’s.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>Today they hoped to see a variety of interesting animals.</td>
</tr>
<tr>
<td>Low</td>
<td>Today they saw an antelope that looked as tall as a giraffe.</td>
</tr>
</tbody>
</table>

Kevin got some good photographs of everything.
The guide told them it was rare to see a giraffe antelope.
Kevin and Elaine had a great holiday.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>A “giraffe antelope” has a long neck.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A “giraffe antelope” is very tall.</td>
</tr>
</tbody>
</table>
7. Lake Film

**Story:**
George was trying to develop expertise in filming.
He had already invested heavily in camera equipment.
He went into a video equipment store to purchase some items.
The clerk was interested in George’s latest film project.

<table>
<thead>
<tr>
<th>Level</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>George said, &quot;I'm making a film about lakes in the area.&quot;</td>
</tr>
<tr>
<td><strong>Unbiased</strong></td>
<td>George explained that this was why he needed more equipment.</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>George said, &quot;I'm taking a boat out on the lake to shoot my next film.&quot;</td>
</tr>
</tbody>
</table>

George described his ideas to the clerk.
The clerk agreed that the lake film was a great idea.
George hoped to become famous with his films.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>Level</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>A “lake film” is made on the subject of lakes.</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>A “lake film” is made on the surface of a lake.</td>
</tr>
</tbody>
</table>

8. Rhinoceros Horse

**Story:**
John and James were out on the farm.
James had shown John most of the farm animals.
They went to the field where new types of horses were being bred.

<table>
<thead>
<tr>
<th>Level</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>John noticed one horse had a large horn like a rhinoceros.</td>
</tr>
<tr>
<td><strong>Unbiased</strong></td>
<td>They both enjoyed getting out in the fresh air for a while.</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>John thought one horse had very thick skin like a rhinoceros.</td>
</tr>
</tbody>
</table>

James was pleased that John liked the animals on the farm.
John asked “where did you get that rhinoceros horse?”
James gave John the full story.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>Level</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
<td>A “rhinoceros horse” has a big horn.</td>
</tr>
<tr>
<td><strong>Low</strong></td>
<td>A “rhinoceros horse” has very tough skin.</td>
</tr>
</tbody>
</table>
9. Skunk Squirrel

**Story:**

Bill and Jim worked in the military research and development centre. They had worked together for many years. They designed new types of bombs for the army.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Suddenly a squirrel ran by that smelled very bad like a skunk.</td>
</tr>
<tr>
<td>Unbiased</td>
<td>They both heard a rustling noise coming out of the forest.</td>
</tr>
<tr>
<td>Low</td>
<td>Suddenly a squirrel ran by that had a black and white stripe like a skunk.</td>
</tr>
</tbody>
</table>

Peter and Joe moved out of the way very quickly. Joe said, “I’ve never seen a skunk squirrel like that before”. They continued walking for another few hours.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>A “skunk squirrel” has a very unpleasant smell.</td>
</tr>
<tr>
<td>Low</td>
<td>A “skunk squirrel” has a stripe on its back.</td>
</tr>
</tbody>
</table>

10. Sugar Line

**Story:**

David was giving an art presentation in school. He had worked very hard on the project. David was talking about the different types of paintings that artists create.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>One artist imitated snow by sticking sugar in a line onto the canvas.</td>
</tr>
<tr>
<td>Unbiased</td>
<td>He showed a painting where the artist had used a lot of perspective.</td>
</tr>
<tr>
<td>Low</td>
<td>One picture was of people standing in line for sugar during The Depression.</td>
</tr>
</tbody>
</table>

The teacher thought that this particular painting was very interesting. The students were impressed by the picture of the sugar line. David got a good grade for the project.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>A &quot;sugar line&quot; is a thin row of sugar on a canvas.</td>
</tr>
<tr>
<td>Low</td>
<td>A &quot;sugar line&quot; is a queue of people waiting to buy sugar.</td>
</tr>
</tbody>
</table>
11. Tulip Sled

**Story:**
Oscar brought his granddaughter, Betsy, to the new toy store.
He told her that she could have any toy she liked.
Oscar watched as Betsy roamed the store.
Betsy led Oscar over to a display of sleds.

<table>
<thead>
<tr>
<th>High</th>
<th>The first sled Oscar saw looked like a large tulip.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>Oscar was looking at one of the sleds that was hanging from the roof.</td>
</tr>
<tr>
<td>Low</td>
<td>The first sled Oscar saw had a picture of a tulip on it.</td>
</tr>
</tbody>
</table>

Betsy was looking in the same direction as Oscar.
Betsy said, "Could I please have a tulip sled?"
Oscar said, "By all means," and reached for his wallet.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>A “tulip sled” has the same shape as a flower.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A “tulip sled” has an image of a flower on it.</td>
</tr>
</tbody>
</table>

12. Vodka Face

**Story:**
Ronald and his friends were on holidays in Russia.
Earlier in the evening they had gone to a restaurant.
There they had a nice meal before going to a pub.
They tried lots of Russian drinks they had never tasted before.

