Acquiring Phrasal Lexicons from Corpora

Colin Bannard

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2006

Abstract

This thesis deals with the acquisition from corpora of information about the phrasal lexicon. While varying greatly in their treatment of such items, all schools of linguistic thought acknowledge that the lexicon contains not only single words, but a considerable number of multiword sequences (multiword expressions or MWEs). As for individual words, an adequate description of MWEs must be based on the systematic study of real usage, which can best be achieved with the use of corpus data. However, lexicographers cannot manually inspect every combination of words in the corpus with a view to deciding whether that phrase should be included in the lexicon. There is therefore considerable need for well-motivated ways of automatically identifying units of interest.

We begin this thesis by looking at the range of arguments that have been made for the inclusion of different kinds of multiword expression in the lexicon. We identify arguments for inclusion made on the separate grounds of frequency, syntax and meaning. Subsequent chapters look at acquiring information about these three different dimensions, and demonstrate that different varieties of expression require different techniques.

In chapter three, we look at the relative frequency of multiword sequences. We consider arguments that MWEs are too inconsistent in occurrence for their frequency profile to be adequately described using existing corpora. We provide experimental evidence to the contrary, showing that counts for sequences of up to at least length seven are as stable as for individual words of equivalent frequency. We further show that these counts are informative, by providing evidence that they are reflected in speakers' knowledge of the language.

In chapter four, we look at the problem of syntactically fixed units. The linguistic literature contains considerable evidence for the existence of phrases that do not allow one or more syntactic variation that we would expect given their phrase type and consequently require phrasal lexical entries. We present a method for measuring the syntactic flexibility of instance of one phrase type and use this to rank items according to their fixedness. We provide evidence that this technique highlights a set of phrases that are not only valid MWEs, but crucially could not be identified purely on the grounds of relative frequency of occurrence.

In chapter five, we look at the phenomena of non-compositionality. Languages contain a great many units whose meaning cannot be derived analytically from the meanings of their component words. We look here at the example of the verb-particle

construction. Building on previous work, we use lexical context to model the meaning of words and phrases and show that by quantifying the similarity between the contexts of phrases and the contexts of their component words across the rest of the corpus, we can usefully identify those verb-particle combinations whose meaning cannot be analysed compositionally.

Acknowledgements

Buccleuch Place has been a great place in which to conduct research and to think over the last four years. Many many thanks to my supervisors Alex Lascarides and Mirella Lapata for their encouragement, insight and detailed comments. Thanks to my peers, and in particular Markus Becker, Chris Callison-Burch, Josh Schroeder, Matt Smillie and Andrew Smith for being ever encouraging and interested. I have also benefitted greatly from conversations with Frank Keller, Miles Osborne and Scott McDonald.

My debts are by no means restricted to Edinburgh. It was as a visitor to the LinGO lab at Stanford University during my masters degree that I first became interested in multiword expressions - thanks are due to John Beavers, Ann Copestake, Dan Flickinger, Ivan Sag, Tom Wasow and particularly to Tim Baldwin from whom I learned so very much about doing corpus research. The work described in chapter 4 was started while I was a visitor to NTT Communication Science Laboratory in Kyoto. Thanks in particular to Francis Bond and Takaaki Tanaka. The research described in chapter 3 was carried out while I was visiting the Psychology department at Stanford University, and I would like to thank Michael Ramscar and his lab, particularly Nicole Gitcho and Dan Yarlett, for being so generous with their time, space and ideas.

Thanks to Casey Butterfield for being so supportive and so uniquely able to make me forget about all the work I had to do. And thanks finally to my parents for providing the best possible home for me to write in over the last few months.

This research was made possible by the financial support of the Economic and Social research council and grant PTA-030-2002-01740.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

[redacted]

(Colin Bannard)

Table of Contents

Chapter 1

Introduction

This thesis concerns the acquisition from corpora of information about multiword expressions in English. As with individual words, an adequate description of such items must be based on real usage. This can be accomplished using corpora. However, in order for corpus analysis to be possible it is necessary to first identify the expressions that need to be described. This thesis provides evidence that statistical corpus processing can be useful in identifying exactly which sequences need to be accounted for. This chapter introduces the problem and provides an overview of the work that is to follow.

1.1 Introducing the problem

The lexicon is a crucial part of any language description. By this we mean at least a list of the conventional linguistic forms that can be combined to form utterances (we remain agnostic on exactly how this may or may not relate to or overlap with other levels of linguistic description). Deciding what these forms are and consequently what to put in the lexicon is an issue that must be confronted by anyone involved in the creation of such a description, from the writer of pedagogical materials through the theoretical linguist to the builder of computer language processing systems. The easiest answer is that the lexicon contains words, where words are sequences of characters surrounded by space. However, even leaving aside the question of segmentation (in spoken language the "spaces" are far from obvious, and many writing systems simply do not use space in this way) for anyone looking at real language usage it soon becomes apparent that such a solution will not do.

One significant problem with this overly simple definition of lexical items concerns sub-lexical morphology. Many word-forms are best treated not in isolation but as a set

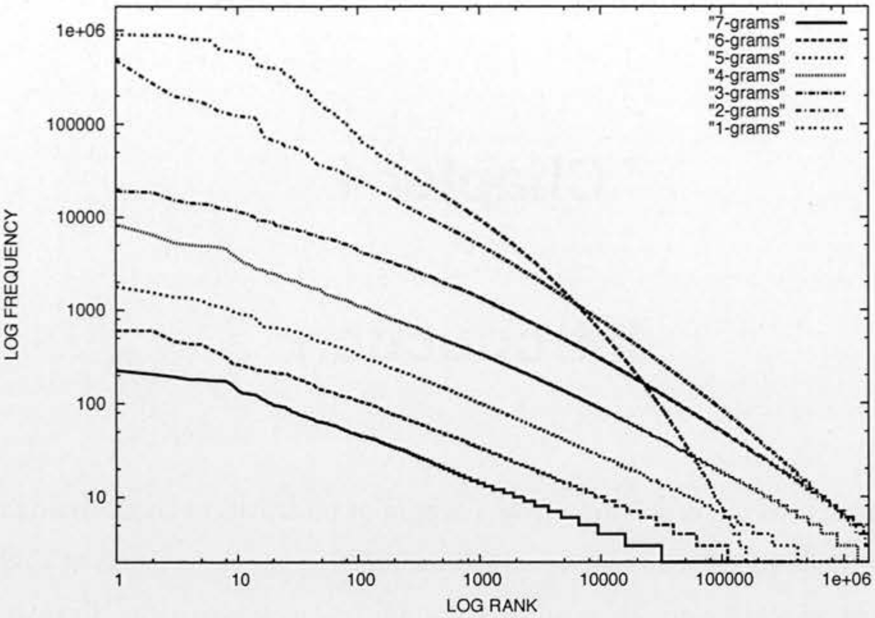


Figure 1.1: Zipf curve for all words and n-grams of up to length 7

of related forms which can, for descriptive purposes at least, be grouped under a single underlying "lemma". By treating each word form as a unique lexical item one misses out on important linguistic generalisations. Another problem, and that which we are concerned with in this thesis, is the fact that all human languages contain many linguistic units consisting of multiple "words" that must be entered into the dictionary as a single lexical entry. The linguistic literature contains a vast range of terms for referring to such phrases. We will use the term "multiword expression" (we will abbreviate this to MWE where appropriate) to refer to all combinations of two or more words (adjacent or otherwise) that need, for any of the reasons we will go on to discuss, to be included in the lexicon as whole separate entries. As we will see, the arguments for the inclusion of units varies enormously, with syntax, semantics, pragmatics, processing efficiency and style being variously given as justification for their inclusion. Add to this the fact that the most conservative estimates for the number of such units is as equal to the size of the single word lexicon (see Mel'čuk, 1995 or Jackendoff, 1995 for estimates of the size of the multiword lexicon), and it should be apparent that this poses a significant problem for the lexicographer.

The art of lexicography has changed a great deal over the last two decades. Whereas previous generations of lexicographers had to work largely from intuition in discovering and analysing units of language when creating a dictionary, the rapid growth in computer processing and storage capacity has meant that lexicographers have been able to exploit large text corpora in selecting and analysing their entries. The publication in 1987 of the Collins COBUILD dictionary of English (Sinclair *et al.*, 1987; see Sinclair, 1987 for an account of its creation), marked the arrival of a new kind of corpus-driven dictionary. As well as being the first to include the words that are actually in use rather than thought to be in use, this saw the inclusion for the first time of information about a word's place in the greater linguistic environment, such as its relative frequency and the linguistic contexts in which it habitually occurs.

The use of a corpus is of considerable assistance in the task of deciding what to put in the dictionary. Instead of pure intuition, lexicographers can now base their entries on real data.From the point of view of compiling lists of graphologically discrete words this is extremely useful. From the point of view of compiling lists of multiword units it should be extremely useful too. However creating corpus-driven multiword dictionaries is far from straightforward.

This thesis will explore the extraction of information about MWEs from a single widely available text collection: the 89.39 million word written component of the British National Corpus (BNC; Burnard, 2000).[1] This contains approximately 200,000 unique sequences of characters that occur more than once in the corpus and consequently might be candidate words. About 14,000 of these can be discarded as sequences of numbers, and many others can be ruled out as noise, using knowledge of the kind of sequences that exist in English words (e.g. those strings that contain only consonants can be ruled out). For a large scale lexicographic effort, the task of sorting through these word lists and compiling evidence is considerable but manageable. By contrast, the number of repeated strings of more than 1 word in the corpus is over 12 million. Having a staff of lexicographers sort through this list manually is clearly not feasible. And it is important to note that this list is not just made of 2 word combinations. The frequency of longer word sequences can be seen in figure 1.1, which shows the rate of occurrence of all repeated strings of between one and seven words in this corpus, plotted against their rank.

[1] The complete BNC consists of approximately 100 million words, encompassing an additional 10.58 million words of transcribed speech. There are significant difference between spoken and written English (Biber 1988). These often make results reported over the combination of the two kinds of language difficult to interpret. We therefore choose to explore a single variety, and utilise only the larger written section.

The information above concerns just repeated strings of adjacent words. A further problem is that the kinds of fixed expressions one is interested in will not necessarily be contiguous. Rosamund Moon, one of the original COBUILD lexicographers, and the author of the most substantial corpus-based study of multi-word units to date states the problem that they present to corpus lexicographers as follows:

> "There is too much data for manual analysis, and pre-processing is required in order to make use of the information that such corpora contain. That is, routines are run over the data in order to identify certain features of the component text. FEIs [Fixed Expressions and Idioms] present a particular problem for preprocessing routines. Semantically and often syntactically they function as units rather than as arbitrary sequences, but they need not be contiguous or uninterrupted. They may be syntactically or lexically ill-formed, breaking conventional rules or valency patterns." (Moon 1998:52)

Moon despairs of extracting units directly from the corpus, and instead bases her study on a pre-compiled list of phrases of interest, of which she simply identifies occurrences in a corpus by searching.

The problem is not just practical, but is of theoretical importance too. As we will discuss in chapter 3, there are still considerable doubts among researchers as to whether corpora can ever really give us the evidence we need about multiword expressions. As we will see, researchers have argued that the phrases of interest are often too rare or inconsistent to be effectively covered by any corpus.

The challenges in using corpora to learn about the multiword lexicon are clearly great. This thesis will argue, however, that valuable information can be extracted from the BNC about a range of different varieties of MWE by utilising corpus processing techniques from computational linguistics. As we alluded to before, and will argue in detail in chapter 2, one of the major problems with creating multiword lexical resources is that there exist more than one variety of MWE and consequently there are multiple reasons why a lexicographer might want to include phrases in lexica. We will demonstrate that computational linguistic techniques can provide valuable information about three different varieties of MWE, utilising information about frequency, syntax and meaning. We will present evidence that these techniques can be useful to lexicographers in the task of identifying which phrases in a corpus need to be considered for inclusion in lexical resources.

1.2 The contribution of this thesis

The central claim of this thesis is that statistical corpus processing can be useful in identifying examples of a range of MWE varieties that need to be accounted for. Its specific contributions are as follows:

First we will use the linguistic literature in order to outline the different varieties of MWEs that need to be covered by lexica, and to delineate the different problems that computational linguists will need to tackle if we are to provide the much-needed support to lexicographers. We will then look at the acquisition of information about three overlapping but distinct varieties of MWE.

The first kind of corpus-derived information we will look at is frequency data. We will provide evidence that counts for MWEs derived from the BNC are at least as stable as for individual words of equivalent frequency. We will then demonstrate the importance of such frequency information by showing that the rate of occurence of multiword sequences is reflected in their processing by language users.

We will next move on to look at the syntactic dimension of MWEs. There exist in English a great many phrases that are formally fixed, and do not allow the range of syntactic variations we would expect given their phrase type. We will show that by mining corpora for such fixed units units one can automatically discover a range of MWEs that are not available using purely frequency-based approach.

Finally we will consider the semantic dimension of MWEs. English contains many phrases whose meaning cannot be derived from the meaning of their component words. Such phrases must be consequently be included in the lexicon as whole units. We will look at how corpus processing techniques might help to identify such items, and will show that the lexical contexts of phrase can provide us with information about their degree of semantic analysability.

1.3 Overview of the thesis

1.3.1 Overview of chapter 2

This chapter will provide an account of the kinds of MWEs that have been identified in English, and review the case that have been made for their inclusion in the lexicon. It will provide a review of the empirical evidence that exists for the range of phrase varieties.

1.3.2 Overview of chapter 3

Frequency has long been recognised as an important dimension of the single word lexicon. Valuable work has been done in extracting frequencies from corpora, which has been reported in published lexical resources and utilised by psychologists in conducting experiments. However the significance of the frequency of multiword sequences remains largely unknown. Researchers that are interested in the multiword lexicon continue to express doubt as to the relevance of frequency information. And while there has been some excellent psycholinguistic work exploring the effect of frequencies on the processing of two word combinations, there has been no conclusive work on the human processing of longer sequences.

This chapter will report on a set of experiments that show that corpus frequencies for multiword phrases of various lengths can be as stable and reliable as for large proportions of the single word vocabulary. It will then go on to provide evidence that these counts are representative of the linguistic environment in ways that are reflected in the knowledge and performance of language users.

1.3.3 Overview of chapter 4

Work on the identification of MWEs in corpora has to date concentrated almost solely on the identification of statistically significant "collocations". While this is an important dimension of the phrasal lexicon, there exist a great many phrases that are restricted in terms of the morphological and syntactic variations they allow. For example, an idiom such as "kick the bucket" (which occurs a total of five times in the written part of the BNC, and consequently would not appear in the lexicon on the basis of frequency alone) can be morphologically varied ("kicked the bucket"), but does not allow for passivisation ("the bucket was kicked"), or internal modification ("he kicked the big bucket"). An example of an even more rigid phrases is *by and large* which allows no variation at all, we say that something is *by and large* true, but we cannot say something is *by and quite large* true, or *by and largest true* . Such phrases must have special entries in the lexicon. This chapter describes a technique for detecting such inflexibility using a parsed corpus. We show that this can be used to identify expressions that cannot be discovered by purely frequency based extraction methods.

1.3.4 Overview of chapter 5

A phrase is said to be non-compositional when the meaning of the whole cannot be recovered by any operation over the meanings of its component parts alone. This chapter describes how non-compositionality is reflected in the lexical contexts in which such phrases appear. It tests the hypothesis that the lexical contexts in which a given multiword expression occurs in a corpus will be more similar to those of a given component word if that component word is contributing an independent meaning to the phrase. It reports a significant correlation between this distance and expert and non-expert compositionality judgements for verb-particle constructions, and presents evidence that this information can be used to useful rank lists of candidates MWEs as an aid to lexicographers. Part of this work has previously been published in Bannard *et al.* (2003) and Bannard (2005).

Chapter 2

What's in a lexicon?

This chapter provides a survey of the arguments that have been made for including sequences of more than one word in the lexicon (together with the supporting empirical evidence). The main contribution of this thesis is in providing evidence that statistical corpus processing can be useful in identifying exactly which sequences need to be accounted for. The role of this chapter is to delineate the types of phrases about which information is needed, and to thereby provide a justification for the information that we go on to acquire in later chapters.

Before beginning this survey it is important to note a limitation on the scope of this thesis. Much of the literature outlined in this chapter is concerned with the encoding of lexical information, with how this information interfaces with other levels of linguistic description, and with the role that it might play in human language comprehension. This discussion will be extremely useful to us in guiding the kind of multiword sequences we will go on to acquire. It is important to emphasize, however, that the output of the techniques described in this thesis is not going to be fully realised linguistic descriptions. We are concerned simply to identify items that need to be described and not with the details of that description. Furthermore, there are very necessary debates in linguistics and psychology concerning where the line between the lexicon and other levels of linguistic description lies, and indeed whether it even makes sense to talk about distinct components or levels of processing at all (see e.g. Sinclair, 1991, Goldberg, 1995, Tomasello, 2003). We will argue only that the information we are acquiring from corpora needs to be accounted for somehow in any description of a language, employing the term lexicon for descriptive convenience, and do not intend to engage with these debates.

The structure of this chapter will be as follows. We will begin by discussing the two broad arguments for including multiword items in the lexicon. We will then review the varieties of multiword expressions that have been discussed in the linguistic literature, and the case for their inclusion. We will then move on to discuss the crucial empirical evidence that has been provided for the existence of the phrasal lexicon in various fields of the cognitive sciences.

2.1 What is in the lexicon?

Before looking at the different kinds of multiword unit in English, and the case for their inclusion in the lexicon, we want to consider the more abstract discussion of what belongs in the lexicon. One obvious place to look for discussion of what belongs in the lexicon is lexicographic tradition. English Lexicography has long been a largely atheoretical pursuit. However, Béjoint (1989:1) introduces the french notion of "codée", which he translates as **codedness**, and defines as follows:

> "Roughly speaking, a sequence of graphemes or phonemes is coded if it is recognised as an "established" unit of the language by the members of the community...Codedness is an important notion in lexicography, because the word list of a dictionary can only be made up of coded units; if a dictionary recorded uncoded units, it would not be a dictionary at all. What is excluded is fairly obvious: most sentences, random sequences of discourse like *What is,excluded is fairly,obvious most sentences random* etc, and also perhaps *hapax legomena*; but a positive definition is difficult."

He cites definitions of the term with reference to frequency : "A sequence of graphemes or phonemes is coded when it is frequent enough, frequency in a corpus being evidence of the fact that it is indeed used by the language community at large" (p.2). However, he then acknowledges that this is rather problematic, pointing out that the items he defines in the above sequence as random sequences are more frequent than some words (we will consider this in more detail in chapter 3).

Béjoint (1989:2) next cites attempts to define codedness qualitatively, citing both semantic ("[a] sequence is coded if it is stored as a unit, produced as a "ready-made" whole and decoded without any analysis of its constituents....") and syntactic features ("[a] coded sequence is one in which no (or very little) syntactic or lexical variation is allowed; it belongs to the language user's competence, not to performance") that make a sequence coded. Pawley (1986) similarly refers to both qualitative and quantitative justifications for inclusion. However, he attributes the different justifications to different traditions, the qualitative definition to the tradition of structural linguistics:

> "Parsimony of description being highly valued, any one form-meaning pairing should be specified only once in the grammar, whether it be in the lexicon or by the syntactic rules which apply to lexical items....If a given form-meaning pair (expression) cannot be predicted by the productive rules of the grammar, it must be listed in the lexicon; if it can be predicted, it does not belong in the lexicon." (p.99)

And the quantitative to the tradition of lexicographers:

> "...dictionaries may include any composite form if it is a common usage, i.e. if it is recognized by members of the language community as a standard way of referring to a familiar concept or conceptual situation." (p.101)

This notion of codedness will guide the approach to the construction of the lexicon in this thesis. We will also accept this distinction between qualitative and quantitative codedness, and argue that the two distinct varieties need to be approached separately in the construction of lexicons, and crucially in the acquisition of information about the phrasal lexicon from corpora. The terminology we use to make the distinction will be taken from the linguistic literature on word formation. Bauer (1983) outlines three stages in the history of the word. The first stage is **nonce formation**. This is when a new complex term is coined in order to fit a communicative need. In terms of word sequences, this would just be when a sequence is creatively produced and is not coded as a whole at all. The next stage is **institutionalisation**. This is when a complex word comes to be an established term, widely used and understood in the language community. The institutionalisation of a term is most clear when there exist seemingly synonymous ways of realising the meaning of that term that are chosen rarely or not at all. A often quoted example is the word combination *strong tea*. The phrase *powerful tea* would seem to carry much of the same meaning, and yet it is almost never seen. This appears because the term *strong tea* has become the established or institutionalised way of realising the concept. In terms of multiword sequences this is equivalent to Bejoint's notion of a phrase being used by "the community at large" or Pawley's notion of a phrase being "in common usage". As with the equivalent notions from the previous writers, institutionalisation is characterised quantitatively in the main.

The third and final stage in the development of a word for Bauer is **lexicalisation**. This occurs when "because of some change in the language system, the lexeme has, or takes on, a form which it could not have if it had arised by the application of productive rules" (p.48). This change can take place at any level of linguistic analysis. Multiword examples of lexicalised forms are expressions whose meaning cannot be derived from to the meanings of their parts alone or which have a syntactic form which would be ungrammatical if used for a newly formed phrase. An example of the former is the phrase *red herring* which is conventionally used to refer to an intentional distraction. The meaning has nothing to do with fish or the colour red, and could not arise through the application of the productive principles of English. An example of a form which is lexicalised at the syntactic level is *by and large*. There are no other preposition plus adjective conjunctions in English that serve an adverbial function, and therefore we can say that the phrase could not have arisen from the combinatorial patterns of the language. The notion of lexicalisation is exactly analogous to Bejoint's qualitative definition of codedness. Although the terms are diachronically defined for Bauer, he uses them in synchronic analysis of the vocabulary of English, and they seem very useful in discussing the key distinctions in considering the phrasal lexicon.

Lest the notion of "codedness" remain entirely theoretical and abstracted from the task of lexicography, we will give an analogy from an applied field. One application for which an accurate description of language is particularly important is computer natural language processing. NLP systems need to handle real language data and invariably need to be able to respond to a great variety of different kinds of input. In an introduction to lexicography for NLP, Boguraev and Briscoe (1989:4–5) state that "the lexicon provides the information not predictable from the rules, which "feeds" the rules and ensures they function correctly". They give a clear example of what they mean by this. They outline five kinds of knowledge that are "potentially relevant to NLP systems", which they label as phonological, morphological, syntactic, semantic and pragmatic. They state that "[e]ach of the five broad types of knowledge we have introduced can be characterised to a large extent in terms of sets of general rules". They give the example of a rule that adjectives can appear after the verb *to be* in English. However they then point out that the "rule breaks down in the case of an adjective such as man-eating", and state that "[d]ealing with such exceptions is the province of the lexicon...".

The key here is the suggestion that the lexicon must contain information that is "not predictable" from the other components of the language description. The lexicalised phrases that we mentioned, the semantically opaque phrase *red herring* and the syntactically fixed expression *by and large* are not predictable from the semantic and the syntactic components of the language description respectively, and would consequently need to have phrasal entry with unique semantic and syntactic information. For the purely institutionalised expression *strong tea*, the syntactic and semantic aspects of the phrase are predictable. However the form in which it is realised is not. Grammar engineering (and generative approaches to language in general) has tended to ignore such questions of lexical form. However if the system is to account for and generate natural-sounding speech there needs to be some knowledge of this, and this would require the addition of a special phrasal entry with unique phonological information.

Distinguishing between institutionalisation and lexicalisation is not always straightforward. As one might expect given the diachronic process of language change that underlies the distinction, it is to some extent a matter of degree. They are, nonetheless, very useful distinctions, which any approach to lexical acquisition must take into consideration. As we will explain later in the thesis, almost all work to date on acquiring multiword sequences from corpora has concentrated exclusively on institutionalised word sequences.

2.2 The dimensions of the phrasal lexicon

This section will detail the varieties of multiword expression that have been detailed in the linguistic literature, and the evidence that has been presented for their existence. We will relate the varieties back to the defining features that we discussed in the last section and to the larger aims of the thesis.

The linguistic literature contains a great number of different attempts to categorize and label the varieties of multiword expressions. and a truly remarkable range of terminology. Despite this diversity, however, there is considerable consistency in the basic distinctions being made. This chapter will discuss the three main dimensions along which most writers agree that MWEs vary. We will refer to these dimensions as **collocation**, **syntax** and **semantics**.

2.2.1 Collocations

According to the most reductionist generative conceptions of language, linguistic competence can be accounted for in terms of two distinct components - a grammar and a finite lexicon. The lexicon lists the syntactic category of each word, and the grammar uses this categorial information to combine the words to form a sentence. This view is outlined by Stephen Pinker as follows:

> “The way language works, then, is that each person’s brain contains a

> lexicon of words and the concepts they stand for (a mental dictionary) and a set of rules that combine the words to convey relationships among concepts (a mental grammar)...When people learn a language, they are learning how to put words in order, but not by recording which word follows which other word. They do it by recording which word *category* – noun, verbs, and so on – follows which other category." (Pinker 1995:85,93-94)

The basic underlying principles of word combination are assumed to be down to categorial information and have nothing to do with the contingencies of the particular word form chosen. Such contingencies are not considered to be a part of linguistic competence, but are instead part of the less interesting category of performance, which is not regarded as of interest to the linguist.

These basic assumptions dominated linguistic departments, particularly in the US, from the 1950s on. There were, however, dissenting voices all along. One tradition of dissent was the line of thought in British linguistics that began with J.R.Firth. In a paper first published in 1951, Firth introduced to serious linguistic enquiry the term "collocation" (Firth, 1957 ; the term was of course not new, and indeed the Oxford English Dictionary details a use of the word that seems to correspond to Firth's sense as early as 1873, but the popularity of the term can be attributed to Firth). Collocation refers to the fact that words are not chosen purely according to their categorial status, but that particular words characteristically occur with other particular words. In a description of this tradition, Mitchell (1971:47–48) emphasizes the important "interdependence of grammar and lexicon" stating the following:

> "[T]he division between morphology and syntax is in fact a great deal less clear-cut than is often assumed and may even be otiose. Many of the roots and affixes, inflections and derivations of morphology have their implications as to choices made elsewhere in word + domains, and vice versa; good (with zero suffix) is by no means the singular of *goods* and will not therefore appear in such associations as *consumer* —- or —- *and chattels*, while *goodness* not only does not occur indiscriminately with any kind of following verb (cf. the impossibility of **goodness hates him*) but also excludes pronominal forms other than those of the first person singular from exclamations like (*my*) —— and —— *gracious* (*me*)."

In other words, in language production particular lexical choices have an effect on other lexical choices elsewhere in the discourse in a way that has nothing to do with their categories.

How is this relevant to MWEs? Well the idea is that lexical selection sometimes occurs not at the level of the word but at the level of groups of lexical items. In the traditional view, the stored units of language are conventional lemmas for which a generative system of morphology can produce the correct form for the context. What Mitchell is suggesting is that language data should instead be thought of as stored at the level of the phrase or the sentence, with what we usually think of as syntax being similar to morphology in the way in which it generates the correct form of the conventional unit for the context. A competent speaker stores many such conventional units from the word to the sentence to the whole discourse. As the major proponent of such a view in the US, Dwight Bolinger, put it "our language does not expect us to build everything starting with lumber, nails and blueprint, but provides us with an incredibly large number of prefabs, which have the magical property of persisting even when we knock some of them apart and put them together in unpredictable ways." (Bolinger 1976:1).

This view of language has received a great boost from the rise in the use in corpora that was made possible by the increasing availability of computing power. Prior to the existence of corpora, linguistic analysis was usually based upon intuition and isolated examples. When trying to account for a small set of examples in this way, linguists clearly required maximally abstract principles of word combination, as clearly the principles of combination that are driven by such relatively rare events as individual words is not going to be apparent. Corpora, by contrast, allow the analysis of a very large collection of sentences, meaning that specific lexical patterns become apparent, making them unavoidable data that linguistic theory needs to account for. The linguist can no longer ignore the fact that as Pawley (1986:110) says "[m]any forms are called by the grammar but few are chosen. A grammar may allow a familiar idea to be expressed in many ways, but of the various paraphrases we often find that one (or perhaps two) are used 99 percent of the time and the others rarely if ever."