<table>
<thead>
<tr>
<th>High</th>
<th>Some people were looking very drunk from having too much vodka.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased</td>
<td>Some of the group began to sing and dance in the pub.</td>
</tr>
<tr>
<td>Low</td>
<td>Some people made a strange face when they drank vodka.</td>
</tr>
</tbody>
</table>

Ronald was having a good time with his friends.
Ronald knew that he had a vodka face too.
A good time was had by all.

**Paraphrase Judgement:**

<table>
<thead>
<tr>
<th>High</th>
<th>A “vodka face” is a drunken expression from consuming too much alcohol.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A “vodka face” is the expression someone gets from tasting strong alcohol.</td>
</tr>
</tbody>
</table>
APPENDIX D – EXPERIMENT 5, 6 & 7 MATERIALS

The following high- and low-plausibility descriptions were used as materials in Experiments 5-7.

<table>
<thead>
<tr>
<th>Description</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A jar made of glass that has chutney in it</td>
<td>High</td>
</tr>
<tr>
<td>2. A jar made of glass that has flies in it</td>
<td>Low</td>
</tr>
<tr>
<td>3. An accident caused by a flood from a river</td>
<td>High</td>
</tr>
<tr>
<td>4. An accident caused by a bubble from a river</td>
<td>Low</td>
</tr>
<tr>
<td>5. A monkey that eats honey made by bees</td>
<td>High</td>
</tr>
<tr>
<td>6. A monkey that eats wax made by bees</td>
<td>Low</td>
</tr>
<tr>
<td>7. A uniform worn by a guard from prison</td>
<td>High</td>
</tr>
<tr>
<td>8. A uniform worn by a guard from a school</td>
<td>Low</td>
</tr>
<tr>
<td>9. A game played by children on the street</td>
<td>High</td>
</tr>
<tr>
<td>10. A game played by children on the roof</td>
<td>Low</td>
</tr>
<tr>
<td>11. A page from a book made from reed</td>
<td>High</td>
</tr>
<tr>
<td>12. A page from a book made from plastic</td>
<td>Low</td>
</tr>
<tr>
<td>13. A house near the sea used for holidays</td>
<td>High</td>
</tr>
<tr>
<td>14. A house near the glacier used for holidays</td>
<td>Low</td>
</tr>
<tr>
<td>15. A shoe with tassels used for dance</td>
<td>High</td>
</tr>
<tr>
<td>16. A shoe with tassels used for polo</td>
<td>Low</td>
</tr>
<tr>
<td>17. A cup made from clay used for plants</td>
<td>High</td>
</tr>
<tr>
<td>Description</td>
<td>Plausibility</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>18. A cup made from clay used for soup</td>
<td>Low</td>
</tr>
<tr>
<td>19. A pet owned by sailors found on an island</td>
<td>High</td>
</tr>
<tr>
<td>20. A pet owned by sailors found on a ship</td>
<td>Low</td>
</tr>
<tr>
<td>21. A box that contains files belonging to lawyers</td>
<td>High</td>
</tr>
<tr>
<td>22. A box contains sandwiches belonging to lawyers</td>
<td>Low</td>
</tr>
<tr>
<td>23. A kangaroo that hops over bushes during summer</td>
<td>High</td>
</tr>
<tr>
<td>24. A kangaroo that hops over cars during summer</td>
<td>Low</td>
</tr>
<tr>
<td>25. A bowl made of wood for storing fruit</td>
<td>High</td>
</tr>
<tr>
<td>26. A bowl made of concrete for storing fruit</td>
<td>Low</td>
</tr>
<tr>
<td>27. A store that sells fruit in a mall</td>
<td>High</td>
</tr>
<tr>
<td>28. A store that sells guns in a mall</td>
<td>Low</td>
</tr>
<tr>
<td>29. A shop that sells chickens fed on cornmeal</td>
<td>High</td>
</tr>
<tr>
<td>30. A shop that sells parrots fed on cornmeal</td>
<td>Low</td>
</tr>
<tr>
<td>31. A bed in a surgery used by patients</td>
<td>High</td>
</tr>
<tr>
<td>32. A bed in a surgery used by visitors</td>
<td>Low</td>
</tr>
<tr>
<td>33. A bag made from cotton for storing laundry</td>
<td>High</td>
</tr>
<tr>
<td>34. A bag made from steel for storing laundry</td>
<td>Low</td>
</tr>
<tr>
<td>35. A wall made of stone with glass on top of it</td>
<td>High</td>
</tr>
<tr>
<td>36. A wall made of bamboo with glass on top of it</td>
<td>Low</td>
</tr>
<tr>
<td>37. A bin that holds chemicals made from glass</td>
<td>High</td>
</tr>
<tr>
<td>38. A bin that holds chemicals made of plastic</td>
<td>Low</td>
</tr>
<tr>
<td>39. A wax that melts in fire for sealing letters</td>
<td>High</td>
</tr>
<tr>