So what we have encountered here are claims that speakers of a language employ not just individual words, but longer sequences and patterns of words. As we will go on to discuss, these patterns may have additional features that make them coded in some way[1]. However many do not, and even without such features they are important.To use the terminology introduced earlier these are **institutionalised** sequences. Such units are essential in accounting for production, because without knowledge of the preferred lexical realization, a speaker could not produce natural sounding English. As we will see they can also be important in comprehension. Any description of a language must

[1] Some lexicalised forms are common phrases, and so we might say that they are both institutionalised and lexicalised. However not all are and we prefer throughout this thesis to use the term institutionalised to refer only to expressions that are purely institutionalised and not lexicalised.

contain knowledge about these if it is going to account for real language use.

2.2.2 Syntactic fixedness

In the last section we outlined claims that English contains many phrases and formulae that are frequently repeated by different speakers. While we need to account for their frequency of occurrence, and seemingly need to posit that speakers have some stored knowledge of the combination, it is often the case that the meaning and syntax of such sequences can be accounted for in terms of the categorically defined principles of word combination of the type described in the quote from Pinker (1995) seen above. For this reason, generative grammarians have tended to ignore them, assuming that they have no place in the lexicon and can be explained entirely by the generative capacity of the grammar. While regularities may exist, they are assumed to have nothing to do with language *per se*, and exiled to the domain of performance. Ray Jackendoff (1995:136), a linguist who works very much within the generative tradition but recognises its limitation in this respect, characterises such a view as follows:

> "There are a vast number of such memorized fixed expressions...They are hardly a marginal component in the use of language. How is all this material stored? The received wisdom gives us an immediate reaction: "I don't know how it is stored, but it certainly isn't part of the lexicon. It belongs to some other more general purpose part of memory, along with pragmatics, facts of history, maybe how to cook."

Jackendoff goes on to criticise such a view, pointing out that "when it is integrated into the speech stream, we have no sense that suddenly a different activity is taking place".

While set expressions of this kind can be dismissed as a class by generative linguists, there are varieties of recurring expressions that cannot. These are expressions which have become lexicalised. That is they have assumed a form which "could not have arisen by the application of productive rules" (Bauer 1983:48). According to Bauer these can occur at any level of linguistic analysis. In this section we will discuss those items that display syntactic behaviour that cannot be accounted for by general grammatical principles.

The literature contains reference to two such kinds of phrases. The first of these is that variety labelled by Fillmore *et al.* (1988:505) as **extragrammatical idioms**, and described as constructions that the grammar cannot account for. They offer the following list:

(1) *first off, sight unseen, all of a sudden, by and large, so far so good*

In addition to the one overlap *by and large* Nunberg *et al.* (1994:515) offer the following list:

(2) *No can do, trip the light fantastic, kingdom come, battle royal, handsome is as handsome does, would that it were, every which way, easy does it, be that as it may, believe you me, in short, happy go lucky, make believe, do away with, make certain*

While all of these phrases would be familiar to and interpretable by most native speakers of English, they are completely idiosyncratic in that they cannot be accounted for using a grammar rule that is generalisable to other phrases. A number of these can be seen in the sentences from the BNC included below. Take for example the phrases *believe you me*, which we find in sentence 3 [2]. This is an imperative sentence of the form Verb Object Subject (VOS). This particular sentence is interpretable and grammatical, but there is to my knowledge no other sentence of English with the form VOS that would be judged grammatical, and certainly none that would be both grammatical and interpretable. Another sentence that is acceptable itself but cannot be accounted for by any generalisable grammatical rule can be seem in sentence 4. This has the form Negation-Auxiliary-Verb. There is no other acceptable sentence of English with this form. In example sentence 5, we have the phrase *by and large*. This functions as an adverbial. And yet it has the syntactic form Preposition-Conjunction-Adjective. There is no other sequence of words of that form that could serve as an acceptable adverbial in English.

(3) "Believe you me, anything can happen", said Mr Kronweiser, with gloomy relish.

(4) After a brief pause for thought, Stuart Baxter said, "No can do, Vic".

(5) They are by and large perfectly sound.

These kinds of extragrammatical idioms are very interesting and seem to be irrefutable evidence for the existence of MWEs in English. However, they are a fairly restricted set. The second kind of phrase with idiosyncratic syntax is far more common: these are word sequences that have a particular canonical form that can be accounted for using generalisable grammar rules, but that is restricted in terms of the kind of

[2] Unless otherwise noted, all linguistic examples in this thesis are taken from the British National Corpus (BNC).

syntactic variations on that form that it will allow. Pawley (1986:109) describes this variety of MWE as follows:

> "It is characteristic of a large class of phraseological units that, while they are syntactically well formed, they are not freely variable according to the phrase structure rules. That is, particular grammatical or lexical constituents cannot be substituted or expanded on without changing the status of the phrase from lexical item to free expression."

Weinreich (1969) discusses the case of idiomatic adjective-noun combinations such as *white lie*, *hot potato* and *blind date*. For a productive adjective-noun combination such as *white belly* it is possible to generate the predicative form as seen in sentence 6, and the nominative form as in 7 with adjective retaining the same sense. However, for the three listed idiomatic examples it is not possible to do this.

(6) The belly is white

(7) The whiteness of the belly

(8) *The lie is white

(9) *The whiteness of the lie

(10) *The potato is hot

(11) *The hotness of the potato

(12) *The date is blind

(13) *The blindness of the date

These are, then, examples of syntactically fixed MWEs. There are many such phrases across the syntactic spectrum.

One variety of construction whose potential to become inflexible is widely discussed is the transitive verb phrase (we will discuss this in much greater detail in chapter 4). Take for example the phrase *speak volumes* as seen in sentence 14. This, like many verb phrase idioms, does not undergo passivisation as seen in sentence 15.

(14) Whether it's a boilersuit or a power suit, what you wear speaks volumes about you.

(15) *Whether it's a boilersuit or a power suit, volumes are spoken about you by what you wear.

Meanwhile other verb phrases have restrictions on the kind of modification of their elements they will allow. For example the phrase *give birth* as seen in sentence 16 is not seen with the kind of adjectival modification seen in sentence 17.

(16) In the spring of 1977 I did indeed give birth to a boy who later grew that shock of fair hair.

(17) *In the spring of 1977 I did indeed give c-section birth to a boy who later grew that shock of fair hair.

It is important to note that syntactic restrictions are often a matter of degree, with some being very strongly disallowed, and others simply dispreferred. Take for example the adjectival modification of volumes in sentence 18. While this is an unlikely combination it seems to me to be more acceptable than the internal modification in sentence 17. A constructional variation that is very often discussed in terms of preference is the placement of the particle in verb particle constructions. In such constructions, the particle can appear either before or after the object of the verb. Some verb and particle combinations allow both positions equally, while others have a very strong preference for the placing the particle before the object. Even greater restrictions can be found with established verb particle and object combinations such as *let off steam* as seen in sentence 19. As we can see in sentence 20 this phrase does not comfortably occur with the particle after the noun.

(18) ? Whether it's a boilersuit or a power suit, what you wear speaks great volumes about you.

(19) Most just let off steam by shouting and screaming, but one in five admitted to lashing out.

(20) ? Most just let steam off by shouting and screaming, but one in five admitted to lashing out.

This is just a selection of the kind of syntactic restrictions that exist for MWEs.

Before moving on, we would like to make a point about the form of MWEs that relates to all varieties but is best introduced in this section. Until now all of the examples that we have given have been of fully specified sequences of words. However there are many examples of recurring sequences in which one or more of the words can vary. Fillmore *et al.* (1988) note this, referring to the former kind as **substantive** idioms and the latter as **formal** idioms. [3] A particularly productive example discussed by Fillmore *et al.* (1988) is the *the Xer the Yer* construction. This is seen in the substantive idiom

[3] It should also be noted that although Fillmore *et al.* have in mind lexicalised items, there also exist many purely institutionalised items of this kind. (Moon 1998) refers to lexically specified units as phraseological collocations and lexically open units as frames. Another more recent term which has been used to refer to such items is snowclone, which Pullum (2004) defines as "some-assembly-required adaptable cliché frames".

the bigger they come, the harder they fall but can be used productively as in for example *the more money a political party spends, the more likely they are to win an election* or *the more Jenny worked out, the better she felt.* In the terms we outlined above this is an extragrammatical idiom.

2.2.3 Non-compositionality

In this section we will introduce phrases that are semantically lexicalised. We will refer to these as non-compositional phrases. Before we can describe them, however, we will need to introduce the idea of compositionality.

Competent speakers of a language are capable of great linguistic creativity and produce and comprehend novel utterances which obey the grammatical and semantic constraints of their language on a daily basis. It is clear then that we have the ability to combine our units of meaning (words or phrases) in novel ways, and for communication with such novel productions to be successful. In this respect human languages are often said to be **compositional**.

As a principle of linguistic explanation compositionality is usually said to go back at least as far as the work of Frege. While a number of writers have pointed out that nowhere in his work does Frege actually explicitly state compostionality as a principle of language (see Janssen, 1997) the notion of compositionality is evident in observations such as the following:

> "It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thoughts corresponding to the parts of a sentence, so that the structure of the sentence serves as the structure of the thoughts" (Frege, 1977 quoted in Janssen, 1997:420)

The dominant contemporary tradition in linguistic semantics, following the work of Richard Montague (Montague, 1970;Montague, 1973) defines compositionality formally. Compositionality for Montague entails the existence of a syntactic algebra accounting for the combination of the words and semantic algebra representing the meaning, and, crucially, a homomorphism mapping all elements in the syntactic algebra to the elements in the semantic algebra. Montague was concerned with language as a formal language rather than with how people actually do things, and it is difficult to reconcile his approach with in-context communication between individuals. [4]

In this thesis we will take a broad definition of compositionality, assuming that an expression is compositional if its meaning can be recovered through some combination of its parts. How exactly the meaning is so derived varies greatly from linguistic account to linguistic account, and as we wish to have a broad a definition as possible this will be left unspecified here (although presumably it has something to do with their form and/or the way in which they are ordered). We should also make the point that in focussing on analysability we are making no commitment to how the phrase **is** actually processed by humans. As we discussed in the previous section, it has been suggested that speakers of a language have form-meaning mappings at above the word level for constituents of all kinds of sizes whether analysable or not. We are remaining agnostic on this in our definition of compositionality. When we say that a phrase is analyzable or compositional, we are not saying that in all cases it will be processed compositionally, but simply that it can be. When we say that a phrase is non-compositional we are, however, saying that we do not believe it can be analysed compositionally and therefore that it cannot conceivably be processed in a compositional manner.

[4] Interpretation for Montague involved mapping expressions to elements in models, and entails that there is a closed world described by a fixed set of primitives. In other words it requires that the words or components have a set range of possible meanings that are identifiable prior to their occurrence in a particular context that can be used to explain their meaning in that context. The problem with this is that meaning always depends upon the specific context in which communication occurs, and the knowledge of the specific communicants. Frege recognised this problem when he wrote that "[o]ne should ask for the meaning of a word only in the context of a sentence, and not in isolation." (Frege, 1961:x quoted in translation by Janssen, 1997:420).

We should make clear that we are not (and Frege was not) just talking about the interaction of a word with its linguistic context; linguists have of course proposed a range of ways to deal with context by having underspecified lexical entries that can be resolved to a meaning in context within a compositional framework (e.g. Pustejovsky, 1995). The point is rather that the meaning of a word or a constituent of English is something that depends upon the knowledge or understanding of the language user (or the language users and the negotiation between them) in the specific context in which it is uttered or read. The problem is stated by Tomasello (2001:96–97) as follows:

> "Many theorists, stretching back many centuries in the Western intellectual tradition, describe acts of linguistic reference in terms of just two items: the symbol and its referent in the perceptual world. But this view turns out to be quite inadequate.... especially in its inability to account for the acquisition of and use of linguistic symbols whose connections to the perceptual world are tenuous at best, that is to say, most linguistic symbols that are not proper names or basic-level nouns...linguistic reference can only be understood within the context of certain kinds of social interactions that I will call joint attentional scenes"

This kind of skepticism about determinate linguistic meaning has a long history in the philosophy of language and is associated most strongly with Ludwig Wittgenstein who famously wrote that "the meaning of a word is its use in the language" (Wittgenstein 1953:43)

Having accepted that some aspect of the meaning of sentences in natural languages can usually be attributed to the form and meaning of the component words and the way in which they are ordered, we can get to an account of non-compositional expressions. There exist in all languages some conventional mappings between sequences of words and meaning that cannot be accounted for compositionally. These are constituents that have become semantically lexicalised. Take the pair of phrases *hit the photographer* and *hit the roof* as seen in sentences 21–22. The speaker in 21 seems to be concerned that *Kenneth* will violently assault the photographer. The meaning of the constituent *Kenneth would hit the photographer* can be recovered from our knowledge that the verb *to hit* means to make impact with, that a photographer is a person who takes photographs for a living, that the object usually follows the verb in English and is the element acted upon by the verb, and our knowledge of the intentions and consequences associated with one person hitting another person. So the meaning of the whole phrase might be said to be a function of our knowledge of the meaning of the words *hit* and *photographer* and of the transitive construction, all of which are part of our knowledge of English. For the other phrase, *hit the roof*, the situation is not so straightforward. In sentence 22 we are told how two people either individually or collectively repeatedly made contact with the ceiling of the cabin of what the greater context tells us is an aircraft, and, as with the last sentence, we can recover the meaning using our knowledge of the words *hit* and *roof* together with the transitive construction. However, this knowledge is not sufficient for us to understand sentence 23. This tells us that when he heard about a particular event, *Donleavy* became enraged. Recovering this meaning however requires more than knowledge of the parts. It requires, we suggest, knowledge of a mapping between the whole phrase *hit the roof* and the meaning *become angered*.

(21) I felt so trapped, belted into my window seat with my knees tucked under
my chin, and was terrified Kenneth would hit the photographer

(22) Gillroy and Davies hit the roof several times before managing to strap
themselves down in the cabin where all they could do was sit
ashen-faced and pray.

(23) When he reported this encounter to Control, Donleavy hit the roof
and summoned him to Frankfurt.

So *hit the roof* seems to have two meanings. One is a function of the meanings of the parts (its compositional meaning). The other is a conventional meaning associated with the whole (its holistic meaning). The phrase occurs 21 times in the BNC, and for 16 of these it has the holistic meaning, with the remaining five having the compositional meaning. The reader is able to disambiguate which meaning is appropriate in the context. It is not always the case, however, that MWEs have a compositional meaning that is plausible in addition to their holistic meaning. Take for example the phrase *bite the bullet*, as seen in sentence 24. By convention this means to *accept something unpleasant*. The words do of course have a compositional interpretation. However it is unlikely that in the course of a political discourse we would be told of a politician clamping his teeth on a bullet. In fact it is an act for which it is difficult to see any motivation, and accordingly while there are 8 occurrences of the phrase *bit the bullet* and 20 of *bite the bullet* in the BNC there are no cases where the meaning is compositional.

(24) And wouldn't it be ironic if John Smith, after much agonising about
Labour's soul and purpose, bit the bullet, and, abandoning what many say
was the reason for failure last time, promised not to put up taxes –;
and nobody believed him.

The successful use of these phrases in communicational situations can only be accounted for if their holistic meaning is included in our lexicon. Both of the phrases that we have discussed occur once for every five million words in the British National Corpus. They are therefore recurrent phrases, but not highly frequent ones. And they both consist of component words that are relatively frequent, meaning that it is not clear that they have "distributional privileges of occurrence" (Mitchell 1971:50) of the kind that we discussed before. What makes them important is their non-compositional meaning. Although there may be some overlap with collocations, in that many non-compositional units may be collocations, they are a quite distinct phenomenon. As John Sinclair the linguist most responsible for the popularisation of the notion of collocation over the last 30 years notes:

> "The individual words which constitute idioms are not reliably meaningful in themselves, because the whole idiom is required to produce the meaning. Idioms overlap with collocations, because they both involve the selection of two or more words...we call co-occurrences idioms if we interpret the co-occurrence as giving a single unit of meaning. If we interpret the occurrence as the selection of two related words, each of which keeps some meaning of its own, we call it a collocation." (Sinclair 1991:172)

We should note at this point that pure co-occurrence does have some impact on meaning, in that through repetition an otherwise ambiguous phrase may come to have a strongly preferred interpretation[5]. However such phrases still have a transparent

[5] Bauer (1983) points out that *institutionalisation* has a notable semantic aspect:

meaning that can be recovered from the knowledge of their parts, and are qualitatively different from non-compositional expressions, which do not. Pawley (1986:111-112) makes this distinction, noting the existence of **semantic idioms** which he says "have two meanings, one of which is not predictable from a knowledge of its constituents", and a separate group which he says "admit of several literal readings, of which one the common or lexicalised one" (note that his notion of lexicalisation is different from that we have been using, and encompasses what we have, after Bauer, 1983, been calling institutionalised phrases). Some writers in the computational linguistics literature, however, take a different approach. Manning and Schutze (1999:172), for example, claim that collocations are all non-compositional to some degree:

> "The meaning of a collocation is not a straightforward composition of the meanings of its parts. Either the meaning is completely different from the free combination (as in the case of idioms like *kick the bucket*) or there is a connotation or added element of meaning that cannot be predicted from the parts. For example, *white wine*, *white hair* and *white woman* all refer to slightly different colors, so we can regard them as collocations"

There are significant problems with this analysis. The first is simply the pragmatic reason that if one collapses the terminology in this way then we have no way to make the distinction between the two different kinds of "non-compositionality" they note, the first of which requires special lexical knowledge for any interpretation to be possible, while the second has a difference in meaning that is clearly inferable by the hearer. Secondly regarding non-compositional phrases as collocations is problematic, as many kinds of non-compositional expressions are not at all common (*kick the bucket*, for example, occurs just twice in the British National Corpus, and *kicked the bucket* just three times). The justification for inclusion in the lexicon of such a phrase must be based on factors other than purely the rate of cooccurrence, and similarly for practical purposes the method of extraction must be different too. Thirdly what they are observing in the *white wine* etc examples is simply the context specific nature of word meaning which is by no means unique to MWEs. The word *green* refers to different shades of colour in

"Typical of this stage (especially for compounds) is that the potential ambiguity is ignored, and only some of the possible meanings of the form are used (sometimes only one). Thus, for example, there is nothing in the form *telephone box* to prevent it from meaning a box shaped like a telephone, a box which is located at/by a telephone, a box which functions as a telephone, and so on. It is only because the item is familiar that the speaker-listener knows that it is synonymous with *telephone kiosk*, in the usual meaning of *telephone kiosk*." (p.48) the two phrase *green apple* and *green lawn*, and the meaning of *good* is different in the two sentences *the game Matt saw on friday was good* and the *Santa Claus knew the boy had been good*, but we would not want to argue that they are all non-compositional, or argue for their status as institutionalised phrases. We here prefer to adopt the approach taken by the majority of the linguistic literature and keep them separate.

It should be obvious to say that according to the definition we have given, non-compositional phrases need to be included in the lexicon, and overlap with but are distinct from collocations and syntactically fixed expressions. Rather than giving the impression that they are a homogenous and straightforward group however, we want to give some idea of the range of semantic complexity and diversity of non-compositional expressions. The examples that we have given so far are all pretty straightforward. Whereas for most sequences of words in a language, the meaning of the phrase is a result of the meanings of the individual component words, here the meaning is simply a result of a conventional association between the whole phrase and a meaning. As such, one might think that they can just be entered into the lexicon as large words which contain spaces. Take the phrase *hit the roof.* In its compositional meaning, it is a straightforward transitive construction. There are two referents, the agent and the roof, and a verb which describes a relationship between them. For the purposes of explication it might be useful to see how this would be represented in a semantic notation. In a simple event-based semantics (ignoring quantifiers) the phrase *Gillroy hit the roof* would be represented as follows:

$$\text{hit}(e1,x,y) \wedge \text{Gillroy}(x) \wedge \text{roof}(y)$$

The situation with the holistic usage is somewhat different. There is an agent, *Donleavy*, but there is no referent corresponding to the roof. So the apparently transitive construction exhibited in the phrase *hit the roof* is actually equivalent to an intransitive construction with the agent *Donleavy* and a single verb *hit_the_roof.* This can, then, be represented as a one-place predicate as follows:

$$\text{hit\_the\_roof}(e1,x) \wedge \text{Donleavy}(x)$$

For a certain class of MWEs, an analysis such as that will do. However there is a very large class for which it will not. As we describe in section 2.2.2, despite their formulaic nature, MWEs can take part in many of the syntactic varations available to freely combining expressions. This makes it impossible to deal with them as words-with-spaces. Take for example the phrase *strike a chord*. This has a conventional meaning "to have significance or familiarity". So we can say that a book we relate to, or a poem we recognise *struck a chord* with us. However the phrase is not completely rigid. One variation that can occur is the addition of an adjectival modifier of the noun *chord*. The phrase makes 24 appearances in the BNC, for 2 of which it has such a modifier. The related form *struck a chord* has 52 occurrences, for 12 of which it has an adjectival modifier. An example can be seen in sentence 25. Here *chord* is modified by the adjective *nostalgic* producing the meaning that the cultural artifact discussed has a nostalgic significance.

(25) East London during and just after the war is lovingly portrayed, with an eye and ear for detail which strike a nostalgic chord.

What is happening then is that the meaning of a component of the MWE is being modified by the meaning of the adjective. This is clearly inconsistent with the idea that the meaning of the phrase is completely holistic. A holistic treatment cannot produce the correct analysis in either of the representational languages we have looked at. If *strike a chord* is dealt with as a single predicate, we cannot accurately represent the ways in which its meaning is affected by adding modifiers to the noun.

Nunberg *et al.* (1994) recognise this problem. They refer to unanalysable expressions such as *hit the roof* as **idiomatic phrases**. However they argue that there is separate class of items "whose parts carry identifiable parts of their idiomatic meanings" (p.496), which they refer to as **idiomatically combining expressions**. What is special about such phrases, is that unlike normal compositional phrases, where the meaning of the component words are completely independent of the phrase and can be licensed in many other contexts, in **idiomatically combining expressions** the components have a special meaning that they take on only in the context of the phrase. That is they have a meaning that they can have only when the other elements of the phrase are present. Using the same conventions as before, then, the meaning of *strike a nostalgic chord* might be represented as follows:

$$\text{strike\_idiomatic}(e1,x,y) \wedge \text{AGENT}(x) \wedge \text{chord\_idiomatic}(y) \wedge \text{nostalgic}(y)$$

An alternative explanation of the way that the parts of a phrase such as this relate to its meaning is that they are components in the metaphorical meaning of the phrase which is available to speakers. Some writers (e.g. Gibbs (1994)) have emphasized the live metaphorical content of expressions such as these. However, as Nunberg *et al.* (1994:497) point out, such an analysis can't explain the fact that a phrase such as *spill the beans*, where the metaphorical origins of the phrase are completely obscured, allows such a variation:

> "Note that to call an expression an idiomatically combining expression is not the same as saying it is 'transparent' - that is, saying that speakers can wholly recover the rationale for the figuration it involves... saying an expression is an idiomatic combination doesn't require us to explain why each of its parts has the figural interpretation it does, so long as we can establish a correspondence between it and the relevant element of the idiomatic denotation. When we hear *spill the beans* used to mean 'divulge the information', for example, we can assume that *spill* denotes the relation of divulging and *beans* the information that is divulged, even if we cannot say why *beans* should have been used in this expression rather than *succotash*."

Such phrases, then, require knowledge of unit specific meaning and therefore must, along with the collocations and syntactically-fixed expressions, be included as units in the lexicon of any language description that is to account for language use.

2.3 Empirical evidence for the phrasal lexicon

The previous section introduced the varieties of MWE that have been discussed in the linguistic literature. The argument for the inclusion of these items in the dictionary has been made on the basis of linguistic examples. The presence in real usage of word combinations that cannot be accounted for by either productive syntactic processes or semantic composition provide very compelling evidence for the need to include lexicalised items in the lexicon. However, because the presence of some kinds of MWE in the lexicon is controversial (in particular the storage as a whole of institutionalised phrases by speakers of the language), it will be useful to describe some of the compelling evidence that has been provided by experiments or large scale observational studies that both varieties of "coded" phrase are reflected in the linguistic knowledge of language users. The following section will discuss the evidence that has been provided in various fields.

2.3.1 Evidence for the storage of lexicalised phrases from online experiments

The first significant targeted online study of the phrasal lexicon was that of Swinney and Cutler (1979). They were concerned to evaluate a model of idiom processing that was popular at the time, and which they call the "Idiom List Processing Hypothesis". According to this view, idioms are processed using different resources from those used in processing the rest of language. The view, advocated by Bobrow and Bell (1973), held that idioms are stored in a list that is separate from the regular lexicon. It speculates that all language is first processed "literally" using the individual word meanings, and when this fails a special mode of processing is called on which involves retrieving the phrase from a special idiom store. Swinney and Cutler set out an alternative model, which they call the "Lexical Representation Hypothesis". According to this view, idiom comprehension does not involve any special mechanisms. Instead idioms are retrieved from the lexicon as with all words, and the idiomatic meaning is recovered simultaneously with the "literal" meaning, in the same way that various researchers have shown two different word meanings can be activated simultaneously given an ambiguous context.

Swinney and Cutler evaluate these hypotheses by examining the time taken to process idioms. They use a lexical decision task. Subjects were presented with a series of idiomatic phrases, and control phrases which were identical except for a single word. For example subjects saw the pair *break the ice* and *break the cup*. The phrases appeared on a computer screen and they were asked to judge whether the phrase was a meaningful acceptable phrase of English. The computer recorded the time taken to make the decision. The rationale behind this is explained as follows:

> "If idiomatic meanings are computed by reference to a special idiom list, via some special mode of processing which is instigated following an attempt at literal computation, the phrase classification decisions should take longer, or at least no less time, for grammatical idioms than for nonidiomatic phrase controls. If on the other hand the Lexical Representation Hypothesis holds, decisions made to idiomatic strings should be faster than those made to literal word string controls." (Swinney and Cutler 1979:526)

They indeed found a faster reading time for their "idiomatic" expression and this was taken as support for the Lexical Representation Hypothesis.

The idea that processing time could tell us about idiom comprehension was picked up by other researchers exploring other theories about the lexicon. One important issue that has been explored is decomposability, which we introduced in section 2.2.3. Gibbs *et al.* (1989) presented subjects with a lexical decision task much like that we saw above. They predicted that non-decomposable idioms should be processed in less time than decomposable ones, as decomposable idioms require compositional processing while non-decomposable idioms are simply retrieved as a whole from the lexicon. In actual fact what they found was not only that decomposable idioms are processed faster, but that the non-decomposable idioms in their experiment took longer to process than non-idiomatic "literal" phrases. They interpret this data as evidence that when people read idioms they attempt by default to perform compositional processing. In the case of the decomposable idioms the compositional processing aids the understanding of the idiom, while for non-decomposable idioms, the compositional processing fails and this causes a delay in the processing.

While the result of Gibbs *et al.* (1989) is interesting, it is unclear to me that we should draw detailed conclusions about processing strategies from lexical decision tasks. For example it might be that the subjects in this case are performing a compositional analysis of all phrases simply in order to perform the verification task at hand. It is not clear that this would also be the case in normal language use, where one's goal is not to make a linguistic judgement but to interpret another's intentions.

A number of more convincing results have been provided in the last few years using eye-tracking technology. Titone and Connine (1999) are again interested in whether non-compositional expressions are all processed by retrieving the whole from memory, or whether some compositional processing is performed that assists comprehension. They hypothesise that when an idiom is encountered, both the figurative and the literal meanings are activated. They test this using an eye-tracking experiment. They took a set of 16 decomposable and 16 non-decomposable idioms and situated them in literally biased, figuratively biased or neutral contexts that either preceded or followed the phrase. They consider that if the reading time for idioms is higher when they are preceded by a biasing context than when followed, they can conclude that there is a processing cost associated with selecting between the two active (literal and figurative) readings. They further anticipate that this difference will be significantly less for decomposable idioms where both the literal and figurative meanings are relevant to comprehension than for non-decomposable idioms where it cannot be used. And indeed this is the result that they find. They take this as evidence for their dual-activation model of idioms processing and for the cognitive reality of the linguistic idea of decomposability.

One problem with all these experiments is that, presumably because of the lack of availability to researchers of adequate corpora, they do not control for the frequency of the phrases. It may be that the difference in reading time could be explained away as an effect of frequency (assuming people process familiar phrases faster than unfamiliar ones). Apart from this consideration however, these results seem to provide evidence that non-compositionality affects processing.

2.3.1.1 Evidence for the storage of institutionalised phrases from online experiments

We saw in section 2.2.1 that institutionalised phrases are recurrent in language use. We hear and read them more than other phrases. And they seem to have some privileged status in speakers' knowledge relative to other phrases. We might expect, then, to find evidence of some processing advantage. This section will discuss what evidence there is for this.

The first thing to establish is that the effect of frequency on word processing gives us probably the most robust result in the psychological literature on language. Howes and Solomon (1951) presented a number of words to subjects, first of all for very short durations, then incrementally increasing the display time, and asked them to identify the word. They found that the mean time taken by subjects to recognise the words correlated significantly with the log frequency of that word as measured over a corpus of a few million words of English. Howes (1957) examined the ability of hearers to identify spoken words that were disguised by noise. He found that the noise threshold at which subjects were able to identify noises correlated with log frequency, with each log unit of word frequency dropped requiring a drop in noise of about 4.5 db for identification to occur.

These results suggest that frequently occurring words have a processing advantage over infrequent words. This finding has been repeated for both written and spoken words using a range of different experimental paradigms. For example, Forster and Chambers (1973) asked subjects to read aloud visually presented words, and found that the time taken to do so was shorter for words than non-words, and for high frequency words than for low frequency bands. Rubenstein *et al.* (1970) asked participants to distinguish English words from non-words that were visually presented. They found that the response times (or the time taken to begin speaking) were faster for words of high frequency than of low frequency. These results are all for individual words. There has been far less work exploring the effect of frequency on the processing of multiword sequences. The following paragraphs will describe the key works.

Bower (1969) was concerned to examine the "chunking" hypothesis. This suggests that the key to memory performance is how many chunks of information are involved. Bower is concerned with words and sequences of words. He suggests that because of our chunking ability, subjects should be able to store and recall units of three related words (which can be chunked) as efficiently as they can individual words, while units of unrelated words should be less easy to recall. Subjects performed a free recall task. They were presented with a list of 24 units consisting of 12 critical words and 12 fillers. In one condition the fillers were simply 12 nouns presented individually. In a second condition the fillers were 12 groups of 3 unrelated nouns. In the third condition the fillers were 12 familiar (intuitively thought to be high frequency) three word clichés (*ball-point pen*, *mail order catalogue*, *Rose Bowl Parade*, *Birth Control pill*, *ice cream cone*, *Bay Area transit*, *tick-tack-toe*, *turtle neck sweater*, *happy new year*, *fair-weather friend*, *great salt lake* and *good old days*).

Bower (1969) found, as hypothesised, that there was no significant difference between the recall accuracy for the single words and the familiar chunks, but a decrease in accuracy was found for the unrelated word triples. He concludes that the clichés are treated in every respect like singles words in recall. He concludes that the limit on recall is down to chunks and not words. The reason this result is of interest to us here is that the special performance that is shown for frequently recurring chunks would seem to be down to the subjects having some prior memory trace which gives them a processing advantage. It suggests that people have memories of frequently occurring chunks that are stored in the same ways as individual words.

This study is very interesting. However a) its objective is not to provide evidence for the storage of institutionalised language, and consequently uses a memory task rather than a natural language process, and b) it uses intuitive rather than objective assessments of phrase frequency. Another study to rely on similarly intuitive assessment is that Lancker-Sidtis and Rallon (2004). They are concerned to show that people have stored representations of formulaic sequences. They obtained a list of items that are formulaic and not, by having annotators classify the utterances from the script of the film "Some Like It Hot". They then had subjects perform a phrase completion (cloze) task. A significantly greater accuracy was found for the formulaic sequences than for the non-formulaic. This suggests that the subjects had stored memories of the formulaic sequences. The remaining work in this chapter all use objective estimates of frequency taken from corpora.

Lapata *et al.* (1999) were concerned to find whether the distributional characteristics of adjective-noun combinations determine the extent to which speakers find them "plausible". They found a very strong relationship between the frequency of combination and the plausibility judgements of their subjects. Their materials were 120 such combinations, constructed by selecting 30 adjectives, extracting all adjective noun combination featuring them in the BNC, and selecting for each a high, medium and low frequency pair using an equal division of their log-transformed frequencies over the corpus. They obtained plausibility judgements by presenting these 120 pairs to 24 subjects and asking them to indicate how plausible they found them using a magnitude estimation procedure.

This procedure involved participants assigning an arbitrary number to a modulus item presented to them, and then assessing the plausibility of each combination by assigning them a score relative to the modulus. The judgment for each combination is then taken to be the ratio of this score and the modulus item. The geometric mean of the judgements was 2.966 for the high frequency items, 2.660 for the medium frequency items, and 2.271 for the low items. An ANOVA showed a significant difference between the ratings for the groups at a $p < .001$ level both by items and by subjects. The authors then performed a correlational analysis of the relations between the plausibility scores and the frequencies, as well as other distributional characteristics. They found the strongest correlation between plausibility and frequency, with a pearson's *r* of .570, followed by the relationship between plausibility and log-likelihood ratio (see section 4.5.1) which also had a correlation significant at .01, and the conditional probability of the noun given the adjective [6] which was significant at .05. This study suggests that a speaker's assessment of the plausibility of a word combination is determined by their remembered experience of previous encounters with the combination, and that this is affected by the frequency with which it occurs in the linguistic environment of which the BNC is taken as representative.

MacDonald (1993) conducted reading time studies looking at the resolution of lexical category ambiguities. She is interested in discovering what information is used by the reader in deciding the lexical category of an ambiguous word, and specifically whether syntactic information alone is involved or whether they also use lexical or se-

[6] This was calculated as follows:

$$P(noun|adjective) = \frac{freq(adjective, noun)}{freq(adjective)} \tag{2.1}$$

mantic information. She looks at the comprehension in context of word pairs that are are ambiguous between noun-noun and noun-verb interpretation. One example provided is the word pair *desert trains*, which can either be two sequential nouns as in the sentence *I know that the desert trains could resupply the camp* or a verb followed by a noun as in the sentence *I know that the desert trains soldiers to be tough.* She studies online comprehension of this by looking at varying reading times in the ambiguous region of the sentence. assuming that increased reading time immediately after the word pair is indicative of comprehension difficulties.

Of interest to us here is that MacDonald looks at the effect of frequency information, and particularly at the effect of the frequency with which the word pair occurs as two nouns on the reading time for sentences in which the second word is a verb. Her hypothesis is that a reader will have have greater difficulty in reading the latter variety of sentence if the word pair is familiar to the reader as a noun-noun pair. In the study she indeed found that the frequency with which the pair occurs in the nominal interpretation in her corpus correlates with reading time for between 4 and 6 words after the second word in the pair, and a marginal correlation up to as many as 8 words after. MacDonald interprets this result as evidence that readers store memories of the syntactic properties of specific sequences of two words as well as of the individual words and of abstract lexical categories.

Perhaps the most convincing result concerning the processing of frequent word pairs is that of McDonald and Shillcock (2003). They were interested in factors that determine the duration for which a reader's gaze is fixed on a word during reading. In particular they are interested in whether this gaze is determined by the predictability of a word given its preceding or its following word. They quantify this using a the British National Corpus, and look at two related measures, the forwards and backwards transitional probabilities. The forwards transitional probability is the likelihood of seeing a word y given the preceding word x. This is calculated as follows:

$$P(y|x) = \frac{freq(xy)}{freq(x)} \tag{2.2}$$

The backwards probability is the probability of seeing a word given the word that follows it. This is calculated as:

$$P(x|y) = \frac{freq(xy)}{freq(y)} \tag{2.3}$$

They test this by comparing this with a corpus of eye movements during reading created using an eyetracker. In a regression analysis, they found that the total gaze duration (the sum of the total fixations on a word) decreased as both the forward transitional probability and the back transitional probability increased and that the first fixation time also decreased as the forward transitional probability increased. This result very strongly suggests that readers do store memory traces in some form of the pairs of words that characteristically go together, and that this gives a processing advantage. This is evidence that certain word combinations become institutionalised in the memories of speakers.

These last three pieces of work provide very good evidence that language users have memory of frequent two word chunks. However, languages contain a great many frequently occurring chunks of language of more than two words. There is to my knowledge only a single piece of work to date looking at the storage of longer chunks. Bod (2000) reports on a study looking at the processing of three word sequences of varying frequency. He extracted 50 frequent subject-verb-object sentences from the British National Corpus and the world wide web. All sentences were judged by the researcher to be semantically transparent and non-idiosyncratic. For each sentence, three additional sentences were created by substituting each of the three words in turn with a roughly equally frequent word of the same category and length. Each of these were low frequency. These were split into two lists of sentences with each containing 25 of the high frequency sentences. Subjects were presented with the phrases on a computer screen and had to decide whether they were an acceptable English sentence or not. A shorter reaction time was found for the frequent sentences than for the infrequent. From this Bod concludes that frequent three-word sentences are stored and retrieved from memory rather than processed online.

As these results are currently only available from a presentation and a later abstract (Bod 2001), many of the details of the study are unavailable. Nonetheless it is potentially a very interesting result. It seems to provide good evidence for the storage of some aspects of the sentence. However no control is performed for the bigram frequencies that make up the three word sentences. We saw above that there is good evidence that frequent two word combinations are stored somehow. As Bod does not ensure that these are not controlled for his three word sentences, we cannot be sure that the decreased reading time is not simply a result of these component transitional probabilities.

To summarise the findings described in this section, we have very convincing evidence for the storage of recurrent two word sequences. We are, however, yet to have any solid evidence that speakers have any memory trace of longer chunks.

2.3.2 Other evidence for the storage of institutionalised phrases

We saw above that a number of researchers have reported apparent experimental evidence for the storage of lexicalised phrases of English. However, the experimental support for the storage of institutionalised phrases of more than two words is less complete. We noted before the different status that the two kinds of MWE have in the linguistic literature. The necessity of including lexicalised phrases in the lexicon is more apparent from a small number of examples. For example, it is straightforward to see that a non-compositional phrase needs to be included in the lexicon in order that its meaning be accounted for. The existence of institutionalised phrases on the other hand only becomes irrefutable over large sets of data, the like of which have been available for just a couple of decades. For this reason the storage of lexicalised phrases has been more widely accepted, and the subject of more experimental work. We will therefore outline here the evidence from other fields of research that institutionalised phrases belong in the lexicon.

2.3.3 Evidence from first language acquisition

Thus far we have discussed adult language. However a notable body of evidence for the storage of multiword units comes from the study of child language acquisition. For many years, the predominant approach to first language acquisition was to treat their linguistic knowledge as quantitatively different from that of adults, but qualitatively the same. That is to say that they were assumed to be processing language in the same analytic fashion as adults. However, there is a growing body of evidence that children's early language use is significantly holistic. Tomasello (2003:137-138) explains this as follows:

> "The child's major symbolic vehicle at this early stage is what is often called a holophrase: a single-unit linguistic expression intended as an entire speech act...most children begin language acquisition by learning some unparsed adult expressions as holophrases - such things as "I-wanna-do-it," "Lemme-see," and "Where-the-bottle"."

There is a long literature reporting observational evidence for this, with varying degrees of formality. Clark (1970) reported on a particular habit in the developing language of her son, Adam. She reported that a great number of Adam's utterances directly incorporated the adult's previous utterance. For example, when told *We're all very mucky*, Adam replied *I all very mucky too.* And when he was trying to put his coat on by placing his hands in the wrong end of the sleeves, and was prevented at told *That's upside down*, Adam grabbed the coat back and replied *No, I want to upside down.* Clark suggests that "not all of the constituents of his utterance were necessarily being processed at all three linguistic levels, phonetic, syntactic and semantic, but some sequences may have been taken over as unopened packages from the previous adult utterance" (p.3) . She reports that at the age of between 2 years 9 months and 3 years, Adam's speech seems to consist in part of "routine unproductive sequences". She further reports that many productive rules seem to originate as unproductive sequences, suggesting that the learning of complete multiword units is a crucial stage in the development of language.

In a seminal paper, Peters (1977) describes the speech of a child she calls Minh, whom she recorded for up to an hour a week from the age of 7 months to 2 years and 3 months old. She recounts how from the age of 17 months, he demonstrated a style of speech that she calls "mush-mouth". This was mumbled production, where phrases were approximated by their intonational contours, despite unrecognisable segments. These aim at the reproduction of whole sentences or utterances rather than individual words. Many of his utterances were very frequently repeated phrases for which context was very useful in determining meaning, such as *look at this!*, *what's that?* or *open the door.* However he also attempted to reproduce set repetitive utterances, using an established filler sound for the unclear parts. Peters suggests that this kind of "gestalt" learning of speech is an important part of language which tends to be overlooked by researchers who look in child speech for the same units and levels that are supposed for adult language.

An early more formal study is that reported by Tomasello (1992). This work is based on a diary of the author's daughter's speech in her second year, and is concerned with her learning of verbs. It describes how the child's language use can be best accounted for by positing that her knowledge consists of a repository of stored constructions that are specific to individual verbs rather than of abstract syntax. It is reported that she used many verbs only in very simple constructions (e.g. *cut* ____), while others are used in multiple constructions of greater complexity (e.g. *draw* ____ , *draw* ____ *on* ____ , *draw* ____ *for* ____ , ____ *draw on* ____) without appearing to generalise knowledge across verbs. The strongest evidence for this is that the greatest predictor of her use of a particular verb on any day was not the use of other verbs on that same day, but rather her use of that exact verb in the days before. This idea that a child's early syntactic knowledge is based around specific lexical items in this way is referred to and has become widely known as the **Verb Island Hypothesis**.

Lieven *et al.* (2003) find a similar dependence on item specific constructions in a very intensive corpus of the speech of a single child recorded over a six week period, beginning on her second birthday. This child, known as Annie, was recorded for an hour per day for five days per week. The mother also kept a diary of any new utterances that she noticed. The researchers then took the transcripts of the last day of the 6 week period, and studied the creativity of Annie's speech by looking at how many of the child's utterances over this hour had been heard before over the last 6 week period. Their results are quite remarkable. They discovered that of the 295 utterances, 63% had been said before at some point during the six weeks of recording by either the mother or the child. Of the remaining 109 utterances, 74% only differed from the nearest previous utterance by a single edit operation of substitution, addition, deletion or swap. This suggests that she is depending to a very large degree upon a stored repository of formulaic, lexically specific constructions.

This section has reported on a body of research which suggests that early child language consists to a large extent of stored multiword utterances. The reason this is relevant is that if it is the case that children acquire language by storing such sequences, which they then use to develop the more creative adult linguistic competence, we might expect that they will continue to store traces of multiword utterances where it is efficient and useful for them to do so as they become fully competent adult speakers.

2.3.4 Evidence from language pathology

The literature on language pathologies and neurological disorders is full of accounts of individuals that seem to have stored established sequences of words. The tenor of most of these accounts is apparent in this very early account (from 1683) from Peter Rommel:

> "She could say no other word, not even a syllable, with these exceptions: the Lord's Prayer, the Apostle's Creed, some Biblical verses and other prayers, which she could recite verbatim and without hesitation, but somewhat precipitously.... Then we tried to determine whether she could repeat very short sentences consisting of the same words found in her prayers. However she was unsuccessful in this." (quoted in Benton and R.J.Joynt, 1960:209–210)

Many patients who are completely unable to produce novel sentences frequently produce set utterances, often with no regard for the conversational context. This is usually accepted as evidence that they have a store of such utterances. Serious discussion of this subject dates back at least as far as the writings of John Hewlings Jackson in the second half of the nineteenth century. Jackson (1879a:174) makes a distinction between propositional and non-propositional language, and describes how patients that he calls speechless, are often capable of non-propositional utterances but not of propositional ones:

> "The recurring utterance is sometimes a phrase. In one case "come on," or sometimes that patient uttered "Come on to me". In another case, just mentioned, it was, "Oh! my God!" In another case, mentioned to me by Dr Langdon Down, "Yes, but you know"...These phrases which have propositional structure, have in the mouths of speechless patients no propositional function. They are not speech, being never used as speech; they are for use only compound jargon; they or their tones are at best of interjectional value only. The man who uttered " Come on to me," uttered it on every occasion when he made a rejoinder to anything said to him"

As with the example from Rommel, since these patients are not capable of spontaneous or creative speech, the recurring utterances of which Jackson speaks are considered to be stored chunks of language which are simply reproduced from memory without any intention or meaning. Jackson (1879b) argues that speakers have two separate faculties, one for this kind of automatic speech and one for creative propositional speech, with the former being situated in the right hand side of the brain, and the latter in the left hand side. In the patients he describes the left hand side of the brain has been damaged, while the right was intact. The literature on aphasia is full of such accounts. A large scale survey of the literature is offered by Code (1987).

While most reports are purely from clinical records, Lum and Ellis (1994) provide experimental evidence of an ability for non-propositional language in aphasics. They tested them on six paired tasks where in one they were asked to use language propositionally, and in the other they were asked to recall language from memory. One task was to identify 15 objects (e.g. a tree), where there were either given a cue in the form of an incomplete "familiar" phrase (e.g. He barked up the wrong ____), an incomplete "unfamilar" phrase (e.g. He climbed up the tall ____), or no cue at all. A significant increase in performance was found for familiar phrase condition relative to both the unfamilar phrase and the no cue conditions. This is taken as evidence that the patients find processing "non-propositional" language easier that the "propositional" examples.

To summarise, there is a vast literature reporting that brain damaged patients show an ability to reproduce frequent chunks of language even when their ability to produce original sentences has completely gone. This suggest that speakers of a language have a repository of recurrent sentences and phrases in that language, and thus provides support for the idea that institutionalised word sequences are stored in the lexicon.

2.4 Chapter summary

This chapter has introduced the varieties of multiword expression that we are going to consider in future chapters. We began by outlining the two different kinds of argument, quantitative and qualitative in nature, that have been used to justify the inclusion of MWEs in the lexicon. We introduced the terms institutionalisation and lexicalisation to describe these. We then discussed institutionalised phrases and made a case for their inclusion in the lexicon, before introducing the two different kinds of lexicalised phrase that we are also going to consider in the thesis, the syntactically fixed phrase and the non-compositional phrase. Having presented the linguistic evidence for each of these varieties of MWE, we gave an outline of the published experimental and observational evidence for the storage of MWEs by language users.

Chapter 3

Frequency and the phrasal lexicon

The role of frequency in language has been the object of increasing attention in recent years. In her 2005 presidential address to the Linguistics Society of America, entitled *From usage to grammar: The mind's response to repetition* Joan Bybee gave relative frequency its place as one of the most important factors in the cognitive, historical and social dimensions of language. She writes that:

> "A usage-based view takes grammar to be the cognitive organization of ones experience with language. Aspects of that experience, for instance, the frequency of use of certain constructions or particular instances of constructions, have an impact on representation that are evidenced in speaker knowledge of conventionalized phrases, and in language variation and change." (Bybee 2005:1)

This focus has been made possible by the widespread availability of large and well designed corpora.

Frequency has been of particular import in the study of the single word lexicon. Frequency information is now available not only in dedicated volumes (e.g. Francis and Kucera, 1982), but has also come to be included in lexical resources for both specialist (e.g Miller *et al.*, 1990, Baayen *et al.*, 1993) and general audiences (e.g. Sinclair *et al.*, 1987, Soanes and Stevenson, 2005). Furthermore this information is now widely used in the conducting of experiments and the design of teaching syllabuses. As we saw in section 2.3.1.1, the processing advantage discovered for frequent words over infrequent words is one of the most robust findings in cognitive psychology. The use of corpora was essential to these experiments.

It might seem logical to assume that corpora and the interest in frequency would also have a significant impact on the study of the phrasal lexicon. Armed with a corpus and some basic programming skills the lexicographer can very easily count the occurrences of words and phrases in even the largest of corpora. However, while there has been some good corpus-based work on the processing of frequently occurring two word combinations (see section 2.3.1.1), there has been very little real empirical work on the frequency of phrases of any greater length by either lexicographers or psychologists. While dictionaries of phrases and idioms might make use of corpora in their development process, the exact contents of such dictionaries are still to a large extent based on precedent. There exists no published list of phrase frequencies of the kind that exists for words. And psychological studies of the effect of frequency in the processing of lexical units of more than one or two words remain rare and limited in scope. As Jurafsky (2003) notes: "...the vast majority of frequency effects that have been studied involve lexical structure. A small number of studies have looked for frequency effects for larger (supralexical) structures, but the results are relatively inconclusive".

The significance of frequency information in accounting for the phrasal lexicon has been questioned by some key scholars of phraseology. Alison Wray writes the following:

> "...it may be premature to judge frequency as a defining feature of formulaicity. It has yet to be established that commonness of occurrence is more than a circumstantial associate. There are certainly many formulaic sequences whose culturally-based familiarity belies their comparative rarity in real text (e.g. That's another fine mess you've gotten me into; Time for bed, said Zebedee; Here's one I made earlier). As Hickey (1993) notes, "we must not rule out the possibility that an utterance which does not occur repeatedly is a formula" (p. 33). In other words, "phraseological significance means something more complex and possibly less tangible than what any computer algorithm can reveal" (Howarth, 1998, p. 27)." (Wray and Perkins 2000:6)

Further doubts as to the usefulness and reliability of corpus frequencies have come from inside the corpus linguistic community. Perhaps the largest scale corpus study of multiword phrases is that of Moon (1998). This conducts a substantial corpus study of 6776 "fixed expressions and idioms". However, rather than selecting the phrases of interest based on a study of the corpus, Moon begins with a list of phrases taken from previous lexical resources (which were based on linguistic intuition rather than data) and proceeds to locate them in the corpus by hand and examine them. She acknowledges that it is unfortunate that she must bring this external resource to bear, but argues that it is essential:

> "Ideally the FEIs [fixed expressions and idioms] would be identified automatically by machine, thus removing human error or partiality from the equation. There is, however, no evidence that this is possible given the current state of the art. It is difficult to see exactly how progress can be made...The problems arise because in so many cases FEIs are not predictable, not common, not fixed formally, and not fixed temporally (that is, they are often vogue items like slang). They are dynamic vocabulary items, whereas - at least at present - corpus processing requires givens and stability"

She highlights two main factors here - rarity and inconsistency. She suggests that on the one hand items of interest do not occur with adequate frequency to be found in corpora, and on the other that they occur with such inconsistency that they will not be reliably represented in any corpora.

This thesis is concerned with exploring what kinds of information we can learn about phrases from text corpora using computational tools. If what Moon states is correct then it is unclear that we can acquire any useful information at all. And if what Wray says is true then it is not clear that any information collected would be of interest. It is important then to consider their arguments. This chapter will explore the stability and usefulness of corpus phrase counts. The first half will look directly at the question of adequacy, examining the frequency and consistency with which phrases occur, aiming to establish whether corpora of manageable size can offer us counts for units of more than one word that are stable and informative. The second half will look at the question of utility, attempting to determine whether the distribution of phrases found in corpora are reflected in speakers' knowledge and/or processing of language, by establishing whether frequently occurring sequences have the same kind of processing advantage over infrequent sequences that we see for frequent over infrequent words.

3.1 Establishing the informativity and reliability of Corpora

The arguments for limiting the role of corpora in the description of the multiword lexicon concern two factors - rarity of appearance and instability of occurrence. Our main aim in this chapter will be to look empirically at the second issue. Over the next four sections we will provide evidence that the multiword sequences that are found in corpora can be reliably counted so that information about their frequency in the corpus is indicative of their salience in a language. However before moving on to this question, we want to discuss the question of rarity and whether or not a significant number of the MWEs in a language can be expected to be found in corpora in the first place.

A key assumption in this thesis is that a significant part of the MWEs that are used in English will be found in corpora. The null hypothesis, then is that the majority of MWEs are too rare to be reliably found in corpora. This is actually extremely hard to reject. Clearly if they are not there in corpora then we cannot reject the null hypothesis by showing that they are. Our only alternative is to question whether such a substantial group exists at all. We assess the plausibility of the argument by examining what we know about the distribution of events in language, and looking at whether the null hypothesis is consistent with this.

The first point to make is that there are many repeated word sequences in corpora that occur with very considerable frequency. Figure 3.1 shows the rate of occurrence of strings of up to seven words in the written component of the British National corpus. We can see that many phrases occur with a considerably higher frequency than many single words that are regarded as part of a speaker's central vocabulary. So we can see that the phrase *only the tip of the iceberg* occurs almost ten as many times as *ransack*, and *at the end of the day* eight times as many as *cookie* and *at the same time*, almost 100. If we take the standard approach to frequency, and created frequency bands by dividing the log-transformed frequency scale for individual words equally into three groups (high, medium and low), we find that in the BNC there are 70633 sequences of two words or more and 23019 phrases of three words or more that occur in the medium frequency band for words or above.

There are, then, a great many multiword sequences found in corpora. However this doesn't preclude the possibility that there are also great many items of interest which are not found. Corpora are snapshots of the linguistic environment. The larger available corpora are assumed to provide a coverage of the vocabulary of English that is broad enough that they can be used for the creation of dictionaries. We know that there are many words that are used in the language that will not occur in even 89 million words of English. However, these are by and large assumed to be so infrequent that they are not part of the central vocabulary of the language or not necessary to be included in reference works. In fact one of the major arguments that is made for the use of corpora in lexicography is that they allow one to focus on words that are actually

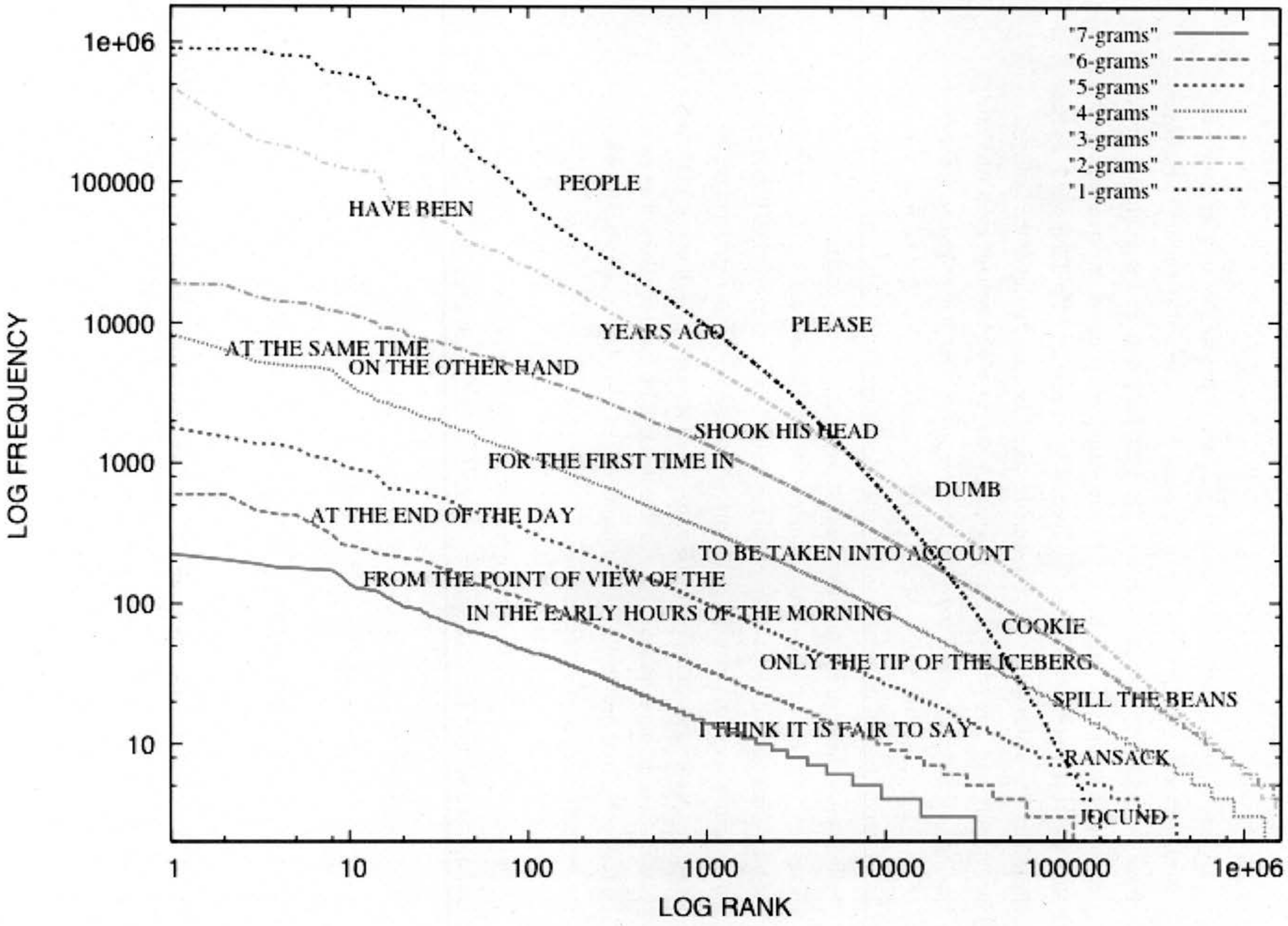


Figure 3.1: Annotated Zipf curve for all words and n-grams of up to length 7

used [1]. How different should we expect the situation for MWEs to be?