<td>40. A wax that melts in water for sealing letters</td>
<td>Low</td>
</tr>
<tr>
<td>41. A gun that shoots arrows with feathers attached</td>
<td>High</td>
</tr>
<tr>
<td>42. A gun that shoots arrows with leaves attached</td>
<td>Low</td>
</tr>
<tr>
<td>Description</td>
<td>Plausibility</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>43. A spider that kills birds with its fangs</td>
<td>High</td>
</tr>
<tr>
<td>44. A spider that kills birds with its web</td>
<td>Low</td>
</tr>
<tr>
<td>45. A bird that pecks apples for their pips</td>
<td>High</td>
</tr>
<tr>
<td>46. A bird that pecks apples for their skin</td>
<td>Low</td>
</tr>
<tr>
<td>47. An applause for a play in a theatre</td>
<td>High</td>
</tr>
<tr>
<td>48. An applause for a poem in a theatre</td>
<td>Low</td>
</tr>
<tr>
<td>49. A chair where people get their shoes polished</td>
<td>High</td>
</tr>
<tr>
<td>50. A chair where people get their buttons polished</td>
<td>Low</td>
</tr>
<tr>
<td>51. A fossil of an insect that has a sting</td>
<td>High</td>
</tr>
<tr>
<td>52. A fossil of an insect that has a tail</td>
<td>Low</td>
</tr>
<tr>
<td>53. An owl that eats mice for their meat</td>
<td>High</td>
</tr>
<tr>
<td>54. An owl that eats cats for their meat</td>
<td>Low</td>
</tr>
<tr>
<td>55. A hamster that from his mouth emits a squeak</td>
<td>High</td>
</tr>
<tr>
<td>56. A hamster that from his mouth emits a whistle</td>
<td>Low</td>
</tr>
<tr>
<td>57. A teacher that teaches spelling to children</td>
<td>High</td>
</tr>
<tr>
<td>58. A teacher that teaches spelling to adults</td>
<td>Low</td>
</tr>
<tr>
<td>59. A dog that catches sticks in the park</td>
<td>High</td>
</tr>
<tr>
<td>60. A dog that catches leaves in the park</td>
<td>Low</td>
</tr>
<tr>
<td>61. A disease caused by flies in a swamp</td>
<td>High</td>
</tr>
<tr>
<td>62. A disease caused by flies in a school</td>
<td>Low</td>
</tr>
<tr>
<td>63. A lawn shaped like a heart used for weddings</td>
<td>High</td>
</tr>
<tr>
<td>64. A lawn shaped like a cake used for weddings</td>
<td>Low</td>
</tr>
<tr>
<td>65. A note left for a milkman on a doorstep</td>
<td>High</td>
</tr>
<tr>
<td>66. A note left for a milkman on a pole</td>
<td>Low</td>
</tr>
<tr>
<td>67. A pencil carved from wood used for art</td>
<td>High</td>
</tr>
<tr>
<td>Description</td>
<td>Plausibility</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>68. A pencil carved from stone used for art</td>
<td>Low</td>
</tr>
<tr>
<td>69. A song played at a concert requested by fans</td>
<td>High</td>
</tr>
<tr>
<td>70. A song played at a concert requested by bouncers</td>
<td>Low</td>
</tr>
<tr>
<td>71. A tank used during a war covered in camouflage</td>
<td>High</td>
</tr>
<tr>
<td>72. A tank used during a parade covered in camouflage</td>
<td>Low</td>
</tr>
<tr>
<td>73. A ruler with an edge for opening letters</td>
<td>High</td>
</tr>
<tr>
<td>74. A ruler with a scissors for opening letters</td>
<td>Low</td>
</tr>
<tr>
<td>75. A disease spread by rats that afflicts binmen</td>
<td>High</td>
</tr>
<tr>
<td>76. A disease spread by fish that afflicts binmen</td>
<td>Low</td>
</tr>
<tr>
<td>77. A hamster that lives in a cage that feeds on nuts</td>
<td>High</td>
</tr>
<tr>
<td>78. A hamster that lives in a desert that feeds on nuts</td>
<td>Low</td>
</tr>
<tr>
<td>79. A guitar that has strings and is played in a street</td>
<td>High</td>
</tr>
<tr>
<td>80. A guitar that has bells and is played in the street</td>
<td>Low</td>
</tr>
<tr>
<td>81. A bruise on a finger caused by a pencil</td>
<td>High</td>
</tr>
<tr>
<td>82. A bruise on a toe caused by a pencil</td>
<td>Low</td>
</tr>
<tr>
<td>83. A cut to the back caused by a seat</td>
<td>High</td>
</tr>
<tr>
<td>84. A cut to the nose caused by a seat</td>
<td>Low</td>
</tr>
<tr>
<td>85. A scratch caused by a mosquito biting an arm</td>
<td>High</td>
</tr>
<tr>
<td>86. A scratch caused by a ladybird biting an arm</td>
<td>Low</td>
</tr>
<tr>
<td>87. A lobster that catches shrimp with its claw</td>
<td>High</td>
</tr>
<tr>
<td>88. A lobster that catches shrimp with its web</td>
<td>Low</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