A corpus like the British National Corpus is a large snapshot of a language. Estimating exactly what portion of language use is represented is not easy. However some writers have proposed estimates. The most well-motivated example was given for children's linguistic experience. Extrapolating from limited duration recordings, Hart and Risley (1995) estimates that a child hears at most 11 million words and at least 3 million words a year, varying with the socio-economic status of the family. It is not entirely clear how to use this to make estimates for adult language users who exist in linguistically more varied and often intensive settings, but it does give us some idea of the magnitude of the figure. Moore (2003) estimates that an adult hears an average of approximately 14 millions words a year. Assuming these figures are in the right ballpark, it is clear that a corpus like the BNC represents a very long period of linguistic experience. Based on Moore's estimate, the 89-million-word written component of the corpus that we are using in our experiments is equivalent to more than 6 years of an adult's linguistic experience (assuming that the corpus is representative of the language people encounter. We will discuss this in section 3.1.1). If our definition of a MWE is a recurring sequence of words that is stored by many speakers, then surely it is reasonable to expect that a reasonable number of these sequences would occur in 6 years worth of linguistic experience.

In order to realise quite what a significant portion of linguistic experience this is, it is important to realise a fact about the distribution of events in language that applies to both words and phrases. There is actually an empirical basis for the assumption that a very significant part of the vocabulary of English will be seen in the scale of corpora currently used in lexicography. This is founded in the frequency distribution of human language. The best known account of this distribution is that of Zipf (1935). This describes how the frequency of any wordform is inversely proportional to its rank in

[1] Sinclair (1991:38) writes that many hand written dictionaries contain many "[f]orms and/or meanings which have lapsed into disuse, but are not so indicated" and "[f]orms and/or meanings which are constructs of lexicography, and do not really exist, in the sense that there is no textual evidence for them". He states that "lexicographers should be scrupulous in extirpating these items", and that no word should be included in a dictionary purely on the basis of lexicographic tradition unless there is independent evidence for it, by which he appears to mean its occurrence in a corpus.

the frequency-ranked list of words in its language[2]. Or most importantly for us, that there are a few very frequent events in human languages and a great many rare events. This has significant consequences for the number of new events that we can expect to see in a corpus of a given size. In small samples, the addition of small amounts of new data greatly increases the number of unique items. However, because of the frequency spectrum of languages, and the disproportionate percentage of language that is made up of items from the higher part of the frequency range, there is a curve so that the number of new items that will be seen for each incremental increase in data decreases as the total amount of data increases.

This fact makes the finite nature of corpora more palatable to researchers. It means that in any corpus of reasonable size we will see a good part of the vocabulary that we would see over a corpus of a much larger size. Given the observed pattern of vocabulary growth, there are ways of extrapolating exactly what part that is. Using a version of the Zipf-Mandelbrot model described in Evert (2004) and implemented in Evert (2004), we can extrapolate from the observed vocabulary growth that doubling the size of the written component of the BNC would only increase the vocabulary size by a third. According to this model, in order to obtain a coverage of twice the vocabulary size we would need to a corpus of five times the size.

Given these facts about language, the assumption that corpora of only 89 million words are adequate to study the vocabulary of English seems sound. It is very interesting to note then that the situation for multiword sequences in not qualitatively different. Ha *et al.* (2002) report that the Zipf-Mandelbrot law is not only extendable to multiword sequences, but that in fact the law more accurately describes the pattern for such sequences than for words. This can be observed in figure 3.1, with the longer sequences displaying a smoother line that individual words [3]. The consequence of this is of course the same pattern in vocabulary growth. The same curve as for words can

[2] In its original formulation this stated that the frequency f for any word could be calculated as follows, where r is its rank, and k is a constant for the text or collection of texts:

$$f = \frac{k}{r} \tag{3.1}$$

This has been reformulated many times since. The most often used reformulation is that of Mandelbrot (1953), which introduces two additional text/corpus specific constants α and β:

$$f = \frac{k}{(f+\alpha)^{\beta}} \tag{3.2}$$

[3] The Zipf-Mandelbrot law entails that a double logarithm plot should give us a straight line.

be observed with vocabulary growth decreasing as corpus size increases.

Again using the Zipf-Mandelbrot model implemented in the UCS toolkit (Evert 2005), we can obtain estimates of the number of unique strings of different sizes we would expect if we had larger corpora. According to this model, in a corpus 10 times the size of the BNC one would only observe 5 times as many unique two-word strings. Thinking about this in a different way, if the estimates of the volume of speech heard are correct (and the BNC is representative), then the number of two word strings contained in the BNC is equivalent to almost one fifth of the two word strings one can expect to hear over 60 years of adult life. Similarly this model tells us that the BNC contains more than one ninth of the unique four word strings one can expect to encounter in a corpus of that size. This is a not insubstantial portion of the multiword sequences that one will hear. And it seems more so when one considers that the top two thirds of the frequency range for four word strings in the BNC accounts for less that one hundred thousandth of the total items seen. Assuming the frequency of multiword strings is stable (we will get to this question shortly), it should be clear that all of the multiword strings that one will encounter with anything greater than very very low relative frequency in a lifetime will occur in a corpus of much smaller size than the BNC.

The point of these arguments is to show that while there are doubtless MWEs that belong in the lexicon that are not found in corpora on the scale of the BNC, a substantial proportion will be. Furthermore rather than treating the absence of a traditional MWE from the BNC as a fault of the corpus or a result of its inadequate size, it should be recognised that a corpus of the size of the BNC represents a very significant snapshot of linguistic experience, and if an item is not found in there, while we should not reject it out of hand, we must at least question whether it is really part of common vocabulary and stored by a large community of speakers, and not rather an item that has been passed down to us by lexicographic convention and has ceased to be part of the living language.

One significant as yet unmentioned problem is that the figures from the BNC we are discussing are for written rather than spoken language. There are many differences between the two varieties (Biber 1988), thereby making generalisation problematic. There will almost certainly be MWEs that occur much more frequently in spoken language, and any description of the phrasal lexicon would need in reality to examine both. Nonetheless spoken language is usually reported to be more rather than less repetitive than written language, as seen, for example, in the significantly lower type/token ratio (Yates 1996), and so we would expect a corpus of spoken English of equivalent size to contain at least the same proportion of linguistic experience.

We are now going to turn to the question of whether the occurrence rate of MWEs is stable enough that we can obtain accurate counts from available corpora. Section 3.1.1 outlines the problem of assessing the stability of corpus counts, and introduces some measures that have been proposed. Sections 3.2 and 3.3 describe experiments that employ these measures in order to evaluate the stability of multiword counts taken from the BNC.

3.1.1 Measuring burstiness

When we talk of the statistics of language we often talk about it in idealised terms. We talk about words or constructions as being frequent or infrequent in a given language, or a word combination as being particularly probable relative to another in a given language. This is a reasonable and usually useful way to think about things. It is, however, not a completely accurate picture. While the language of a football commentary is similar enough to the language of textbook in microbiology that we can distinguish that they are both English and not French or Korean, they are in many respects very different.

Language variation across genres and topics is of interest in many fields of linguistics. In the field of Applied Linguistics, for example, there is a concern with the construction and analysis of corpora covering specialist texts (e.g. Bowker and Pearson (2002)). And in various subfields of computational linguistics there is an active interest in creating statistical models of language that are either designed for specific subdomains, or can adapt to different domains according to need. However for many purposes we want to be able to take a single corpus and draw conclusions about the language as a whole. One solution to this problem has been the creation of multi-purpose corpora that span genres.

The aim according to McEnery and Wilson (1996) is as follows:

> "We are therefore interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. What we are looking for is a broad range of authors and genres which, when taken together, may be considered to "average out" and provide a reasonably accurate picture of the entire language population in which we are interested."

Such corpora enable one to address a wide range of tasks. For example, one of the most widely used corpora in computational linguistics is the British National Corpus. This is a 100 million word collection which was constructed so as to "ensure that the corpus contained a broad range of different language styles...so that the corpus could be regarded as a microcosm of current British English in its entirety, not just of particular types." (Burnard 2000:6). It has been used for tasks such as thesaurus extraction (Curran 2003) and ambiguity resolution (Lapata and Lascarides 2003a), which achieved high performance when evaluated against independent published recourses and subject judgements. The field of psycholinguistics has also tended to employ a single mixed-genre corpus. The most widely used corpus resource in experimental work remains the published word frequencies from the one million word cross-genre Brown Corpus (Francis and Kucera 1982). And, crucially for us, in the field of lexicography, the corpora that are used by the two largest British producers of dictionaries (The Bank of English as employed by Harper Collins (Sinclair *et al.* 1987), and the Oxford English Corpus as employed by Oxford University Press (Soanes and Stevenson 2005)), both cross genres and claim to be representative of the scope of British English.

So it is for many purposes considered possible to obtain a snapshot of a language from a corpus that is sufficiently representative of a whole language that it can be used to describe it. And yet what Moon suggests is that while this counts for most language and crucially for individual words, it is not appropriate for MWEs. While for words it may be the case that things "average out" in a mixed-genre corpus, she suggests that MWEs are more sensitive to the contingencies of register, genre and fashion and that no corpus can give a reliable snapshot. This chapter will look at this empirically.

We can think about this more formally. When we talk about frequencies in language we are talking about their rate of occurrence. What we want to know is how many times any word x will be seen in any text of N words. According to the idealised view of language that we discussed above, the rate of the occurrence of linguistic events such as words is constant across all contexts and varieties of language. Take for example the word *entrance*. This occurs 2818 times in the written section of the British National Corpus. These consist of 89.39 million words. Given these statistics, if the rate of occurrence of the word was constant, we might naively expect to find approximately 32 occurrences of the word in every 1 million words. The written component of the BNC is made of a total of 3144 independent texts, giving an average text size of 28431. This means that we have a mean occurrence of 0.9 per text[4].

[4] We can in fact be more precise about the predictions of this naive view. What this view assumes is

We can observe directly whether or not this is a good approximation of the situation by looking at the patterns of occurrence. The most straightforward way to do this is graphically. The occurrence of the word *entrance* in the BNC is plotted in figure 3.2. Remember that the mean length of text in the corpus is a little over 28,000 words. The corpus was split into chunks of 28,000 words and the frequency recorded for each chunk. Each line in the figure corresponds to such a chunk. The frequency in each chunk is shown moving from the beginning of the corpus at the leftmost end, and the end of the corpus on the right. The y axis shows the frequency for that chunk. Remember that if the rate of occurrence is constant then we would expect the lines to be of a consistent length hovering about the mean. Based on 2818 occurrences in a corpus of 89.39 million split over 3144 documents we have a mean of 0.9 occurrences per chunk. So we would expect the lines to be consistently around this point on the *y* axis. In fact what we see are per chunk frequencies as high as 37, and frequently running above 5. Clearly then the picture is quite different from what we might naively predict.

In reality , then, the rate occurrence of words is much less constant than predicted by the binomial distribution. The two main reasons for this are a) the probability of occurrence of a word is not in fact independent of the other words that occur in the text, and b) the probability of occurrence of a word actually depends upon other variables such as the identity of the speaker, the present genre, and the topic under discussion. In computational linguistics this latter effect has become know as **burstiness**.

What we have described here is a way to examine the consistency of occurrence of linguistic events, such as words and phrases. This allows us to ask simple questions about the difference between words and phrases. We described above how various writers have suggested that accurate quantitative information cannot be acquired from corpora because phrases are more subject to the contingencies of genre, style and subject than words are. We have seen the extent to which words are subject to these factors. What we want to do now is to look at the occurrence pattern of phrases. If

basically that words have a binomial distribution. A variable has a binomial distribution if its occurrence is the number of successful outcomes of a fixed number of independent trials in each of which it has a equal probability of success. Many techniques in information retrieval and natural language processing assume that words in text have a binomial distribution. According to this the probability of a word occurring *i* times in a text of N words is as follows.

$$P_{Binomial}(i) = \frac{N!}{i!(N-i)!}P(i)(1-P)^{N-i}$$

According to this distribution, the mean occurrence is of course $N * P(i)$, and the variance is $NP(1-P)$.

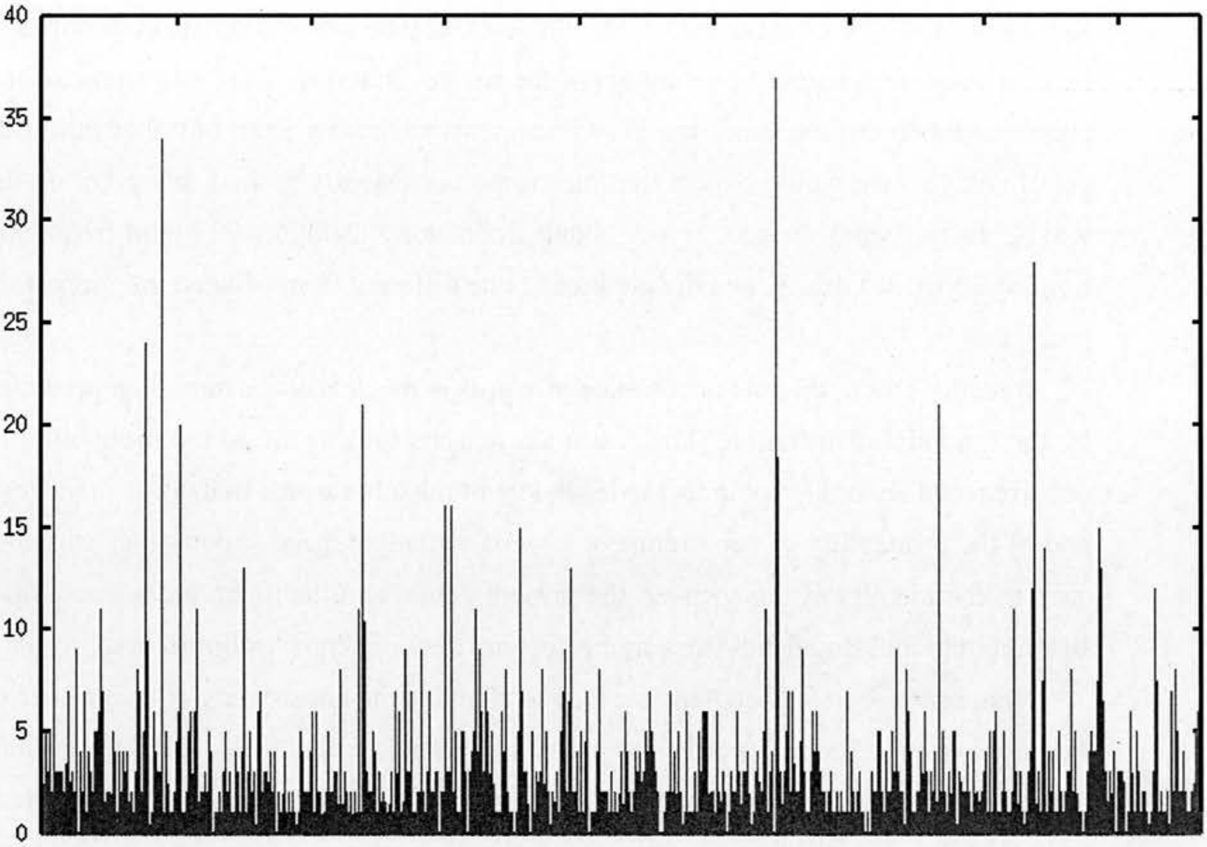


Figure 3.2: Rate of occurrence of the word *entrance*

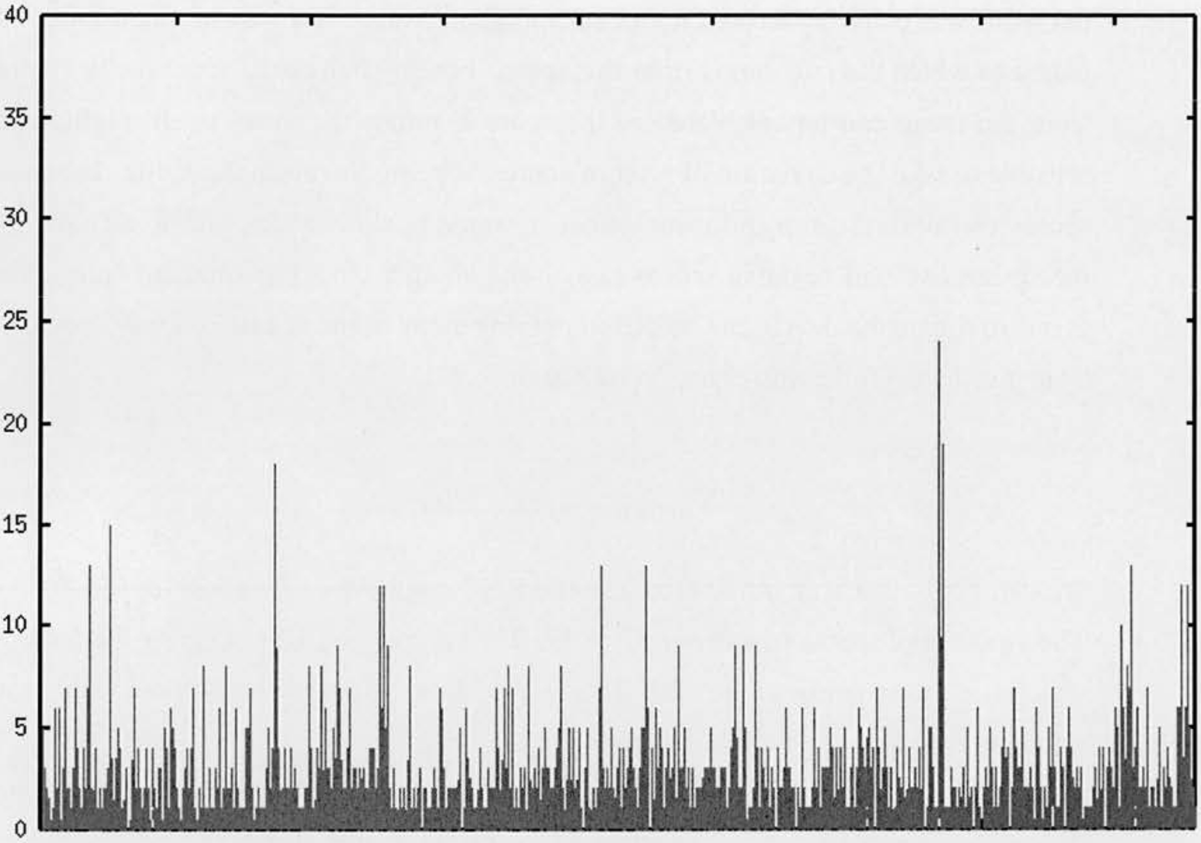


Figure 3.3: Rate of occurrence of the phrase *on the basis of*

these writers are correct, then we will find a much more variable rate of occurrence for phrases than for words; they will (on average) be more bursty than words. Figure 3.3 shows the occurrence rate of the phrase *on the basis of* using the same method as we used for *entrance*. This phrase was chosen from the SAID corpus of idioms, because it has exactly the same overall frequency as the word *entrance*. That is, it too has an mean occurrence of 0.9 per document. And like *entrance* it too shows some deviation from this rate. However, it seems to vary less than the word. Its per text frequency doesn't exceed 25, and goes over 10 on much fewer occasions.

The impression from these graphs, then, is that for this pair of frequency matched items, the rate of occurrence is more constant across portions of the BNC for the phrase than for the word. Striking as this informal picture is however, if we are to conclude anything we are going to need to look at more items. First we are going to need a formal way of quantifying the consistency of occurrence. In statistics, the usual way to describe a set of data like this is in terms of its mean and its variance. The mean is the per document rate of occurrence that we would expect if the occurrence rate of the word was constant across the document. The variance is a way of quantifying the extent to which the rate varies from this mean. For any individual score the deviation from the mean can be calculated as the score X minus the mean μ. It might seem sensible to take the deviation of a set of scores as being the mean deviation. However scores can deviate from the mean either upwards or downwards and if we take the mean, positive and negative scores cancel one another out. The standard solution to this is to square the deviations before combining them, as the square is always positive. This give us the following equation for variance:

$$Variance = \frac{\sum(X-\mu)^2}{N} \tag{3.3}$$

We can apply this straightforwardly to the BNC counts we saw above in the graphs. The variance of the word *entrance* is 4.49. The variance of the phrase *on the basis of* which has the same mean, is 2.89. This tells us that the phrase has a more stable rate of occurrence than the individual word.

Although variance is a standard way of looking at variability of word rates it is not the only approach available. Church and Gale (1994) survey three other scores that are available to us. One score is taken from the field of information retrieval. SparckJones (1972) describes the problem of choosing which index terms for document retrieval are most informative. She proposed a measure that has become known as **inverse document frequency**. This measure can be written in many ways. Defined probabilistically it can be written as follows:

$$IDF_t = -log_2\, P(t) \tag{3.4}$$

where t is the term (word or phrase), and $P(t)$ is the probability of seeing that term in a given document, which can be calculated as

$$P(t) = \frac{n_t}{N} \tag{3.5}$$

where n_t is the number of documents containing the phrase t and N is the total number of documents. IDF is intended as a measure of how characteristic a particular term is of the documents that contain it. In information retrieval a term that has a high IDF is assumed to be specific to certain subject matters and types of document. As such terms that have high IDF can be said to be occur inconsistently across documents or to be bursty. It is therefore considered an appropriate measure of variability in frequency of occurrence.

Another measure that Church and Gale (1994) suggest is the entropy of the frequency of occurrence over the documents. Entropy is a measure of the uncertainty of a random variable. The variable we are interested in here is the frequency of occurrence of a word or phrase t. We measure the entropy of this variable over all the values that it takes over all the segments of 28000 words. It is calculated as follows, where x is a count value from the set of observed count values X:

$$H(X_t) \quad = \quad -\sum_{x \in X} p(x)\ log_2 p(x) \tag{3.6}$$

The probability of a value x is calculated as the number of times a segment has that value divided by the total number of values. A high entropy indicates inconstancy in the rate of occurrence of a word across segments, and a low score indicates a stable count.

The final measure that Church and Gale (1994) discuss is that described in Katz (1996). This is based on the observation that "when a concept named or expressed by a content word is *topical* for the document....then the content word is characterised by *multiple* and then often *bursty* occurrence" (p.18). The point is that when words or phrases have single occurrences in texts, this tends to be because the word is not typical of the topic. When a word or phrase is representative of a topic it will occur multiple times. Accordingly, he suggests, a measure of the burstiness of a word should not only reflect the frequency of occurrence relative to the number of documents, as in the mean, but that this should be seen relative to the percentage of documents in which the term occurs at least once. The measure that he proposes involves calculating the mean frequency of a term, but this is then divided by the probability of seeing that term in a document at least once. This can be written as follows:

$$Burstiness_t = \frac{\bar{t}}{P(x \geq 1)} \tag{3.7}$$

What we have described are four ways of evaluating the burstiness of words and phrases. The next section will described an experiment to compare the burtiness of words and phrases.

3.2 Experiment one

The aim of this experiment is to explore whether, as various writers have suggested, MWEs have more bursty occurrence patterns than words. We will do this by selecting a representative set of phrases from an established dictionary of idioms, matching them with individual words of identical occurrence frequencies that occur in popular lexical resources, quantifying the variability in appearance rate across a corpus using the four measures described above, and then looking for differences in the profiles of the set of phrases from that of the set of words.

3.2.1 Materials

The materials for this experiment consist of lists of the rate of occurrence of a set of 180 phrases, and a set of 180 frequency matched individual words, measured across segments of the British National Corpus. The phrases were selected from the SAID corpus of Idioms. In order to get a balanced representative sample of these phrases, they were selected in the following fashion. Phrases of between two and seven words in length that were found in the BNC were extracted from the collection and put into groups by length. These groups ranged from a set of 73 idioms of seven words in length to a set of 4082 idioms of two word length. Counts for each each group of phrases were then extracted from the BNC. [5] These counts were then log-transformed and the sets of phrases ranked, and split into three categories of high, medium or low frequency for the word length category based on an equal split of the log-transformed frequency range.[6] 30 phrases were then selected for each of the six sets of phrases. For the sets of two to six word phrases, 10 phrases were randomly selected from each frequency band. For the seven-word phrases, which had a particularly thinly spiked frequency distribution, 4 phrases were randomly selected from the high frequency phrases, 16

[5] It should be noted that not all idioms in the SAID group are found verbatim in the SAID corpus. There could be a number of reasons for this. Firstly, as we discussed in section 3.1, it could be that they don't exist. However in many cases it is likely that they do exist, but in some alternative form with lexical or morphological variants. It would have been possible to work on detecting variants. There are, however, two good reasons for not doing this. Firstly, any such attempt would inevitably skew the counts, as some variation forms would be considered and others not, thereby reducing the validity of the counts. Secondly, one would have to make biasing decisions as to what constitutes a variant on the original form and what a different item. The decision was therefore taken for the purposes of this experiment to only count a word sequence as an example of an idiom if it exactly matched the dictionary entry.

[6] Information about frequency is often log-transformed in the cognitive sciences. It is assumed that log frequencies better fit human perception of frequencies of perceived events as they seem to fit frequency effects across modalities better than raw frequencies.

from the medium frequency phrases and 10 from the low.

It is important to note that we count occurrences by simply looking for the form of interest in the corpus. This method of extracting counts is arguably problematic in that each MWE form will have been included in the dictionary with a particular usage in mind and our counts will very likely include occurrences that do not fit this. For example the list contains the phrase *and all,* which we presume is the abbreviation of *and all that* as seen in sentences such as *There was a new independent nation in the world, and its ambassador, striped waistcoat and all, was presenting his credentials to King George III.* It might be argued that that our counts do not strictly reflect the occurrence of the MWE but only of the surface form. However, we take this approach for both practical and theoretical reasons. Firstly manually inspecting each of the many thousands of occurrences of all the MWEs would not be feasible. Secondly the same problem that we are seeing with MWEs arises when dealing with individual words, which can also have multiple meanings. The frequency lists of Francis and Kucera (1982) and Leech *et al.* (2001), the former of which has been used extensively in psycholinguistic experiments, report word frequency on a per-form rather than per-sense basis. For us to make this distinction would make the frequencies incomparable with these resources and would introduce an aspect of individual semantic judgement that would make any result open to question. Furthermore since false positives will occur for both the phrases and the single words, it is reasonable to assume that any effect will be balanced across the two conditions and will not be responsible for the presence or absence of any significance differences between the distributions when we come to make our comparison.

Once this list of phrases had been constructed, we moved on to constructing a comparable list of words. This was done as follows. BNC counts were acquired for all items listed in Francis and Kucera (1982). This word list was chosen for use as frequencies from this source are frequently used in psycholinguistic experiments and so are considered to be amenable to valid frequency analysis. These counts were then used to randomly select for each of the 180 phrases a word that had an identical frequency (i.e. there was a total of 180 words). Where no word had an exactly matching frequency to a particular phrase, a word with a very nearby frequency in the set was selected, ensuring that the frequencies were balanced so that the total frequencies was the same for the phrases and the words for each of the sets of different phrase lengths. This gave us a set of 180 words which matched the set of phrases for frequency.

The next step was to extract the frequency profiles for the set of words and for the set of phrases. This was done by splitting the corpus sequentially into segments and then counting how many times each word or phrase occurred in that segment. The first question then is how to split the corpus. The written component of the BNC consists of 89.39M words split over 3144 texts, giving an average text length of just over 28,000 words. For this reason we chose to split the corpus into 3144 equal chunks of 28,000 words.

3.2.2 Method

Our materials are per-segment counts for a set of 180 MWEs and a set of 180 matching words over the whole of the BNC. The aim is to compare the stability of distribution of the two sets over the corpus. Because the words and phrases are matched for frequency as well as being randomly selected and representative of the frequency range of the phrase, by comparing their behaviour we can test the hypothesis that phrases are more bursty than words. What we want to do then is to quantify the variability of each of the phrases and each of the words. We do this using all four of the methods outlined above. These are Variance, IDF, entropy and Katz's burstiness score. Because the sets are balanced for term frequency, the number of different documents in which the words and phrases occur in is also highly indicative of their relative stability so we also report document frequency.

3.2.3 Results

The output of this procedure is four measures of the variability of each of the set of words, and each of the set of phrases. The frequencies and scores for all 180 items can be see in appendix A. The mean instability measures can be seen in table 3.1

Our first analysis is a comparison of variance over all lengths and frequency bands. For variance, IDF and Burstiness, the words are found to be marginally less stable that the phrases. For the entropy measure they are found to be marginally more so. We next wanted to test whether this difference was significant. As these scores are based on word frequencies which we know are not normally distributed, we need to be careful in our choice of significance test. In order to see whether parametric statistics would be appropriate, we performed a Kolmogorov-Smirnov and a Wilks-Shapiro test on each set of words and each set of phrases of each length for each of the measures of stability. We found that the scores were not normally distributed for all lengths for both phrases for the variation, IDF and entropy scores at a value of $p < 0.001$. For the

	Idiom mean				Matching word mean			
	Var	IDF	Burst	Ent	Var	IDF	Burst	Ent
2-grams	.439	5.282	1.275	.555	1.759	5.668	**1.742**	.498
3-grams	.554	5.250	1.283	.524	1.858	5.638	**1.760**	.654
4-grams	.501	5.685	1.313	.502	1.548	5.922	1.633	.458
5-grams	.025	7.743	1.161	.086	.035	7.969	**1.376**	.080
6-grams	.007	8.488	1.042	.047	.014	8.861	**1.476**	.041
7-grams	.003	9.352	1.048	.0215	.003	9.381	1.073	.021
ALL	.2547	6.967	1.187	.289	.870	7.239	**1.509**	.263

Table 3.1: Instability of frequency distribution for SAID idioms

entropy scores there was a less pronounced by still strong evidence of a non-normal distribution at a level of at least $p < 0.05$ in the majority of cases. We therefore opted to use a non-parametric test to look for significant difference - the Mann Whitney test. [7]. There was found to be a non-significant difference for variance (U = 14846; p = .170), entropy (U = 15505; p = .481) and IDF (U = 15246; p = .333), with the occurrence rate of the words being found to be significantly less stable than the phrases according to the Burstiness measure of Katz (1996) (U = 11178.5; $p < .001$).

These are the overall results. It is also interesting to look at the differences for the various phrase lengths. We see the same pattern of the words being less stable over all sequence lengths across all the measures with the exception of the entropy measure. Again Mann-Whitney tests were performed to measure the significance of these differences. All groups of words or phrases that were found to have a burstiness scores that was higher than their matched set are shown in bold in table 3.1. The words were found to have a significantly less stable rate of occurrence at $p < 0.005$ level of significant according the Burstiness measure for two, three five, and six word sequences. No other significant differences were found.

[7] The Mann-Whitney test measures the significance of any difference between two groups by comparing the ranks that the groups achieve over all the data. The U statistic is the difference between the actual rank of a groups and the maximum ranks it could have got. It is calculated as follows

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2 \tag{3.8}$$

The score of the more highly ranked group (the lower of the of the two U values) is reported.

3.3 Experiment two

Experiment one discovered that idioms of various lengths taken from an existing lexical resource are at least as stable in their rate of occurrence as words of matching frequency. This next experiment looks at whether this pattern is true only for established dictionary-confirmed idioms, or also holds for arbitrary substrings of the same lengths.

3.3.1 Materials

The materials for this experiment were the rate of occurrence of a set of 180 phrases, and a set of 180 frequency matched individual words, measured across segments of the British National Corpus. The phrases were selected as follows. Firstly all substrings that occur more than once in the corpus were extracted, and sorted according to length and frequency. These were then split into sets according to their length in words. The frequencies were log-transformed and the six sets of phrases of between two and seven words in length were sorted, and split into high medium and low frequency bands. 30 phrases were then randomly selected from each of the six sets of phrases. For each of the sets of phrases of from two to six words in length, 10 sequences were selected from each frequency band. As in experiment one, the top of the frequency range for seven-words strings is very sparsely populated and so again these were handled slightly differently with 4 sequences being taken from the high frequency band, 16 from the medium and 10 from the low.

These phrases were then matched with single words on the basis of frequency, and frequency profiles extracted for all sequences and matching words in exactly the same way as for experiment one.

3.3.2 Method

Our materials are per segment counts for a set of 180 randomly selected subsequences of various lengths and frequencies and a set of 180 frequency matched words. Our aim is to compare the stability of the distributions of the two sets over the corpus, in order to test the often made claim that phrases are more bursty than words. This was done in the same way as for experiment one, with all four of the measures outlined above being calculated for all words and phrases.

	Multiword mean				Matching word mean			
	Var	IDF	Burst	Ent	Var	IDF	Burst	Ent
2-grams	6.579	4.322	2.470	1.148	15.565	4.593	2.940	1.091
3-grams	.587	5.942	1.362	.513	2.150	6.168	1.735	.464
4-grams	.125	6.304	1.165	.309	.370	6.690	1.578	.237
5-grams	.3247	6.764	4.011	.208	.209	6.780	1.532	.201
6-grams	.086	7.445	2.207	.133	.056	7.147	1.334	.149
7-grams	.146	9.323	4.642	.428	.079	8.205	1.405	.054
ALL	1.308	6.683	2.632	.457	3.072	6.597	1.756	.374

Table 3.2: Instability of frequency distribution for arbitrary multiword strings

3.3.3 Results

The scores for all phrases and words can be found in appendix B. The mean scores can be seen in table 3.2.

The mean overall scores for this experiment show a higher variance for the words, but a marginally lower variability according the IDF, burstiness and entropy measures. A Mann-Whitney test tells us that the difference is not significant for either of the variance (U = 15752, , p = .650), IDF (U = 15813.5, p = .695), burstiness (U = 14771, p = .145) or entropy measures (U = 15710.5, p = .620).

There is no significant difference between the groups over all phrases. The scores for the different phrases also give a more variable picture than for experiment one. The multiword scores have a lower burstiness for the three shorter groups of phrases, at a level that is significant for phrases of length four. However they have a higher burstiness for the longer strings, at a level that is signicant for 5 and 7 word strings. The multiword strings have a higher entropy for all except the 6 word strings. The shorter multiword strings have a lower IDF than the words for all phrases up to length 5, and a lower variance for all phrases up to length 4. All groups of words or phrases that were found to have a burstiness scores that was higher than their matched set are shown in bold in table 3.2. There is no overall significant difference between the words and the phrases for any of the scores, and no consistent difference across the groups of different lengths.

3.4 Discussion of experiments one and two

One argument that has been made for limiting the role of corpora in studying and building multiword lexicons is that MWEs are simply too sporadic in their occurrence to be usefully studied using such text collections. Experiments one and two have examined this claim and have shown that far from being unstable, multiword substrings of between two and seven words of high, medium and low frequency have a rate of occurrence that is at least as stable as the rate of occurrence of equivalent single words. This was shown to be the case for both idioms taken from a dictionary and for substrings selected randomly from other phrases of their length and frequency.

The factors that produce burstiness are specificity of a word or phrase to a particular topic, genre, register or period of time. The BNC contains a wide range of topics, genres and registers, and it would seem that MWEs are no more specific to any of them than words of equivalent frequency. One interesting outcome from our analysis was the discovery that the SAID idioms were more stable than their matching words (and according to the burstiness measure of Katz (1996) were significantly more so overall) while there was no such tendency for the arbitrary multiword strings. Looking over the list this is hardly surprising. A large number of units of formulaic speech serve a familiar rhetorical function (e.g. *if you know what I mean*), or are aphoristic units of traditional wisdom (e.g. *the best things in life are free*) which are not in any way tied to a particular topic.

Another cause of inconsistency of occurrence is the changes in language that occur over time, with new words and phrases entering the language and other leaving. The BNC is, of course, not designed to be a diachronic corpus and so it cannot tell us about the stability of items over long stretches of time. However the corpus is sampled from a reasonably broad window. 91% of the words in the BNC were originally published between 1984 and 1991, and the remainder no earlier that 1960. Of course the multiword lexicon is subject to change over time. However so is the single word vocabulary of the language. These experiments confirm that over the small window in time covered by the BNC, at least, the multiword lexicon is no more dynamic than the single word lexicon. Further research would be necessary to assess whether this is true over longer time spans. however.

Before moving onto the next set of experiments, it will be interesting to point out a connection between our results here and the question of rarity we considered in section 3.1. Bernoulli's theorem (also referred to as "the weak law of large numbers" or more popularly "the law of averages") tells us that if one has a large enough sample then the probability of an event is likely to be very close to its correct probability. More formally, this states that for any positive number ε, no matter how small there exists an n such that the difference between the observed mean for a sample of that size and the actual expected value μ is guaranteed to be less than n. This is conventionally written as follows:

$$lim_{n \to \infty} P(|\bar{X_n} - \mu| < \varepsilon) \quad = \quad 1 \tag{3.9}$$

As n gets bigger, so $\bar{X_n}$ converges towards μ. What this tells us for our purposes is that for any phrase there is a sample (corpus) size n big enough that we will be able to estimate its rate of occurrence within a reasonable ε of μ. So another way to think of the examination of burstiness is as a test of whether the BNC represents an adequately sized corpus for obtaining reliable counts. Our answer to this question based on the experiments we have described here is that it seems to be at least as adequate a size for describing the distribution of MWEs as it is for individual words.

3.5 Human processing and phrase frequency

In the previous section we saw that stable counts for phrases can be acquired from corpora. In this section we will attempt to further confirm the validity of these counts by looking at whether they are reflected in speakers' knowledge of the language.

We saw in section 2.3.1.1 that frequency has long been recognised as a significant factor in the representation and processing of individual words. We also saw more recent evidence that it is a significant factor for bigrams or pairs of words. We might expect, then, to find an effect of frequency for longer phrases. However there has been no conclusive work looking at the processing of recurrent phrases. This section describes an attempt to fill this gap.

Two experiments are described in this section. They both took the form of a self-paced reading experiment, similar to the design seen in MacDonald (1993). A series of sentences were presented to subjects on a computer screen in a series of chunks of between and 3 and 8 words. Subjects were told that when they had read a particular chunk they needed to press the space bar to move onto the next chunk. The computer recorded the time between the initial display of each chunk and the subject's pressing of the space bar. This was taken to be the reading time for that chunk. The materials were designed so that the sentences each contained a target chunk. This was either a) a frequent word sequence of between 4 and seven words, or b) an infrequent chunk that was matched on key dimensions (the two experiments vary in how the matching was done. This will be described in more detail below).

The hypothesis here is that frequent chunks should be read more quickly than their matching low frequency chunks once other dimensions have been factored out. For individual words it has been shown that more familiar words are identified faster. We are expecting a similar effect for more familiar sequences. In order to facilitate examination of this effect of familiarity, identification was made more difficult for participants by displaying the words in an unfamiliar font.

The idea that reading time reflects processing effort has been crucial to a great deal of sentence processing research. The idea is that increased processing time reflects an increase in processing effort. This has been used to claim evidence for a wide range of phenomena such as garden path syntactic structures, prepositional phrase attachment preferences and various strategies for the resolution of lexical ambiguities. The claim that reduced processing time reflects processing advantages for phrases of particular kinds is not a novel one either. This idea has been explored using both lexical decision (Swinney and Cutler 1979) and reading time (Titone and Connine 1999) tasks.

One thing that distinguishes this work from previous work is its use of corpus materials. We discussed above how an idealised view of the multiword lexicon often dominated previous research. For example we described above how researchers regard dictionaries or intuition as a more reliable source of evidence than linguistic use when discussing occurrence. And we saw when discussing experimental work on idioms, that researchers tend to select the phrases to be explored from dictionaries without regard for actual usage. Such expressions are then often situated in fabricated sentences. Both of these factors then result in stimuli that are far removed from real language use. Moon (1998:32,36) discussing psycholinguistic research looking at the processing of fixed expressions, observes the following:

> "It has to be said that from a corpus linguistics perspective, some of the experiments are suspect. Much of the work elicits responses on the basis of either decontextualized strings or fabricated texts and contexts...Until researchers work with authentically occuring texts, it is very difficult to see whether the various hypotheses accurately reflect what actually goes on during interpretation and processing in real language situations."

Now it is common in psycholinguistic research to make use of invented stimuli, and there are certain merits in doing so. Obviously constructing the data oneself allows the experimenter greater control over the stimuli at the point of its creation. However, as Moon notes, inventing stimuli greatly reduces the strength of one's claim to be discovering facts about real language processing. Particularly when one is examining usage-oriented hypotheses such as frequency effects and collocation, it is very much preferable to employ real naturally occurring data. This may introduce variance into the results, but as long as one is not explicitly manipulating this variance, and one's analysis ensures that any effect is not produced by any factors that are known to affect performance, one is able to assume that the variance is random and will be compensated for by one's statistical analysis as all other random factors are. This experiment will employ as stimuli phrases and sentences that have been obtained from a corpus.

The two experiments in this section vary in how they match high frequency phrases with their low frequency counterparts. Experiment three is more conservative in how it finds matching phrases. Frequent phrases were matched with infrequent phrases that only differed by a single word where that single word was matched for category and frequency. In experiment four frequent phrases will be paired with infrequent phrases by identifying items that have a matching syntactic form.

3.6 Experiment three

3.6.1 Subjects

30 students of Psychology from Stanford University participated in the experiment. All were native speakers of English. They were each paid $5 for their participation.

3.6.2 Materials

The experiment materials consisted of 24 sentences, 12 of which contained a frequent phrase, and 12 of which contained a matched infrequent phrase. Each frequent phrase (e.g. *a state of emergency*) had a matching infrequent phrase (e.g. *a state of pregnancy*) that differed only in having a different final word (in all cases a noun) of similar frequency. These were assigned to 2 groups of 12 sentences, each of which contained 6 frequent-phrase-containing and 6 infrequent-phrase-containing sentences. These two sets were presented to different groups of 15 subjects who were randomly assigned. Matched pairs were always put into different groups, to prevent the confounding familiarity effect that might arise if a subject was presented with two similar chunks. This splitting of stimuli and conditions (frequent and infrequent) between subjects gave a counterbalanced design with both groups of subjects and both sets of stimuli participating in both conditions.

The target phrases in our materials were all of between 4 and 7 words. These were in two groups of frequent (mean = 176, minimum = 44, maximum = 441) and infrequent (mean = 1, minimum = 1, maximum = 5) phrases, with a minimum of one quarter of the log frequency range of phrases of 4-7 words between any frequent item and its matched infrequent item [8]. All selected phrases and their matches can be seen in table 3.3. Again the counts reflect all occurrences of each form with no distinction being made on the basis of meaning. The literature has reported that various factors can effect reading time. Any variance in these factors were either held constant or minimized and recorded so as to be factored into our analyses. There were as as follows.

- **Syntactic ambiguity** - It has been reported that the reading time for words that are ambiguous between multiple syntactic categories can produce an increased reading time. Two factors have been shown to effect this in experiments. One is the bias of the context (MacDonald 1993), and the other the relative frequency with which the given word form occurs with different syntactic categories (Trueswell 1995). These factors were therefore controlled here. All final nouns could occur with only a single category given the preceding word, and all the final nouns in the infrequent condition occurred in a nominal form in more that 75% of appearances in the BNC (occuring in nominal form in a mean of 87% of cases as compared with 85% of cases in the frequent group). It was further checked that over the BNC no phrases were found that were the same but ended with a different syntactic category, meaning that all contexts were strongly biased to a noun interpretation.

- **Word frequency** - We saw in section 2.3.1.1 that the time taken to identify a word is inversely correlated with its log frequency. We might therefore expect reading time to vary with the frequency of the words found in our materials. It

[8] A popular strategy when manipulating frequency as a variable in experimental stimuli is to log transform the frequencies, split the items into bands by equally dividing the log frequency range, and then pick frequent items from the top band(s) and infrequent items from the lower band(s). However due to the more sparsely populated range for phrases (giving, for example, just 4 items in the top third of the range for 7 word strings, all of which contain proper nouns), and the fact that we are working with phrases of different lengths and types, taking such an approach over the set of all phrases was not possible here. All frequent items are taken from the top third of the frequency range for valid phrases of the same phrase type (with the same part of speech tag sequence) in the corpus, and all matched infrequent phrases from the bottom third of the range.

was ensured that the frequency of the final word varied minimally across conditions. The mean natural logarithm of the frequency over the BNC for the frequent phrases was 12.53, and for the infrequent was 12.38. These frequencies were recorded and all variance will be factored into our analysis of the results.

- **Non-compositionality** - Various writers, among them Swinney and Cutler (1979), Gibbs *et al.* (1989) and Titone and Connine (1999) have provided evidence of a reduced reading time for phrases that were independently judged to have a non-compositional meaning. It was necessary therefore to ensure that this could not be a factor in our experiment. All phrases were examined by two native speakers, who were asked to indicate "yes", "no" or "don't know" to the question "are all the words in this phrase contributing a meaning that they can have when seen outside of the phrase". None of the phrases were judged to be non-compositional. One annotator answered "yes" for all phrases. The second annotator answered "don't know" for two phrases ("a matter of form" and "a bit of speed"). As both of these were in the infrequent group of phrases, any bias would favour the null hypothesis.

- **Transitional Probability** - McDonald and Shillcock (2003) reported that the probability of seeing a word given its preceding word inversely correlated with the duration of a readers gaze in an eyetracking experiment. We therefore needed to ensure that this could not produce any effect found in our data. We calculated the probability of seeing each word in the phrase given the preceding word in its sentence using the SRI language modelling toolkit (Stolcke 2002) and the written component of the BNC. As in McDonald and Shillcock (2003) the model was smoothed using the toolkit's Good-Turing smoothing option with the default discounting range for bigrams of between 1 and 7 occurrences. The probabilities were converted to natural logarithms. The mean log transitional probability for the frequent phrases was -4.2638 and for the infrequent was -4.6227. These probabilities were recorded and will be considered in our analysis.

Having selected our phrases, sentences containing these phrases were randomly selected from the British National Corpus to use as our stimuli. All the stimuli sentences can be seen in appendix C. In half of the cases corpus sentences could be used directly. In the remaining sentences it was necessary to alter the sentences by removing or substituting material in order to remove disfluencies, [9], to satisfy the criteria that the target

[9] Group two sentence six was edited from *There were certain problems common to all nineteen*

phrases should not be at the beginning or ending of the sentences [10],or to dispose of redundant material at the end of the sentence [11]. There were two additional factors that we need to consider in using these sentences. These are as follows:

- **Sentence Position** It has been claimed that if all other factors are held constant then reading time decreases as more context becomes available (Keller 2004), meaning that words or phrases that occur later in a sentence might be read faster. For this reason the position of the phrase in the sentence was examined. The phrases in the frequent chunks were found to be preceded by a mean of 7 .0 words while those in the infrequent condition were found to be preceded by a mean of 7.6. Any bias would therefore appear to favour the null hypothesis. These were recorded and will be factored into our analysis.
- **Plausibility** It has been reported that reading time inversely correlates with the plausibility of sentences. In order to check that the overall plausibility of our sentences was not producing any differences in reading time, it was necessary to obtain judgements from subjects as to the plausibility of the stimuli sentences. Seven native speakers of English were asked to judge the plausibility of all the sentences. This was done using a magnitude estimation task. Magnitude estimation has been successfully used to gather plausibility judgements in previous work (e.g. Lapata *et al.*, 1999). The participants were first asked to assign an arbitrary number to a modulus item, and then to assess the plausibility of each of the stimulus sentence by assigning a score relative to the modulus. The judgment for each combination is then taken to be the ratio of this score and the modulus item. A single score for each item was then obtained by taking the mean of the logs of these ratio scores over all subjects. Any effect of these scores will be considered in the analysis of results.

3.6.3 Method

A further 3 practice sentences and 12 filler sentences were added to each of the sets of 12 sentences presented to subjects. All materials were presented on a computer

denominational colleges: the first was the quality of men who entered the ministry and from whom students were recruited.

[10] Group one sentence ten was edited from *I was surprised that he should sound so definite: it was usually I who pinned down occasions with that sort of fact.*

[11] Group one sentence one was edited from *When the firemen went on strike in 1977 a state of emergency was called by the Callaghan government and the army was employed in a breaking capacity with the use of green goddess fighting vehicles.*

Frequent Phrases	Count	Infrequent Phrases	Count
the right place at the right time	62	the right place at the right price	1
in the early hours of the morning	86	in the early hours of the afternoon	1
in the heart of the city	44	in the heart of the woods	1
at the time of writing	319	at the time of review	1
there is no such thing	150	there is no such force	1
only a matter of time	137	only a matter of form	1
as a matter of fact	327	as a matter of sympathy	1
a state of emergency	140	a state of pregnancy	1
the course of time	79	the course of use	1
that sort of thing	441	that sort of fact	1
the quality of life	260	the quality of men	1
a bit of luck	68	a bit of speed	5

Table 3.3: Phrases and Frequencies for Experiment Three

monitor, and response times recorded using the DMDX software (Forster and Forster 2003). After reading instructions, the subjects read the three practice sentences before beginning the experiment. The 2 sets of 12 sentences were then presented to the two groups of subjects. The order of presentation was randomized for each subject. The sentences were broken up into chunks of between 3 and 8 words (as seen in appendix C). Subjects were instructed to read each word sequence and then to press the space bar to move onto the next. A pause screen was presented between each sentence and the subjects were required to press the spacebar to proceed. In order to require participants to pay attention to the sentences, after 12 of the 24 sentences they were asked yes-no questions to which they indicated an answer using the keyboard. The text was displayed using the Vivaldi font in cyan on a white background.

3.6.4 Results

The outcome of this experiment is reading times in milliseconds for each word sequence read. The mean reading time for each item (averaged over subjects) and by condition for each subject (averaged over items) are shown in appendix E. As is standard for self-paced reading experiments, before analysing the data, we want to normalise for phrase length. Although the phrases pairs are all matched for the number of words they contain, in a number of cases there is a difference between the length of the frequent phrase and its matching infrequent phrase in characters. The most straightforward way to do this would be to divide by character length, thereby giving a per-character reading time. However, as Ferreira and Clifton (1986) point out, while it is reasonable to assume that there is a linear relationship between reading time and length, simply dividing by length in this way assumes that the recorded time when the number of characters is zero would be zero, when in fact the time taken to press the button means there is a lower bound on reading time of more that zero. It has therefore become standard to accommodate length variation in the following fashion. A linear regression analysis is performed to obtain an estimate of the zero intercept α and the slope of the function β which describes the relationship between length and reading time. This can then be used to calculate an expected reading time y for each phrase given its length x as shown in the following equation:

$$E(y|x) = \alpha + \beta x \tag{3.10}$$

This estimated reading time value can then be subtracted from the observed time in order to partial out the effect of length, and the residual can be subjected to analysis.

We performed a regression over all the reading times for all the sentence segments for each subject. The parameters for each subject were used to calculate an expected reading time for each target sequence. The mean r value for the correlation over all subjects for all chunks in group one was .552 (df = 52; p < . 00001), and the same in group two was .475 (df = 45; p < .001). The mean r^2 values are .319 and .239, telling us that the regression accounts for a mean of 32% and 24% of the variance in the data for the two groups. This expected reading time value was subtracted from the observed values and the residuals subjected to analysis.

This experiment has a counterbalanced repeated measures design. There were two conditions in two sets of data presented to two sets of subjects, with the subjects and the sets of data balanced over the two conditions. It was ensured that no subject saw both a frequent phrase and its matched infrequent item, thereby avoiding the effect that reading one phrase would have on the reading of its matching phrase. Crucially though, because the subjects and the data were balance across the data we can still perform a more powerful within-subjects analysis (see Pollatsek and Well, 2001 for a good discussion of the assumptions and advantages of counterbalanced designs).

The mean reading times for the two conditions over the two groups of subjects can

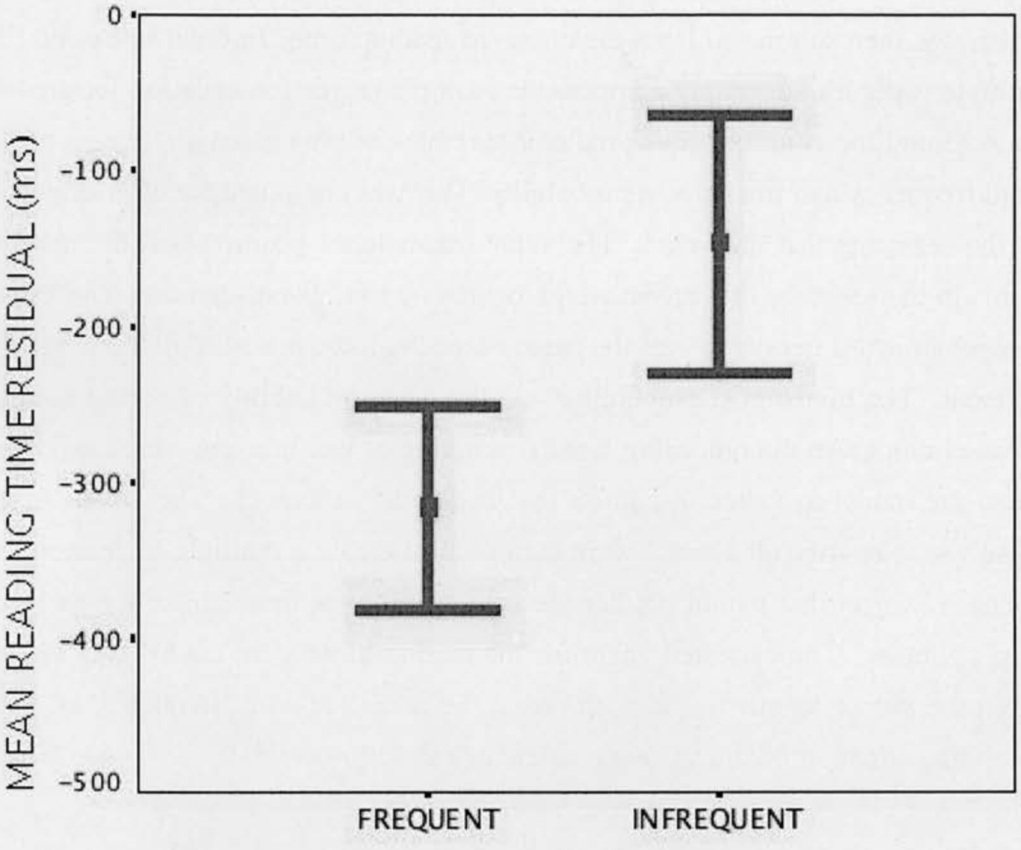


Figure 3.4: Mean experiment three reading time residuals (adjusted for length)

be seen in appendix E in table E.2 and the mean time for each item averaged over subjects can be seen in table E.1. These means are also represented in figure 3.4. We can see that in both groups the frequent phrases have a lower residual indicating a lower reading time. A repeated measures ANOVA was performed on the data, with frequency as a within-subjects variable. There was a significant effect of frequency in both a by-subjects (F1(1,29) = 16.310; $p < .001$) and a by-items analysis (F2(1,11) = 5.216; $p < .05$).

We discussed above how there were various factors that have been argued to affect reading time that we need to consider as potential confounds. These are the position in the sentence that the chunks occur, the frequency of the component and the within-chunk transition probability. We held these as constant as possible using various heuristics. However, because some small variance inevitably remains across conditions it will be useful to reanalyse the data factoring all of these variables out.

These additional variables can be integrated into the analysis in much the same way as word length was. That was accomplished by computing a simple regression model to predict the expected reading time values from the character length of the phrases, which was then subtracted from the observed reading time. In order to extend this to multiple variables we simply compute the multiple regression equation for predicting the reading time from our four predictor variables of phrase length, phrase position, word frequency and transitional probability. This was computed for each subject over all the segments that they read. The value for sentence position was the number of words in the sentence that preceded the beginning of a given segment. The value for component word frequency was the mean of the log frequencies of all the words in the segment. The transitional probability was the mean probability of seeing each word in the chunk given the preceding word (beginning of line markers were introduced to allow the model to factor in chunks that began the sentence). The values of all of these variables over all chunks were then used to create a multiple regression model for each subject that would predict the reading time for that subject for each of the target phrases. The expected value for the reading time y for each target sequence, given the phrase length x_i, the position x_j, the mean log word frequency x_k, and the mean transitional probability x_l was calculated as follows:

$$E(y|x_i,x_j,x_k,x_l) = \alpha+\beta_1 x_i+\beta_2 x_j+\beta_3 x_k+\beta_4 x_l \tag{3.11}$$

The regression had a mean r value of .66 (df = 49; p < .0001) over group one and .576 (df = 42; p < .0001) over group two. It also obtained an r^2 value of .448 for group one and .345 for group two, telling us that the new variables in the regression account for an additional 13% and 11% of the variance in the data for the two groups.

Adjusted reading times were then calculated by again subtracting the predicted reading time from the observed reading time and taking the residuals. The mean residuals can be seen in tables E.1 and E.2 and are represented in figure 3.5. Again the frequent chunks have a lower reading time than the infrequent in both groups. A repeated measures ANOVA was performed on the data, with frequency as a within-subjects variable. There was a significant effect of frequency in both a by-subjects (F1(1,29) = 12.667; p < .001) and a by-items analysis (F2(1,11) = 5.489; p < .05). The introduction of these new covariates into our analysis reduces the by-subjects effect, and slightly increases the by-items effect. It can thus be concluded that the variables cannot account for the difference in reading time across conditions. The significant variance appears to be due to the manipulated factor of frequency.

The last factor that we want to consider is plausibility. We want to check that dif-

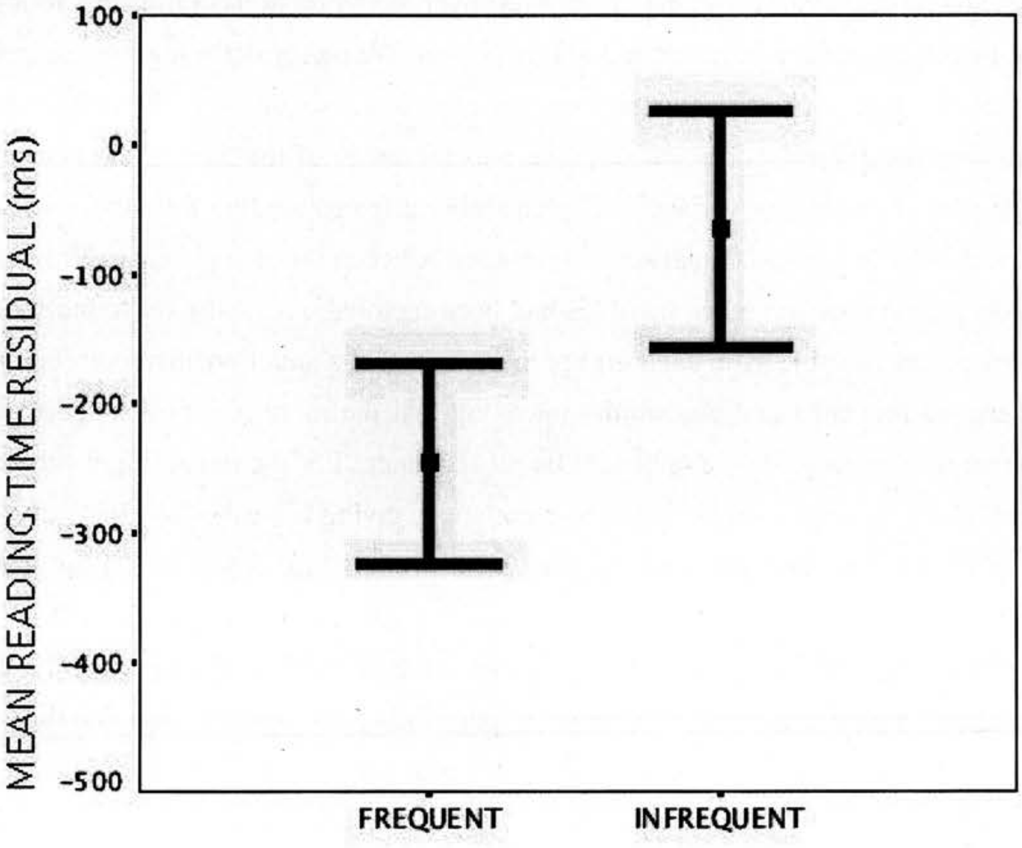


Figure 3.5: Mean experiment three reading time residuals (adjusted for length, position, frequency and transitional probability)

ferences in sentence plausibility across conditions are not causing our effect. We have plausibility judgements for each sentence rather than for each chunk read (plausibility is conventionally assessed at the sentence level, and we concluded that it would not be meaningful to ask subjects to judge the plausibility of each segment in isolation). As a consequence we cannot include them in the regression model along with the other factors. We therefore perform an independent regression analysis looking for a relationship between plausibility and reading time. We did this by performing a multiple regression for each subject and then measuring the reliability of any effects over the subjects by performing a single group *t*-test over the regression coefficients, following the method described in Lorch and Myers (1990). We performed a multiple regression for each subject with their reading time for each target sequence as the outcome variable and the covariates discussed above plus the mean of the logs of the plausibility judgments as predictor variables. If plausibility affected reading time then we would expect to find a significant negative correlation between the plausibility judgments and reading time once the other variables had been factored out. What we found was that for 6 of the 15 subjects in each group, there is in fact a small positive correlation between reading time and plausibility, meaning that the more plausible a sentence the longer it takes to read its component target sequence. For the remaining 9 subjects in each group there is a small negative correlation, giving a mean *r* of -.084 (df = 6; p = .97) for group one, and small negative mean *r* of -.10 (df = 6; p = .81) for group 2. A single group *t*-test for group one gives a non-significant *t* value of -1.431 (df=14, p =.171), and for group two a non-significant value of -.902 (df=14; p =-.382). There is then no significant correlation between plausibility and reading time and thus there is no evidence that sentence plausibility is producing the difference in reading time observed between our two conditions.

3.7 Experiment four

Experiment three showed a faster reading time for frequent over infrequent phrases which were the same except for the final word. Experiment four looks for a similar effect using somewhat different stimuli. Rather than pairing frequent sequences with infrequent sequences that were lexically identical except for the final word, this experiment paired frequent sequences with infrequent sequences that were identical in terms of their length and syntactic form. Our hypothesis was that we would find a faster reading time for the frequent strings when all other factors had been held constant.

3.7.1 Subjects

30 students of Psychology from Stanford University participated in this experiment. All were native speakers of English. They were each paid $5 for their participation.

3.7.2 Materials

The experimental materials again consisted of 24 sentences, 12 of which contained a frequent phrase and 12 of which contained a matched infrequent phrase. Each frequent phrase had a matching infrequent phrase that had an identical syntactic form (they were also matched for other features as we will discuss). The match for syntactic form was performed by using the part of speech annotation provided with the BNC. Each infrequent phrase had an identical tag sequence to its frequent counterpart. As in experiment three, the sentences were assigned to 2 groups of 12 sentences, each of which contained 6 frequent-phrase-containing and 6 infrequent-phrase-containing sentences. The two sets were again presented to different groups of 15 subjects. Given that frequent phrases and their infrequent matches were lexically different we expected no confounding effect from subjects being presented with both an item and its match. Unlike in experiment three, then, the six frequent phrases and their matches were kept in the same group. The target phrases were all of either four or five words in length. There was a minimum of one third of the log frequency range for phrases of 4-5 words between any item from the frequent group (mean = 284, minimum = 32, maximum = 817) and the infrequent group (mean = 1, minimum = 1, maximum = 1). All selected phrases and their matches can be seen in figure 3.4. As in experiment three there are various other dimensions along which our stimuli might vary and that we need to consider. The phrase pairs in experiment one varied only by a single word. Controlling them was therefore simple. By necessity, the stimuli in experiment four have a greater degree of variance. This was, however, kept to a minimum and will be factored out in our analysis.

Phrase pairs were checked for any systematic syntactic ambiguity by quantifying the word category ambiguity as in experiment three. This was done by calculating the percentage of occurrences in which each word was seen with the syntactic category with which they occurred in the phrase. It was found that each word occurred with the relevant category in a mean of 90.58 % of cases for the frequent group, and 92.87 % for the infrequent group. There appears then to be no systematic difference in word category ambiguity across the groups. Compositionality judgements were provided by

Frequent Phrases	Count	Infrequent Phrases	Count
on the other side of	817	on the great slab of	1
in the fullness of time	32	at the nature of work	1
would be grateful if	113	may be ineffective if	1
the turn of the century	452	the skull of a hedgehog	1
on the right hand side	41	for a new engine house	1
be taken into account	517	be used under gravel	1
net profit for the year	103	armed man in a raincoat	1
there can be no doubt	166	there would be a result	1
from the point of view	468	about the role of taste	1
a waste of time	256	a right of sale	1
per person per night	338	from prey to predator	1
be borne in mind	223	be moved by road	1

Table 3.4: Phrases and Frequencies for Experiment Four

two native speakers in the same way as for experiment three and they assessed that none of the phrases were non-compositional.

The groups differed very little in terms of the frequency of their component words, with the frequent sequences having a mean log frequency of 11.92, and the infrequent sequences a mean log frequency of 11.59. There was a slightly larger difference between the groups in the component transitional probabilities (calculated in the same way as for experiment three), with the frequent group having a mean log probability of -4.07 and the infrequent a lower probablility of -5.22. This difference that represents more than one seventh of the log frequency range for all observed bigrams, and consequently it will be vital that we factor this out.

The segmented experimental sentences can be seen in appendix D. These were obtained from the BNC as in experiment three. The mean sentence position was 5.9 for the frequent sequences and 6.75 for the infrequent. Plausibility judgements were obtained in the same manner as in experiment three.

3.7.3 Method

The procedure was as in experiment three.

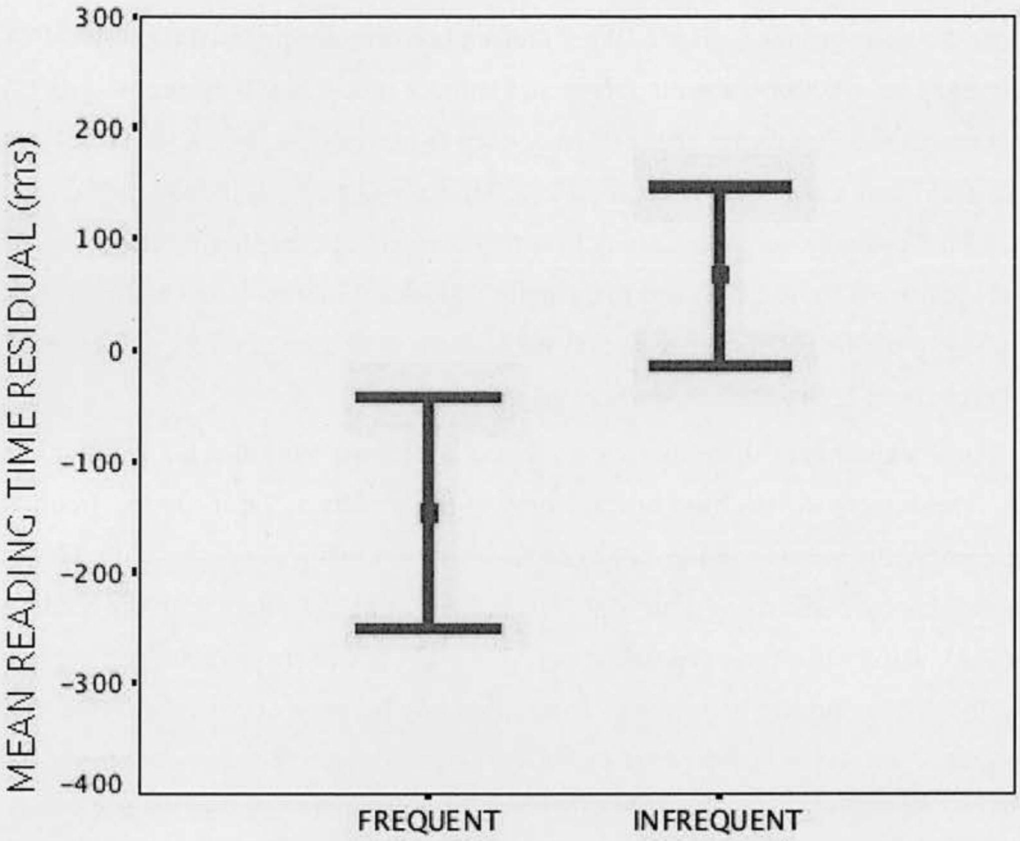


Figure 3.6: Mean experiment four reading time residuals (adjusted for length)

3.7.4 Results

The raw reading time means from this experiment can be seen in appendix F in tables F.1 and F.2. Prior to analysis the raw data was adjusted for word length as in experiment three. A regression was performed for each subject over all the sentence segments with the reading time as the dependent and length as the independent variable. The resulting regression equation was used to calculate an expected reading time for each target sequence. The mean (over all subject) *r* value for the correlation for group one was .41 (df = 59; p < .001), and for group two was .52 (df = 59; p < .0001). The mean r^2 values were .20 and .29 respectively. This expected reading time value was subtracted from the observed values and the residuals were subjected to analysis.

The mean residuals for the two conditions can be seen in tables F.1 and F.2 and are represented in figure 3.6. This shows a faster length-adjusted reading time for the frequent sequences than for the infrequent. Each subject in the experiment was included in both conditions, making a within-subjects analysis appropriate. However there were two groups of subjects reading two unrelated sets of stimuli, which introduces between-groups variance. We therefore performed a mixed design ANOVA with frequency as a within-subjects factor and subject group as a between-subjects factor. This revealed a significant effect of frequency in both a by-subjects (F1(1,28) = 8.822; $p < .025$) and a by-items analysis (F2(1,10) = 6.871;$p < .05$). However it also produced a frequency × group interaction that was significant in a by-subjects analysis (F1(1,28) = 5.576; $p < .05$) and marginally significant in a by-items analysis (F2(1,10) = 4.343; p = .064). This suggests that while there is an overall effect of frequency, it is not consistent across the two subject groups.

As in experiment three there are various additional variables we wanted to factor out. These were the position of the chunk in the sentence, the mean log frequency of the component words and the mean chunk-internal transitional probability. Once again we did this by performing a multiple regression for each subject with the reading time for each sequence as the dependent variable and each of these factors as independent variables. The models had a mean correlation coefficient r of .53 (df = 56; $p < .0001$) for group one and .57 (df = 56; $p < .0001$) for group two. They have mean r^2 values of .29 and .34 meaning that the additional variables account for an additional 9% and 5% of the variance in the data respectively. We then calculated the expected reading time for each subject for each chunk using the resulting regression equation, subtracted this from the actual reading time, and took the residuals as our adjusted times.

The mean residuals can be seen in tables F.1 and F.2, and are represented in figure 3.7. They show a faster reading time for the frequent sequences. We again performed a mixed ANOVA with frequency as a within-subjects factor and group as a between-subjects factor. We found a reduced but still significant effect of frequency by subjects (F1(1,28) = 6.255;$p < .025$) and by items (F2(1,10) = 5.900;$p < .05$). We also found a significant frequency × group interaction in a by-subjects analysis (F1(1,28) = 5.122; $p = <0.05$), and a non-significant interaction in a by-items analysis (F2(1,10) = 4.832; p = .053).

As in experiment three, we want to make sure that our results are not being produced by a difference in the plausibility of the sentences across the two conditions. We check for this in the same manner as in experiment three. A multiple regression was performed for each participant, with reading time as the outcome variable and length, position, mean word frequency, mean transitional probability between the component words, and the mean of the natural logarithms of the plausibility judgments as predictor variables. If plausibility is affecting reading time, then we would expect to find a

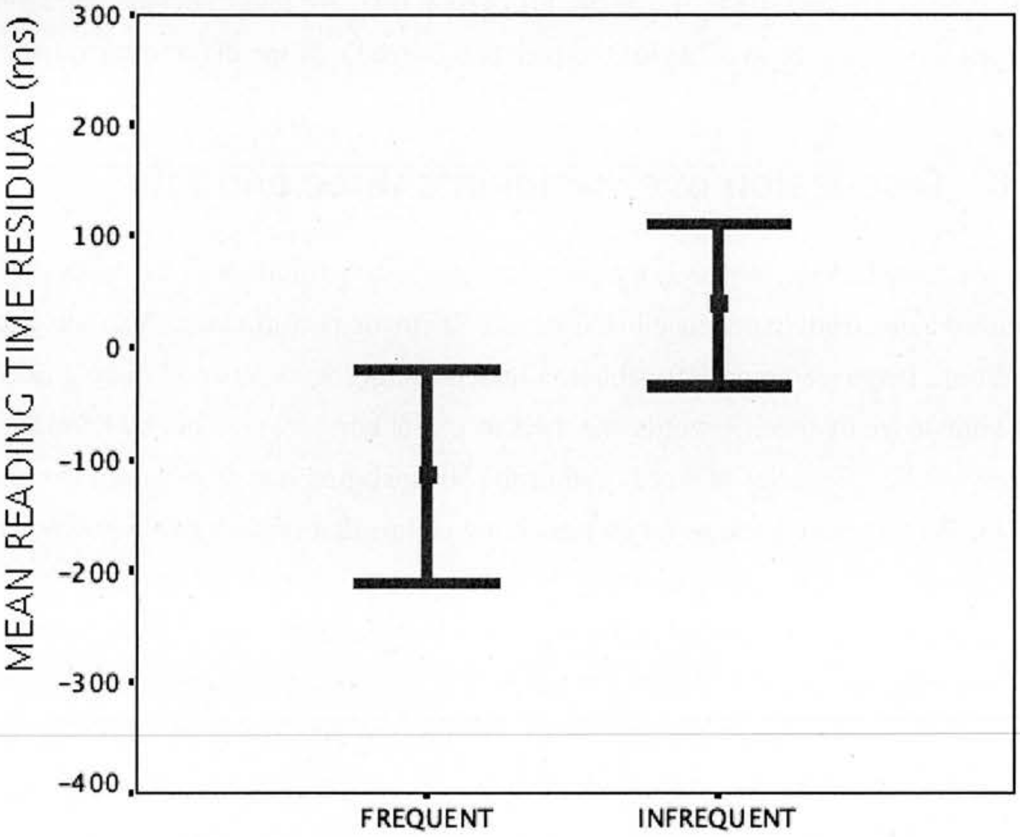


Figure 3.7: Mean experiment four reading time residuals (controlled for length, position, frequency and transitional probability)

significant negative correlation between them once all the other variables have been accounted for. In fact what we see is that in group one, for 9 of the 15 subjects, there is a positive correlation meaning that the reading time for a sequence increases rather than decreases as the plausibility of the sentence increases. For this group there is a mean positive correlation coefficient *r* of .090 (df = 6; p = .79). For group two there is a positive correlation for 6 of the subjects, and small overall negative correlation of -.043 (df = 6; p = .90). We next perform a single group *t*-test for the regression coefficients to look for any significant effect. We obtain *t* values of 1.147 (df = 14; p = .270) for group one and -.563 (df = 14;p = .582) for group two. We thus conclude that sentence plausibility is not responsible for the variation in reading time in our experiment.

3.8 Discussion of experiments three and four

As we described in chapter 2, the dominant generative tradition in linguistics has abstracted away from experience in explaining linguistic performance. According to this tradition, language comprehension and understanding is understood to be guided not by knowledge of specific words and their recurrent combination, but by knowledge of the syntactic categories of words and of the abstract grammar that is used to combine them. Writers have gone as far as presenting claims that readers retain no knowledge of the surface form of language they have heard in short-term memory (Sachs 1967). However in recent years, due to the widespread availability of corpora there have been various attempts to show that the details of linguistic experience are in fact reflected in a speaker's knowledge of and processing of language (see Bod *et al.*, 2003 for surveys of the impact this has had on various fields).

One of the details of linguistic experience that has been shown to affect users' knowledge of their language is frequency. In addition to giving a processing advantage for certain words, frequency has been shown to affect both comprehension (readers' preferences for particular subcategorisation frames for verbs have been shown to vary with the frequency of that frame, e.g. Trueswell, 1995) and production (frequent bigrams have been shown to be more likely to undergo phonetic reduction than infrequent, e.g. Krug, 1998, Bybee and Scheibman, 1999) for larger phenomena. We saw in section 2.3.1.1 that a processing advantage has been shown for frequent two-word sequences over infrequent. Experiments three and four provide evidence of a similar effect for longer repeated sequences.

In experiment three we recorded reading times for a set of frequent phrases and for a set of infrequent phrases each of which was identical to a matched phrase in the frequent group except for the final word. A repeated measures ANOVA revealed a significant difference in reading time between the two conditions in both a by-subjects and a by-items analysis when all other sequence variables were held constant. We take this as evidence that frequent phrases have a processing advantage over infrequent phrases.

In experiment four we again recorded in-context reading times for frequent and infrequent word sequences. Each frequent phrase was matched with an infrequent phrase that was lexically different but had an identical syntactic form. These were presented to two groups of subjects each of whom read both the frequent and infrequent members of the paired sequences. We performed a mixed ANOVA with frequency as a within-subjects factor and subject group as a between subjects factor. We found a significant effect of frequency in both a by-subjects and a by-items analysis. We also found a significant frequency × group interaction.

We can take this as evidence that frequency is having a significant effect on reading time. However the interaction tells us that this effect is not the same across the two groups. In fact, while there is a very large difference between conditions in group two, the effect is much smaller in group one. This is interesting. In experiment three the stimuli sets were split across the subject groups, giving us a counterbalanced design. In experiment four, by contrast, each subject group saw an entirely different set of materials, with each subject reading both the frequent and infrequent items from each pair. This means that the difference in effect across groups could be an effect not only of subject but of item variance. One possible explanation for this is that it is an effect of phrase type. While the conditions were matched for syntactic form, the two groups contain a number of different phrase types. The effect of frequency could vary from phrase type to phrase type. Another possible explanation of the difference in the effect across our groups is that it is down to differences in phrase frequency across the two groups. All infrequent phrases occur only once across both sets. However in group one, the frequent phrases have a mean log frequency of 5.25 while in group two the mean is 5.36.

To what might we attribute the apparent processing advantage for frequent phrases that we have seen here? A substantial literature has shown that non-compositional phrases are processed more rapidly than compositional phrases. As we saw in section 2.3.1, a number of writers, including Swinney and Cutler (1979), Gibbs *et al.* (1989) and Titone and Connine (1999) have taken this as evidence that the non-compositional sequences are being stored and retrieved as a whole. A similar explanation might be provided for the result reported here. One explanation of the frequency effects in both our experiments is that the salience of the chunk in the linguistic environment means that the reader has a recollection of having seen some or part of the sequence before. This memory aids their identification of the chunk. It needn't be the frequency of the entire chunk that causes the effect. It could be a recollection of only a part of it. However, since we controlled for the word and bigram probabilities, it cannot be only the memory of the frequent words or bigrams that is causing the effect. Such a conclusion is supported by evidence from, for example, Lancker-Sidtis and Rallon (2004), who show that people have better recall for frequent word combinations. However, before accepting that memory is responsible for the effect it would be desirable to conduct a tightly controlled memory experiment for the chunks, along the lines of that performed for two-word sequences by Bower (1969).

The idea that frequency affects reading time for extended word sequences may have far reaching implications. However, further work is needed to clarify the significance of the result. By using a difficult-to-read font, we made the task of identification harder and consequently arguably exaggerated the affect of familiarity on processing. It may be that such an effect would not appear when reading text in a familiar font. Further work is necessary before we can draw any conclusions concerning the effect of sequence frequency on reading familiarly presented texts. It is possible that any effect, if it exists, would be smaller, and consequently would require a larger set of stimuli. Given the nature of the data, this might be difficult within a controlled factorial design but certainly the analysis of large volumes of reading data using regression analysis in the mode of McDonald and Shillcock (2003) would be possible.

3.9 Chapter Summary

In this chapter we provided evidence that information about the frequency of multiword sequences obtained from the 89-million-word written component of the British National Corpus is both reliable and informative.

In experiments one and two we provided evidence that counts for multiword sequences found in corpora are at least as stable and reliable as individual words of equivalent frequency. We extracted counts for both acknowledged MWEs found in the SAID collection of idioms, and a selection of arbitrary multiword sequences. These sets covered items of high, medium and low frequency of lengths between 2 and 7 words. We calculated their stability of occurrence across the corpus using various published measures. We similarly calculated the stability of a randomly selected set of words that matched these items for frequency. We found that overall the stability of the multiword sequences was not significantly smaller than that of the words, and that in fact according to at least one measure the SAID idioms were more stable in occurrence than the individual words.

In experiments three and four we confirmed that counts taken from the corpus were informative by providing evidence that they are reflected in speakers' knowledge of the language, as revealed by reading times. In self-paced reading experiments we found a significantly quicker reading time for a set of phrases that occur in the BNC with high relative frequency than for a set of infrequent sequences that were matched on all key dimensions. In experimental three we matched frequent phrases with infrequent phrases that were identical except for their final words, and found that the phrases had an reduced reading time across all phrase items when other variables were held constant. In experiment four we matched frequent phrases with infrequent phrases that had an identical syntactic form. We found that when other factors were held constant the frequent phrases were read more quickly that the infrequent. However we used two different groups of subjects in our experiment who saw different sets of items, and we found a marginally significant interaction between conditions and groups. The two groups of items contained matched phrase pairs of different types, and while there is a clear effect of frequency it is possible that this might vary across phrase types.

Chapter 4

A measure of syntactic fixedness for the identification of multiword units

In chapter 2, we introduced a distinction between institutionalised and lexicalised phrases. An institutionalised phrase is a word combination (contiguous or otherwise) that has become common in a language as a conventional way of communicating a particular idea or series of ideas. As we saw, the argument for the existence of such expressions is made on the basis of their relative frequency. In chapter 3 we provided evidence that stable counts for multiword sequences can be obtained from corpora, and furthermore that these frequencies are reflected in speakers' knowledge of the language. This suggests that the corpus frequency of a phrase reflects its salience in the linguistic environment. It seems therefore that corpora can provide useful information about the degree of institutionalisation of a phrase (we will discuss this point further in section 4.5).

While institutionalisation may be evident from the frequency of a phrase, however, there exist other phrases which belong in the dictionary but are not necessarily frequent. These we identified in chapter 2 as lexicalised phrases. These are phrases whose form or meaning cannot be accounted for in terms of the productive principles of the language that can be used to account for the rest of language usage. Take an English phrase like *darken my door*. In the written component of the BNC phrases of the form *darken (possessive pronoun) door* appear just 4 times, putting this pairing of verb and noun in the lowest third of the frequency range for such combinations. The phrase is included in most dictionaries of English idioms, and yet clearly this cannot be justified on the grounds of frequency. Instead the phrase finds its place in the lexicon because of its meaning and its syntactic behaviour. It has a meaning "to trouble somebody with your presence" that is familiar to a large portion of native speakers and yet which cannot be straightforwardly recovered from the meaning of the component words. It is mostly fixed in form, not being seen with many of the syntactic variations that we would expect of a verb plus object combination, such as passivisation (**My door was darkened by him for months*) or adjectival modification of the object (**He has been darkening my front door for weeks*).

Just as we cannot theoretically justify the inclusion of a phrase like *darken my door* on the grounds of frequency, so we cannot hope to recover such a phrase from a corpus if we focus on frequency alone. If a lexicographer chose to consider verb and object combinations for inclusion in a dictionary in order of their frequency, then s/he would have to read through more than half the list of phrases of this syntactic type (which as we will see in this chapter may be as many as 1 million unique items) before it appeared [1]. An adequate lexicon must contain items not only on the basis of frequency, but also those items that need to be included due to their syntactic or semantic properties.The next two chapters will propose and evaluate methods for directly identifying phrases that have a claim to entry in the lexicon on the basis of meaning and syntax. Chapter 5 will look at the identification of phrases that are non-compositional. The present chapter will look at the identification of phrases that are syntactically fixed.

Section 4.1 will discuss the phenomenon of syntactic fixedness and the explanations that have been offered for it, before introducing the phrase variety that we will deal with in this chapter. Section 4.2 will discuss the previous work that has looked at using syntactic variations to identify MWEs in corpora. Section 4.3 will describe the model of syntactic flexibility that we are going to use to identify fixed phrases. Section 4.4 will describe an evaluation of the technique. Sections 4.5 and 4.6 will describe various techniques that have been proposed for using measures of the association between words in phrases rather than raw frequency in order to identify MWEs and compare the performance of these methods with the syntactic variation measure we propose. Section 4.10 will discuss the results of both experiments.

[1] In section 4.5 we will discuss various methods that have been proposed for refining frequency-based extraction. However, as should become apparent from that discussion these refinements could not hope to capture such a phrase either.

4.1 On syntactic fixedness

The phenomenon we are concerned with here is the preference of certain word combinations for an unvaried canonical form over variations on that form. Human languages are productive. Although certain combinations of words occur much more frequently than others, even the most recurrent word combinations will be seen in many different forms with variations in order or the additional or removal of information to produce a different focus/emphasis or to communicate additional or different information. However, there exist phrases in English where a particular canonical form is strongly preferred such that some or all variations on that form that we would expect given the grammar of the language are rarely or never seen. This chapter is concerned with such phrases.

4.1.1 Syntactic fixedness in English verb phrases

The experiments described in this chapter deal with one particular kind of phrase - verb phrases of the form verb plus direct noun object (e.g. *walk the dog*, *pull teeth*, *take a leaflet*). This variety of phrase was chosen because of its frequency, because of its central place in the literature on MWEs, because of the availability of evaluation resources, and because of the wide variety of syntactic fixedness that has been reported for phrases of this kind. In a survey of the idiomatic phrases listed in the Collins Cobuild Dictionary of Idioms (Collins 2000), Villavicencio and Copestake (2002) found this kind of idiom to account for more of the entries than any other.

Riehemann (2001) investigated the syntactic flexibility of such idioms by performing a corpus analysis of 25 verb plus noun phrase idioms randomly selected from among the Collins Cobuild Dictionary of Idioms, 21 of which consist of a verb and an unmodified direct object. She manually identified examples of all these items in the American News Text Corpus. As expected, she found considerable fixedness, with some phrases allowing no variation at all. She discusses in detail the kinds of variation that are observed. Three kinds of syntactic variation dominate. These are passivisation, adjectival modification of the noun and the variation, addition or dropping of a determiner. I will discuss each in turn, as well one additional variation that Riehemann does not cover: the addition of adverbial modifiers of the verb.

Verbs in English can occur in a range of different constructions, all of which give a different focus on the event described. For most transitive verbs, the dominant construction is its active form (e.g. *Ellen photographed the giraffe*, *Mary kissed John*).

Intuitively, the focus in this form is on the agent and their actions towards the patient. However, it is relatively common to see many transitive verbs in what is known as the passive construction. This construction intuitively switches the focus to the patient (e.g. *The giraffe was photographed by Ellen, John was kissed by Mary*). Such a change of focus is made depending on the discourse context and the situation being described, and is possible for the majority of verb plus object combinations. For many lexicalised verb and object combinations, however, passivisation is rarely or never seen. For example, in over 589 occurrences of the phrase *call the shots* in the BNC, not a single passivisation occurs. The same is true over 169 occurrences of *lead the field*.

Nouns in English can be modified by adjectives. The kinds of adjectives that one will see with particular nouns is of course restricted by semantic plausibility, and speakers have strong preferences for particular combinations. Nonetheless all English nouns can be modified by an adjective. When that noun occurs in certain idiomatic phrases, however, it seems that such modification is either not permitted or is extremely unlikely. For example, Riehemann found over 276 occurrences of the phrase *hit home* in the American News Corpus, not a single example occurs with a adjective modifier. The same is true for the 430 occurrences of *speak volumes*. The other 19 relevant idioms all undergo some adjectival modification. However for many it is greatly restricted. For example, of 386 reported examples of *close ranks* only 5 examples have an adjectival modifier, and of the 198 examples of *bite the bullet* only 2 examples are reported to have a modified object. It is clear that these phrases have a strong preference for a fixed unmodified form.

Determiners in English are used to determine some aspect of the reference of the noun that they precede, such as number, specificity or possession. There are a great many restrictions that apply on the kinds of nouns that determiners can occur with, depending on factors such as whether the noun is mass or count. Nonetheless, nouns can occur with a range of different determiners, or indeed sometimes with no determiner at all, depending on the context of reference (see Bond, 2005 for a detailed account of the factors that dictate determiner usage). However, when the noun is part of a larger set expression that variation is often greatly restricted. For example, in over 518 occurrences of the phrase *turn the tables* Riehemann only notes one occurrence in which *the* does not occur (in this case the determiner is in fact completely dropped giving the phrases *turn tables*).

We want to consider one other variety of phrase variation that Riehemann does not discuss. Just as the meaning of nouns can be added to or altered by the addition of adjectival modifiers, so verbs can be modified by adverbs. Adverbs often add information about the manner or extent of events. However for a phrase like *hit home*, many adverbs that could be used to modify the verb *hit* such as *hard* or *quickly* are not acceptable. We might expect then to see a significant difference in the freedom to attach adverbs to many verb and noun object phrases.

4.1.2 Accounting for fixedness

Before moving on to discuss our approach to modeling the fixedness we have described above, it will be useful to briefly introduce the explanations that have been given as to why this fixedness occurs. One frequent argument that is given is that the syntactic restrictions are a result of the meaning of the phrases. In section 2.2.3, we introduced the phenomenon of the semantically non-decomposable phrases. These are phrases where the component words of the phrase do not correspond to separate parts of the meaning of the phrase. An example is the expression *shoot the breeze* which can be paraphrased as "engage in lighthearted conversational exchange". The words in the expression do not correspond to any part of this meaning. That is to say that there is no clearly separable part of the meaning of the expression that corresponds to shooting or to the breeze. This is also an example of an expression that allows very little variation in form. The phrase is not conventionally passivised (**the breeze was shot by Bob and Samantha*), and the determiner is not dropped or varied (**shoot breezes*, ** shoot a breeze*). Nunberg *et al.* (1994) argue that the non-appearance of the syntactic variations for such expression is down to their meaning. Since the component parts of the expressions are not individually meaningful to speakers, they do not vary the form to put the focus of the uttererance on the object (as passivisation would) or add to its individual meaning (as modification would).

As we noted before, in her examination of 21 verb and direct object idioms, Riehemann (2001) found considerable fixedness in V+NP MWEs, with some phrases allowing no variation at all. Consistent with the semantic explanation for fixedness given above, she found that items which she classified as non-decomposable were notably less likely to be seen in varied form. While her analysis supports the hypothesis that the semantics of the expressions is affecting their form, however, it also highlights the limitation of the hypothesis. While the decomposable idioms discussed are more likely to allow variation, it was also found that many seemingly decomposable examples also show considerable fixedness. For example the expression *deliver the goods* which can be paraphrased as "deliver results", with *deliver* meaning "deliver" and *goods* meaning "results", has a completely unvaried form in 84% of its 176 occurrences. Similarly the decomposable expression *break the ice* occurs in canonical form in 79% of its 183 occurrences and *lose ground* in 70% of the 2350 times that it is seen. While non-decomposable expressions might be less likely to vary than decomposable, it is clear that fully decomposable idioms are also often very fixed in form. To explain this will require a different account.

An alternative explanation of the tendency for certain word combinations to prefer a particular form is that it is a result of a process known as **entrenchment**. The term entrenchment has been most widely used in the literature on acquisition. It has been shown that in learning a language, if a child frequently hears a verb in a particular construction, they are less likely to extend that verb to a novel construction. Theakston (2004) showed an affect of entrenchment for both children and adults. She presented 5-year-olds, 8-year-olds and adults with over-generalisation errors, where verbs were used with apparently ungrammatical argument structures (e.g. intransitive verbs were used transitively in sentences such as *The joke was so funny it really giggled me*), and asked them to make grammaticality judgments. She found that these were judged to be less grammatical by all groups when the verb involved was frequent. This is explained as a result of entrenchment – when a verb is more established in an intransitive usage it is less acceptable when seen in the transitive usage.

The literature on entrenchment is concerned with the existence of restrictions on the application of productive rules for individual words. However, the idea that certain forms become preferred over others can easily be extended to multiword phenomena. Syntactic fixedness could occur simply because particular canonical forms become habitualised, and consequently block variations. Given the processing advantage that familiar phenomena seem to have, we might speculate that once a form becomes established its usage is strongly preferable to variations on the form for reasons of ease of processing and efficiency of communication.

4.2 Previous work

This section will describe the previous work on the automatic detection of syntactically fixed expressions.

Wermter and Hahn (2004) discuss the need for more linguistically motivated approaches to detecting MWEs. They point out the limited modifiability of MWEs and suggest a technique for quantifying this modifiability that when combined with frequency information can help to identify MWEs. They extract all preposition-noun-verb combinations from a 114 million word corpus of German news text. They also identify all the supplementary lexical information that occurs between the preposition and the verb. Their intuition is that if there is a single piece of supplementary lexical material that is particular likely to occur with the phrase then it is more likely to be a collocation. The probability of each piece of supplementary material occurring with the phrase is calculated by dividing its frequency of cooccurrence by the summed frequencies of all its supplementary material:

$$P(PNV, Supp_k) = \frac{f(PNV, Supp_k)}{\sum_{i=1}^{n} f(PNV, Supp_i)} \tag{4.1}$$

The probability of its most frequent supplement is taken to be its degree of fixedness. A final score is then calculated by taking the product of this score and the probability of occurrence of the triple (which is simply its frequency divided by the sum of the frequencies of all triples).

The researchers then looked at the list of items and manually evaluated the precision achieved for the top *n* items at various values of *n* on the task of identifying MWEs. They found that their measure outperformed the t-test, loglikelihood ratio and frequency. While the results here are promising, the intuition behind the work is not entirely convincing. I can see no reason to think that a MWE will contain a single more likely modifier, unless that phrase is part of a larger collocation. Certain kinds of MWEs might be less modifiable, but having a dominant single modifier is not the same thing. One explanation for the apparent success of the method is that it rules out the highly productive example. From a theoretical perspective, though, one might expect the method to work better if one was also able to identify those items which had no or very few modifiers.

The previous paper focused on the limited modifiability of MWEs. There are, however, many other kinds of limitations that can exist on the syntactic variation of MWEs (e.g. passivisation). One work which allows consideration of a greater variety of these within a measure of syntactic fixedness was conducted within the corpus linguistic rather than the computational linguistic tradition. Consequently the analysis is not fully automated and no large scale quantitative evaluation is performed. It does, however, contain some well-motivated ideas.

Barkema (1994) suggests a metric for determining the flexibility of an idiom using corpus frequencies, by comparing the percentage of occurrences for which that idiom has its base form with the percentage of occurrences of all phrases of that type that occur in that form. The method relies on extracting all base forms of the idiom for a corpus, and all variants that exist. The example he provides is *cold war*. He extracts all base form examples of this phrase from a corpus as well as all variations on it such as *cold civil war* and *not-so-cold war* by searching manually. He then calculates the percentage of total occurrences of the phrase that occur in the base form.

The next step is to decide the syntactic type of the phrase from its part of speech tags, and to define all part of speech tag sequences that can occur for that phrase type. Then by obtaining counts from the corpus for the base form tag sequence and all the variants, he calculates the percentage of occurrences of the phrase type that occur in the base form.

By eyeballing the two percentages, Barkema suggests we can get an idea of how fixed the phrase is relative to other phrases of its type. This work doesn't provide us with a single score that we can use to quantify flexibility. However, if we think of this in terms of probabilities rather than percentages, there would many way in which we might quantify fixedness.

4.3 Our model

This section will describe how we use a corpus to discover which phrases are syntactically fixed and which are syntactically flexible.

4.3.1 Identifying variation

In the review of the literature above, we identified four important kinds of non-morphological variation that such phrases can undergo. These are as follows:

- Variation, addition or dropping of a determiner so that, for example, *run the show* becomes *run their show*, *make waves* becomes *make more waves*, or *strike a chord* becomes *strike chords*.

- Modification of the noun phrase so that, for example, *break the ice* becomes *break the diplomatic ice*. We refer to this as internal modification.

- The verb phrase passivises so that, for example, *call the shots* is realised as *the shots were called by*.

- Modification of the verb by an adverb, so that *eat dinner* becomes *eat dinner quickly*. We refer to this as event modification.

In this work we are going to use the BNC to make observations about the extent to which these variations are permitted by verb and noun object phrase combinations in order to obtain a graded assessment of their fixedness. In order to do this we need some way to a) identify such combinations, and b) to identify when they are displaying syntactic variation. In order to do both of these we utilise syntactic parsers. In section 4.3.2 we will describe the measures we use to quantify variation, using counts extracted from a corpus. In this section we will describe how we obtain these counts through the use of syntactic parsers.

Although there have been efforts over the past decade to create standards for the output of different syntactic analysis systems, for theoretical and practical reasons there remain significant differences in the kinds of information systems provide. And clearly there are significant differences in the specific information provided, reflecting the vastly different approaches that are used. For this reason the results produced by any technique that makes use of a parser is going to vary depending on the parser employed. In order to ensure that any successes or failures of the syntactic variability measures we propose are not entirely down to our choice of parser, we are going to utilise two very different parsers , both of which are widely available.

The parser we use in the RASP system (Briscoe and Carroll 2002). RASP is a modular parsing system based on a tag-sequence grammar. Input text is first tokenised and then part-of-speech and punctuation tagged using the detailed CLAW-2 tagset which consists of 155 different labels. This information is then used to lemmatise the text using the technique described in (Minnen *et al.* 2000). The next stage is parsing with a grammar that consists of approximately 400 unification-based phrase-structure rules. This is utilised by a probabilistic LR parser. Probabilities are associated with analyses based on the structure. This is augmented with information about the probability of seeing one of a set of high to medium frequency verbs occurring with one of 23 different subcategorisation frames, and about the probability of certain phrasal verb combinations. Unlike many contemporary parsers then, including the next one we will describe, the parser makes use of no significant information about the probability of seeing particular relationships between lexical items. Since we are looking here for cases where the syntactic behaviour of particular word combinations deviates from general grammatical patterns, the fact that this parser does not utilise statistical information about word combinations makes it particularly suitable for our purposes.

Although RASP does construct phrase structure trees, we choose to make use of its grammatical relation output (this is described in Carroll *et al.*, 1998). The output for the sentence *The largest apartment was immediately bought by a Swedish couple* is as follows:

```
(|ncsubj| |buy+ed:6_VVN| |apartment:3_NN1| |obj|)
(|arg_mod| |by:7_II| |buy+ed:6_VVN| |couple:10_NN1| |subj|)
(|ncmod| _ |apartment:3_NN1| |largest:2_JJT|)
(|detmod| _ |apartment:3_NN1| |The:1_AT|)
(|ncmod| _ |couple:10_NN1| |Swedish:9_JJ|)
(|detmod| _ |couple:10_NN1| |a:8_AT1|)
(|mod| _ |buy+ed:6_VVN| |immediately:5_RR|)
(|aux| _ |buy+ed:6_VVN| |be+ed:4_VBDZ|)
```

Each line tells about the relationship between pairs of words. The relation comes at the beginning of the line. In most cases the head comes next followed by its dependent, both with their part of speech and their position in the sentence. Any additional words that define the relationship between the words appears immediately after the grammatical relation. We use this description to extract all of the candidate phrases that we are going to consider. We extract all verb and nouns pairs connected by a object relation that are found in the corpus. We are interested here in the verb and object relationship between *buy* and *apartment*, and we can use the output to identify the grammatical variations that this combination displays.

The first thing to note is that the phrase is passivised. The parser deals with varying constructional frames as alternations, with a syntactic process modifying the semantic relationship between a head and its semantic argument. Accordingly, two grammatical relations are posited between the head and dependent. The semantic or initial (prior to the application of the syntactic operation) relation appears at the end of the line. So here the *apartment* is described as an object of *buy* by the "obj" relation that appears at the end of the line. Because of the passivisation, *apartment* is also described as a non-clausal subject of *buy* by the "ncsubj" relation that appears at the beginning of the line. This presence of a semantic object that appears as a surface subject tells us that we are dealing with a passive. The "ncmod" relation tells us that the adjective *largest* is a modifier of apartment. The "detmod" relation tells us that *the* is a determiner attached to *apartment*. And finally the *mod* relation between *buy* and the adverb *immediately* describes adverbial modification. We can thus extract information about all three of the phenomena that we are interested in. Rather than simply classifying a phrase as allowing variation or not, we want to measure the extent to which it varies. We therefore make a count over the whole corpus of the number of times each verb-object pair occurs, and the number of times it occurs with each relation of interest.

The second parser that we utilise is Minipar. This is a descendant of Principar (Lin 1994). It is a broad-coverage principle-based parser which utilises a message parsing algorithm. Its grammar is encoded as a network consisting of 35 nodes and 59 links. The system first extracts all lexical entries for the input sentence, and passes each in turn as a message to the relevant network node. The node will then forward the message to other relevant nodes or combine the message with other nodes and then forward the resulting message. This process is guided by constraints associated with each node and link. The result of this process is the creation of a chart at each node containing all structures belonging to that category. Parse trees can then be extracted from this network. The selection of the best parse tree is guided by statistics obtained by parsing an (unnamed) 1 gb newspaper corpus. This proceeds by finding the most probable tree given its root and its component dependency relationship where a dependency relationship is a triple consisting of a head, a modifier and a dependency relation.

Like RASP, Minipar outputs dependency relations. A parse of the sentence *The new teacher was completely flummoxed by the event* looks as follows:

```
flummox V:s:N   teacher
teacher N:det:Det       the
teacher N:mod:A new
flummox  V:be:be  be
flummox  V:amod:A  completely
flummox  V:obj:N  teacher
flummox  V:by-subj:Prep  by
by  Prep:pcomp-n:N  event
event  N:det:Det  the
event  N:mod:A  whole
```

The words on the left are heads of phrases, the words on the right their dependents, and the information between describes the parts of speech of the two words and the grammatical relation between them. We can use this to observe that *teacher* is an object of *flummox*, and thus that *flummox teacher* is a verb and object pair. Such relations provide all of our candidate phrases, the syntactic behaviour of which we are to observe across the corpus. One problem with this is that Minipar makes no distinction between nouns and pronouns. Therefore the phrase *make it* would be described as an object relation between a verb *make* and a noun *it*. We do not want to consider these here. Minipar takes its lexicon from WordNet (Miller *et al.* 1990) which contains no pronouns. In order to filter pronouns from the output we therefore use a different lexical resource. We discard all pairs where the object is listed in the CELEX database (Baayen *et al.* 1993) as a pronoun and does not also appear as a noun in Wordnet 2.0.

As with RASP we use the grammatical relation output to extract all of the information we require about the phrases. We can see that in addition being an object (obj) of *flummox*, the word *teacher* is also the surface subject (as indicated by the s relation). This tells us that this is a passive construction. We can see that *flummox* has a adverbial modifier (amod) in *completely* telling us that the phrase allows event modification. Similarly we can see that object *teacher* has an adjectival modifier in the word *new*. The object also bears a determiner in the word *the*. We can thus obtain all of the information we require.

Passivisation, internal modification and event modification are straightforward to record and quantify. A variation is simply the presence of one of the identifying relations. The addition, dropping or variation of a determiner is not so straightforward. We are interested in how variable a phrase is in its determiner status. A variation is a deviation from its dominant status. We need therefore to determine what this dominant status is for each phrase. We are going to do this as follows. We record the determiner status of each occurrence of a phrase. The determiner status is defined by either the presence or absence of relations. A verb and noun object pair where the noun has no determiner relation is recorded as having no determiner. This is one potential determiner status. The others depend on the kind of determiner that is appended. There is only a single determiner relation and so we use the part of speech tags and the lexical items to learn about the different kinds of determiners used. The RASP parser uses the very rich CLAWS-2 tagset. This provides us with a tag for possessives, 2 different article tags (article and singular article), 6 different basic determiner tags (determiner, singular determiner, plural determiner, wh determiner, genitive wh determiner and wh-ever determiner), 2 before-determiner tags (before-determiner and plural before-determiner) and 6 after-determiner tags (determiner, single determiner, plural determiner, wh-determiner and wh-ever determiner). Although the last two categories of tag are named before- and after- determiners they can in fact occur in the absence of any other determiner and perform a specifier function. We consider each of these tags as a different determiner status. We also make two additional distinctions based on the lexical items that appear as determiners. Firstly for the article tag we introduce a distinction between the definite article "the" and negation of the object in the form of the word "no". For the singular determiner tag we make a distinction between the quantifier "every" and the indefinite article "a". For the RASP parser, then, we have 20 different possible determiner statuses: either no determiner at all, or any one of the determiners described above. Once the determiner status of all occurrences has been recorded, the dominant status for each item is taken to be the status that occurs most frequently. The number of variations in form is then taken to be the number of times that the phrases occurs with any status other than its dominant status.

Minipar has four different relations by which determiners can be attached to nouns. These are the "poss" or possessive relation, the "det" relation, the "post" or post-determiner relation and the "pre" or pre-determiner relation. The presence of each of these, as well as the absence of a determiner is recorded for each word pair. As with RASP, we also use the wordforms to distinguish the definite article, the indefinite article, negation of the noun and the universal quantifier. There are eight different kinds of determiner status that can occur with the Minipar output. The probability of variation is calculated over these in the same way as for RASP.

4.3.2 Quantifying variation

4.3.2.1 The probability of variation

We are interested here in measuring the degree of syntactic variation allowed by each verb-object pair found in our corpus. We will do this probabilistically. Firstly we use the counts that we extracted above to estimate the probability of each variation for each combination. The most straightforward way to do this would be to use maximum likelihood estimation. For a phrase t and syntactic variation V_1 the probability would be calculated as follows:

$$P(\mathbf{V}_1|\mathbf{t}) = \frac{freq(\mathbf{V}_1,\mathbf{t})}{freq(\mathbf{t})} \tag{4.2}$$

This, however, is not entirely satisfactory. Any corpus is limited in size and many of the verb-object pairs infrequent. Many of these will simply not be seen at all with one or more of the kinds of variation. For such word combinations, the above estimate would give us a probability of zero for one or more variety of variation. This is not acceptable for theoretical and practical reasons. We therefore need to smooth our model to account for these unseen events.

One approach to smoothing might be to back off to similar items so that we derive the probability of seeing a variation for a given verb-object pair from the observed probability of seeing that variation for verb-object pairs that have a similar meaning. However this would undermine the very thing that we are trying to observe. We are interested precisely in behaviors that are specific to particular word combinations, and backing off like this would remove this information from the model. Unlike in most situations, then, what we want is a completely naive smoothing method that knows nothing about the expected probability of variation for an item. We therefore use Laplace smoothing. This works by adding one to the number of observed variations for all items and adding the number of possible outcomes (which in this case is two - variation does occur/variation does not occur) to the denominator. Our estimator then is as follows:

$$P_{laplace}(\mathbf{V}_1|\mathbf{t}) = \frac{1+freq(\mathbf{V}_1,\mathbf{t})}{2+freq(\mathbf{t})} \tag{4.3}$$

This assigns a non-zero probability to all possible events.

Thus far we have talked about calculating the probability of individual variations. However, we are interested in obtaining a single measure of variability for an item rather than a number of different probabilities. We obtain such a score by taking the product of all four kinds of variation. What this gives us is equivalent to the probability of seeing an occurrence of the verb-object pair with all four varieties of syntactic variation simultaneously (which we cannot calculate directly because of the limited size of our corpus) under the assumption that the kinds of variation are independent:

$$P_{laplace}(\mathbf{V}|\mathbf{t}) = \prod_i P_{laplace}(\mathbf{V_i}|\mathbf{t}) \tag{4.4}$$

4.3.2.2 Fixedness as deviation from expectation

We have described how to calculate a probability of free variation for a given verb-object pair. There is one last consideration that we need to take into account. We have assumed until now that the flexibility observed for a given word combination is a unique property of the phrase. However, we need to consider the fact that each phrase has a prior probability of variation derived from the probability of variation of the component words. Take passivisation for example. Some verbs are more prone to passivise than others. The degree of passivisation of a phrase will therefore depend to a large extent upon the passivisation habits of the component verb, and is not a unique quality of the phrase. The same can be said for all our varieties of variation.

What we want then is not simply a probability of variation for each verb-object pair, but rather an estimate of the extent to which the probability of variation for that combination deviates from the variation we would expect based on the variation we observe for its component words. For this we use conditional pointwise mutual information. Each kind of variation is associated with a single component word. Passivisation and event modification are associated with the verb. Internal modification and determiner variation is associated with the object. We term this the core element and the other component the secondary element. We calculate the pointwise mutual information of the syntactic variation x and the secondary element y given the core element z, as seen in equation 4.5. In the case of passivisation and event modification the core element z will be the verb and the secondary element y will be the object. In the case of event modification and determiner variation the core element z will be the object.

$$\begin{aligned} I(x;y|z) &= H(x|z) - H(x|y,z) \qquad (4.5)\\ &= -\ log_2\ p(x|z) - [-\log_2\ p(x|y,z)]\\ &= -\ log_2\ p(x|z) + \log_2\ p(x|y,z)\\ &= log_2 \frac{p(x|y,z)}{p(x|z)} \end{aligned}$$

Conditional pointwise mutual information tells us the amount of information in bits that y provides about x (and vice versa) given z. This score tell us how the likelihood of seeing a given syntactic variation for a given verb plus object pair relates to the likelihood of seeing that variation for the relevant component word. If a variation occurs for a given word pair with greater likelihood than we would expect based on the frequency of seeing that same variation with the relevant component word, then the mutual information will be high. We want to find the information that is gained about all the syntactic variations by a particular verb and object combination. We therefore calculate the information gained about all the verb-relevant syntactic variations (passivisation and event modification) by the addition of the object, and the information gained about all the object relevant variations (internal modification and determiner dropping, variation or addition) by the addition of the verb. Summing these, as in equation 4.6 then gives us the total information gained about syntactic variation for the word pair W, and we take this as our measure of the degree of syntactic flexibility for this pair.

$$SynVar(W) = \sum_{i}^{n} I(VerbVar_i; Obj|Verb) + \sum_{j}^{n} I(ObjVar_j; Verb|Obj) \tag{4.6}$$

In simply summing the variation scores for each syntactic feature, we are assuming that they are all equally informative to the lexicographer. It might, however, be the case that a certain value (e.g. non-modifiability) is more useful in picking out items that are of interest to lexicographers. We want to find a way of combining our scores that takes this into consideration. We can do this by applying weights to the features. We introduce these weights by altering our measure to give separate scores for each syntactic feature (rather than calculating the variation score relative to each word as described above) and multiplying the variation information for that feature by its given weight. The measure for a wordpair W is calculated as follows:

$$SynVar(W) = \sum_{i=1}^{n} \lambda_i \, I(S_i; w_{secondary_{s_i}} | w_{core_{s_i}}) \tag{4.7}$$

where λ_i is the weight to be applied for syntactic feature i. These weights can be set empirically, using a method we will discuss in section 4.4.3.

4.4 Experiment five

The central claim of this thesis is that corpora can provide important information about the phrasal lexicon. In this section we will describe an experiment which looks directly at whether a corpus can reveal information about the syntactic fixedness of a phrase in sufficiently reliable manner that the results can be used to automatically identify MWEs. The aim of the measure described in this chapter is to provide a method for highlighting those phrases of potential interest. Here we will think of this a method for ranking the full set of items in the order in which the lexicographer should attend to them.

The evaluation procedure used here (first suggested by Evert and Krenn, 2001 for evaluating measures of lexical association; I will discuss this literature in section 4.5) involves producing and evaluating just such a ranking. The measure of syntactic flexibility will be used to produce a ranked list of phrase items (with the most fixed first). This ranking will be evaluated using existing lexical resources. We will use two different dictionaries of idioms. The ranked list will then be evaluated by measuring how many items from the dictionaries are found in the top *n* items of the ranked lists for different values of *n*. We can produce precision and recall scores for each *n*. We can then compare the method with other methods of ranking, either by comparing plots of *n* against precision/recall or by taking various values of *n* and comparing the scores achieved. We can also then measure the statistical significance of any difference from other methods.

There is a very important point that needs to be made about this evaluation technique. As we saw in chapter 2 there are a great many different kinds of MWE. For a phrase type such as the verb plus direct object VP there is a range of different reasons for inclusion stemming from frequency, syntax and meaning. A core argument of this thesis is that the acquisition of MWEs must involve the use of different methods of extraction for different phrase types. The technique here focuses solely on syntax. It can therefore only be expected to identify those items that need to be included in the dictionary on the basis of their syntax. However the dictionaries that we are using for evaluation will contain a variety of expressions, only some of which are syntactically fixed. First of all this means that we cannot expect the method to identify100% of items. Secondly, while we will compare the performance of the measure with frequency-based techniques, it should be understood that it is attempting to identify a different kind of phrase, and so it is as important to show that the measure is identifying different kinds of dictionary items as it is to show that it is identifying a comparable or better number.

4.4.1 Materials

The materials for this experiment were two dictionaries of idioms and multiple lists of verb and noun object pairs ranked according to the measures described. The dictionaries of idioms we employed were the Longman Dictionary of English idioms (Long and Summers 1979) and the SAID Syntactically Annotated Idiom Dataset (Kuiper *et al.* 2003). The former describes its aims as being "to provide the student of English with a thorough coverage of the most common idiomatic phrases in use" (p.vii). An exhaustive survey of the printed dictionary identified 773 unique verb plus object pairs. 627 of these items were found to exist in the list of verb and object pairs extracted using

RASP and 612 in the list extracted using Minipar. In order for an item to qualify there could be no obligatory lexical material between the verb and the object other than a single determiner. Items were extracted from the SAID list not by an exhaustive survey, but rather by using the phrase structure annotation that is included in the dataset. The same rules for inclusion were used and 593 unique items were identified, 529 of which occurred in the list extracted using RASP and 516 in the list extracted using Minipar.

The aim of this experiment is to show that our method of extraction can effectively identify valid MWEs. In any such evaluation there is a risk that the performance of a technique is particular to the lexical resource used and will not generalise. For this reason we report results using the two separate dictionaries. If the technique is effective then we would expect it to perform well for both resources. We will also report the scores achieved when evaluated against the superset of both dictionaries. Combining the two dictionaries give us a list of 1109 unique verb and object pairs, 914 of which were identified in the BNC using RASP and 893 using Minipar.

The other key materials in this experiment is lists of verb and object pairs which have been ranked using our scores. As explained above, the phrases are extracted by parsing the written component of the BNC using two different parsers. The RASP parser gives us 979,156 different verb object pairs. The Minipar parser (filtered for pronouns) gives us 1,089,210 different verb object pairs. This is a difference of almost 30000 between the two parsers.

In order to evaluate the performance of our technique it will be useful to compare its results with the ranks of scores that can be obtained by other means. The true baseline is of course a random ordering. Without any motivated means of ordering the items, the lexicographer would be confronted with a list of verb and object pairs in the order in which they came out of the corpus. However, given the size of the candidate set relative to our evaluation set, a random ordering would be expected to produce a precision of 1 in 1100, which is so below the minimum acceptable level of performance that a comparison would not be useful. In order to evaluate the usefulness of the measure we instead compare it with a frequency ranked list. As we have discussed throughout this thesis, frequency ordering will highlight a different variety of expression, and is in reality likely to be used in addition to our measure. However, as the most straightforward means of ordering available to the lexicographer, it will be useful to show that our measure is as useful as exploiting frequency information alone in highlighting items that need to be entered into the lexicon.

All of the methods of ranking items described here produce multiple items which have the same score, and hence the same rank. This was handled by ordering any sets of draws randomly. This was done by using Perl to generate a random number between 0 and 1, and using this to rank the drawn items.

4.4.2 Procedure

The BNC was parsed using our two parsers, and all verb and object pairs together with counts for their overall frequencies and their various syntactic variations were extracted from the output. These were then used to calculate our measures of syntactic flexibility, and the measures used to produced ranked lists with the least flexible first and the most flexible last. We then used the various sets of gold-standard items to calculate the precision and recall for each subset of the list moving from the top item down to the full list. We thus have *n* precision and recall scores where *n* is the length of the list of extracted phrases.

Precision and recall are measures devised for the evaluation of information retrieval systems (van Rijsbergen 1979) that have become standard for many NLP tasks. Precision is the percentage of the pieces of information provided (documents in the case of IR, verb and object pairs in the current experiment) that are correct. Recall is the percentage of the total number of available pieces of information available that a system has recalled. These scores are calculated as follows:

$$Precision = \frac{Correct\ phrases\ in\ top\ n\ phrases}{n} \tag{4.8}$$

$$Recall = \frac{Correct\ phrases\ in\ top\ n\ phrases}{Total\ number\ of\ dictionary\ phrases\ in\ total\ candidate\ set} \tag{4.9}$$

In IR the two measures give two perspectives on the performance of the system, recall telling how good the system is at getting the relevant information, and precision how much redundancy there is in its output. For our task the situation is rather different. Precision is very informative. If we think of the ranked list as output to be provided to a lexicographer, the precision tells us what percentage of the items that a lexicographer is examining are found in our dictionary lists. It is therefore a good estimate of the value of our method. Recall, however, is more problematic. We report recall as the percentage of the total number of dictionary items that are found in the sample. This is clearly not entirely satisfactory. Recall would only be truly informative if we had exhaustive knowledge of the valid MWEs to be found in the total candidate set. We know that our lists are not comprehensive and so we must assume that there will be

MWEs in the candidate set that are not in the dictionaries. We therefore report recall with caution. Although reference will be made to it, the majority of discussion will concern precision. We choose to present this rather than some combination such as the F-score because it is straightforward to interpret. It will be useful to note when reading these results that because the number of candidates in each sample is set in our experiment, the relative performance of two methods as measured by recall will be the same as for precision. That is to say that if a method has a higher precision value than another method in our experiment then it will be certain to have a higher recall value as well.

We calculate the precision scores for all n. In particular, the precision scores for n equal to 100, 1000 and 5000 were extracted, and the significance of any difference between our scores and the frequency ordered list was calculated. It is not uncommon in NLP to use regular statistical significance testing that works by calculating the probability that an observed result could have occurred by chance given some expected distribution. The test that is most commonly used in collocation extraction is the Chi-squared test (e.g. Evert and Krenn, 2001). However this test is in fact not appropriate for the task. The Chi-squared assumes that the values it is comparing have been produced using independent samples. Since we are comparing scores that result from the ranking of a single candidate set, they are clearly not independent.

Rather than using a significance test that relies upon an assumed distribution, we will use a computationally-intensive randomization test of significance called stratified shuffling. This technique (Chinchor *et al.* 1993; Cohen 1995) works by estimating the difference that might occur between scores by chance through a simulation. The null hypothesis is that for each of the two list of the top n items, the chance that any given item is going to be found in the dictionary is the same. In order to test this we do the following. First of all the difference between the precision scores obtained for the two groups is calculated. The phrases in the two groups are then randomly shuffled between the groups, and the precision scores of the two groups is recomputed. If the difference between the scores obtained by the shuffle-created groups is equal to or greater than that obtained by the actual groups then a counter c is incremented by one. For an exact randomization this procedure should be repeated 2^n times where n is the size of the list. However, often, as in our case, this number of iterations is not possible, and 10,000 has become the accepted number of randomization to be performed. Once the process has been repeated 10,000 times, the probability of the difference between the two groups occuring by chance can be straightforwardly calculated as (nc + 1 / nt + 1) where n is the size of the list, c is the total number of shuffles on which the resulting difference between the groups equalled or exceeded the observed difference, and t is the total number of iterations. As is standard in experimental statistics, we accept any result of $p < 0.05$ (a less that 1 in 100,000 chance of the difference occurring by chance) as significant.

4.4.3 Results

The outcome of the procedure described above is a list of precision and recall scores for the top n of all candidate phrases, where n ranges from 1 to the size of the whole candidate list.

First of all we calculated the scores obtained when the syntactic variation was calculated using each variety of syntactic variation in isolation (we will refer to these as our features). This was measured using both the simple probability as shown in equation 4.3 and the measure of variation relative to the behaviour of the component words as seen in equation 4.5. Consistent with our intuition, higher precision and recall scores were obtained for all variations when ranking with the latter score. This was true for both parsers. These scores can be seen in figures 4.1 and 4.2. Figure 4.1 provides a plot of the precision score each sample obtains when evaluated using the superset of the two dictionaries for all samples from $n = 1$ to $n = 10{,}000$. Figure 4.2 shows the recall scores obtained.

It can be seen from the plots that there is significant divergence in the performance of the different features. For the RASP parser the best performing feature is passivisation followed by internal modification. Determiner variation performs considerably worse. For Minipar determiner variation is the best performing single feature, followed by internal modification and passivisation. One important thing to note is that the RASP extracted information performs consistently better than that obtained using Minipar. So in fact the performance using determiner information is at least as good with RASP as with Minipar, but the other Minipar-extracted features perform significantly worse relative to this. The relatively high performance of the determiner information for Minipar may be because it was filtered using lexical information. For both parsers the only syntactic variation that performed very badly was the addition of an event modifier, obtaining a precision of less than 2 in 1000 in both cases.

The next step was to examine the results achieved using combinations of the syntactic features. The first score calculated was the combination of all the features. These

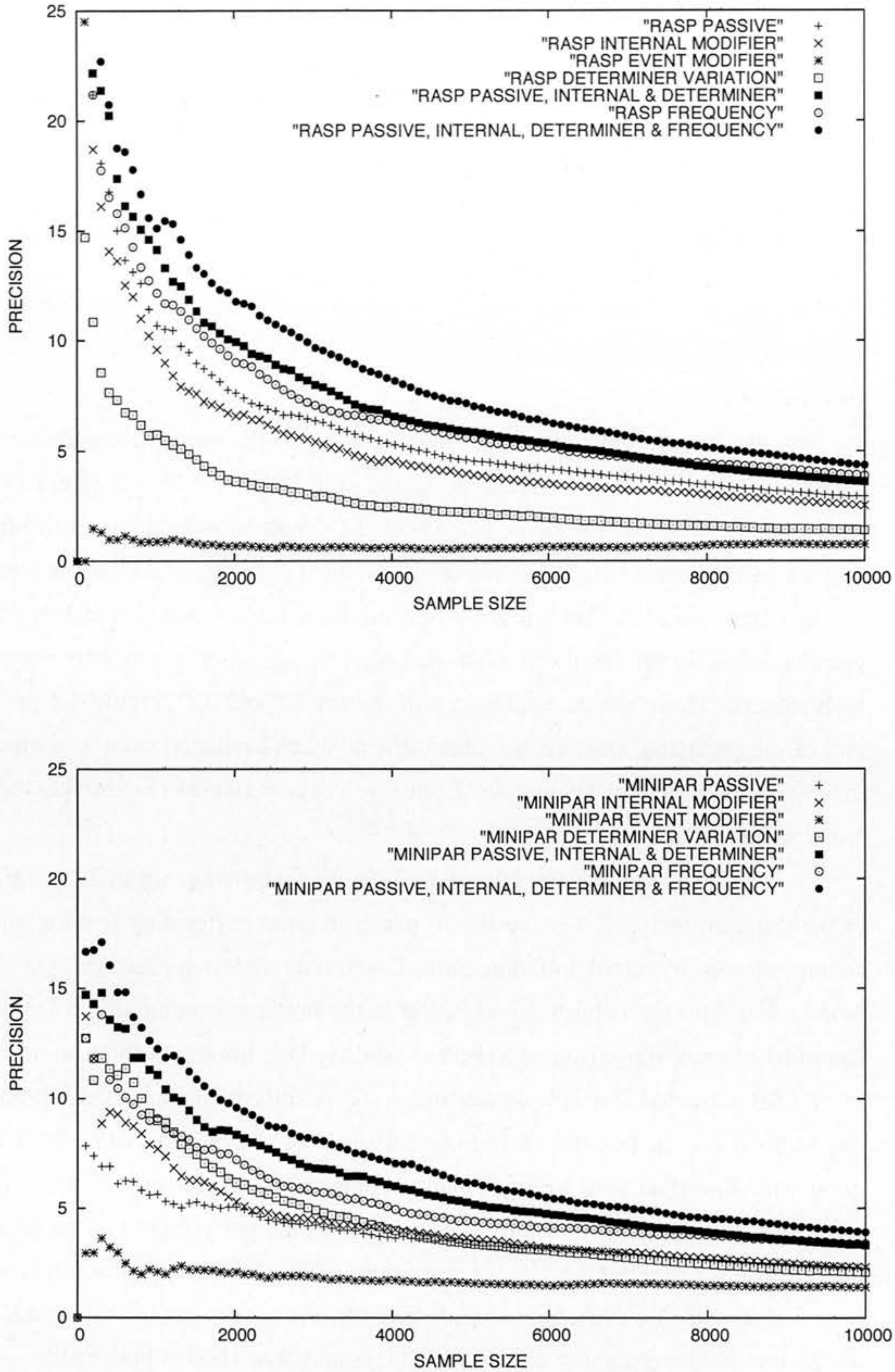
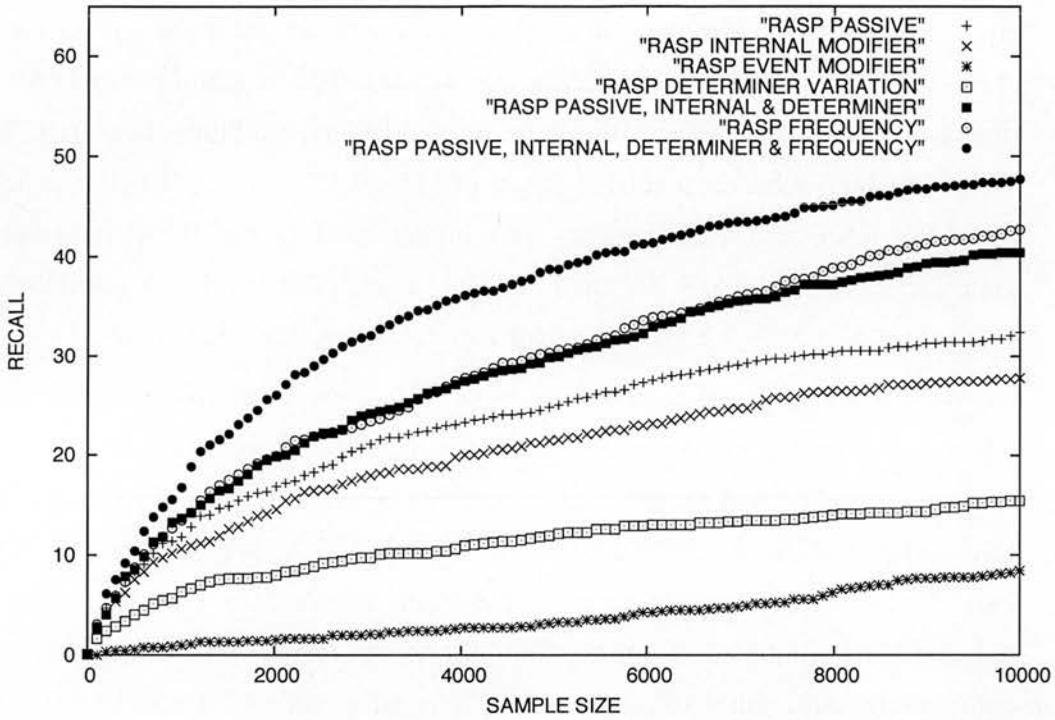


Figure 4.1: Precision score by sample size



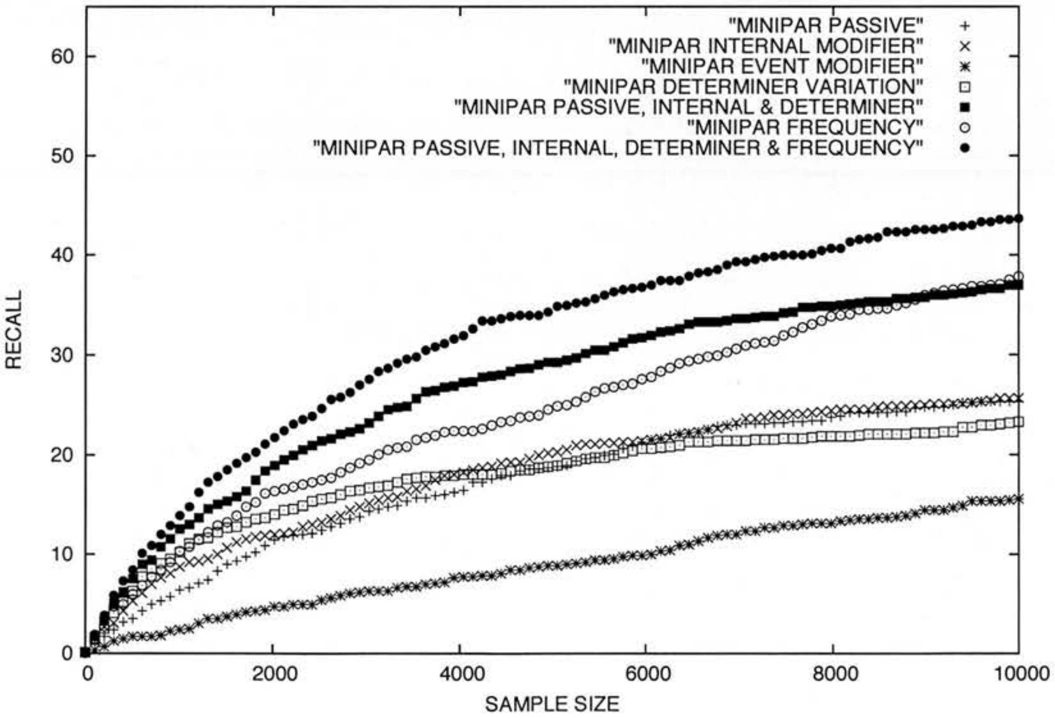


Figure 4.2: Recall score by sample size

were combined both by taking the product of the probabilities as seen in equation 4.4, and by combining the individual conditional mutual information scores as shown in equation 4.6. Using Minipar, the combined probabilities achieved a precision of 5% at n=100, 5.9% at n=1000 and 3.28% at n = 5000. This is significantly better than a random ordering, but it is greatly outperformed by the combined measure of relative variation which achieves precision scores of 11%, 9.4% and 7.32% at the same values of n. For RASP the overall scores were higher but followed the same pattern with precision scores of 13%, 8.4% and 5.72% for the combined probabilities at the various values of n, and 24%, 12.5% and 5.46% for the relative variation score.

We saw above that the various syntactic variations achieved very different scores when used in isolation. For both parsers, the measure of event modification achieved very low precision and recall scores. We therefore want to measure the performance achieved with different configurations of the features. As we would expect given the individual performances, the performance improves when we remove event modification from the combined measure. The best scores for both parsers are achieved by combining the measures of passivisation, internal modification and determiner variation. These are plotted in figures 4.1 and 4.2. Precision scores of 15%, 11.1% and 5.22% at n equal to 100, 1000 and 5000 are achieved with Minipar, and scores of 18%, 14.2 and 5.86% with RASP.

In order to measure the usefulness of these scores we want to compare them with the results achieved using a frequency ranked list. This comparison can be seen in tables 4.2 and 4.1. The two sets of features that we describe above can be seen as "'P,I,E & D unweighted " and "P, I & D weighted" (unweighted refers to the fact that we use unweighted values for all features). Frequency obtains precision values of 28%,12.2% and 5.56% at 100, 1000 and 5000. It can be seen that while the precision is high when taking a sample of the top 100 most frequent items our syntactic variation measure performs better over the more informative sample sizes of 1000 and 5000. Until now we have reported scores over the combined items from the two dictionaries. In these tables we also include the precision scores obtained when evaluated against the two dictionaries separately. We want to ensure that the results reflect how good the method is at extracting MWEs generally and not simply at predicting the contents of a single evaluation set. The syntactic variation measures perform better than frequency over both dictionaries for samples of 1000 and 5000. The good performance against two data sets tells us that the performance does generalise beyond a single resource. Results that are higher than these frequency values at a significance level of $p < 0.05$ are shown

		Unweighted		Weighted		
	Freq	P,I,E & D	P,I &D	P,I,E & D	P,I &D	P,I,D &Freq
Top 100 items						
LDEI	8	6	10	10	9	13
SAID	10	8	10	10	8	12
ALL	13	11	15	-	-	17
Top 1000 items						
LDEI	4.8	7	**8.6**	**8.4**	**8.2**	**8.2**
SAID	7	5.6	6.7	6	5.9	8.2
ALL	9.2	9.4	11.1	-	-	**12.4**
Top 5000 items						
LDEI	2.5	**3.58**	**3.66**	**3.66**	**3.64**	**4.08**
SAID	3.12	3.4	3.3	2.9	3.06	**4.14**
ALL	4.42	**5.22**	**5.22**	-	-	**6.18**

Table 4.1: MiniPar precision

in bold. For the Longman dictionary, the precision achieved by the syntactic variation measure employing the three best performing features ("P, I and D") is significantly higher than that achieved when ranking with frequency for sample sizes of 1000 and 5000. For the SAID data the performance is not significantly better than frequency. We assume that this reflects a difference in the selection criteria of the two resources, with SAID containing more pure unlexicalised collocations than the Longman dictionary.

The results reported so far have all given equal weight to the different variation varieties. As we described in section 4.3.2 it is also possible to vary the contribution of the different features by applying different weights. We want these weights to reflect the usefulness of the feature in predicting whether an item is a good candidate for inclusion in the lexicon. We therefore set the weights using a list of dictionary MWEs (disjoint from the testing set) as a development set. We can try out different weights for the features, evaluating against the development set, and thus find the set of weights that performs best. We have two different dictionary lists at our disposal. We identified the items from the Longman dictionary that are not found in SAID and vice versa. We can then conduct two experiments, one using the Longman items as the evaluation set, with the SAID items that are not in Longman used as a development list to set the weights, and one evaluating against SAID with the non-overlapping Longman items

		Unweighted		Weighted		
	Freq	P,I,E & D	P,I &D	P,I,E & D	P,I &D	P,I,D &Freq
Top 100 items						
LDEI	14	18	21	25	24	15
SAID	21	14	17	12	16	17
ALL	28	24	18	-	-	25
Top 1000 items						
LDEI	6.6	8.9	**10.4**	**10.9**	**10.7**	**10.2**
SAID	9.1	7.7	9	7.3	7.7	9.9
ALL	12.2	12.5	14.2	-	-	**15.2**
Top 5000 items						
LDEI	3.24	3.98	**4.28**	**4.28**	**4.28**	**4.84**
SAID	3.86	3.36	3.56	3.38	3.36	**4.54**
ALL	5.56	5.46	5.86	-	-	**7.68**

Table 4.2: RASP precision

employed for development.

Trying out all possible combinations of weights is too computationally expensive. We therefore set them using an optimisation algorithm known as the "downhill simplex method" (Nelder and Mead 1965). This algorithm searches for the local minimum value of some function (in our case the number of items in the top 1000 items that is not in the dictionary), by starting with particular values for the weights (in our case 1 for each feature) each of which describes the dimensions of an *N* dimensional simplex (a kind of geometric figure), and varying these dimensions in different directions to move through different points of the simplex. The process terminates when the adjustments being made to the dimensions to decrease the function are less than a preset minimum. The positions along the different dimensions are then taken to be the locally best weights.

The precision scores achieved using weights can be seen in tables 4.1 and 4.2. We begin by setting the weights for the four different features of passivisation, event modification, internal modification and determiner variation. When we evaluate against the Longmans data, having set the weights using the non-overlapping SAID items, we find an improvement with both parsers at *n* equal to 100, 1000 and 5000. However, when we evaluate against SAID, having set the weights using the Longman dictionary, the

	FREQUENCY	P,I & D		FREQUENCY	P,I & D
1	take place	follow suit	26	give rise	fit bill
2	have effect	draw level	27	answer question	hold breath
3	shake head	give rise	28	take advantage	return home
4	have time	part company	29	make way	keep track
5	take part	see chapter	30	solve problem	set sail
6	do thing	give moment	31	make attempt	catch hold
7	make decision	open fire	32	see chapter	close point
8	have idea	run counter	33	draw attention	set foot
9	play role	take refuge	34	pay attention	do wonder
10	play part	clear throat	35	provide service	go hand
11	open door	speak volume	36	have child	bear witness
12	do job	please contact	37	take account	report profit
13	do work	leave net	38	make effort	take office
14	make sense	give way	39	close eye	lose count
15	have chance	see page	40	see page	tell difference
16	make use	catch sight	41	have interest	feel welcome
17	ask question	cite argument	42	have difficulty	see appendix
18	spend time	see table	43	make difference	form basis
19	take care	check watch	44	go way	come term
20	have problem	list engagement	45	have advantage	regain consciousness
21	take step	go bust	46	make contribution	thank goodness
22	take time	change subject	47	meet need	gather pace
23	take action	change hand	48	make mistake	bear relation
24	find way	keep pace	49	have experience	go shopping
25	have power	see paragraph	50	make point	return attention

Table 4.3: Top 50 RASP phrases

effect is less impressive. Using Minipar the precision improves with 100 and 1000 item samples, but decreases with a 5000 item sample. Using RASP the performance decreases for 100 and 1000 items samples, achieving only a negligible improvement for a 5000 item sample. We next examine the effect of weighting when we use just the three features of passivisation, internal modification and determiner variation. Here we see that with Minipar performance decreases at all samples sizes in both experiments. With Rasp there is a slight improvement with a sample size of 100 but a decrease in all other cases.

So far we have evaluated our method by comparing its precision with that achieved with frequency ranking. However, as we discussed above, the method is in fact aiming to extract a different kind of item from that aimed at by frequency ranking. The justification for the method is that it is obtaining a kind of item that is not available using frequencies. It is necessary then to look at whether the method is providing novel information or is simply duplicating items. Figure 4.3 shows the 50 top ranked candidates items according to frequency and to our syntactic variation measure. There is an overlap of only 3 items between the two lists. Indeed over the top 100 items of the two lists there is only an overlap of 6 items and over the top 1000 there is an overlap of only 98. For the minipar items there are overlaps of 5, 7 and 105 for the top 50, 100 and 1000 items respectively. This tells us that the measure we propose is pinpointing a different selection of items from those highlighted by frequency ranking.

While there is substantial difference between the kinds of items extracted with our method and with frequency, there is also some overlap. Presumably there are items that are both frequent and fixed. It will be interesting to see whether a ranking method which combines frequency information and fixedness can outperform the two methods in isolation. We test this by ranking our candidate list using frequency and using the most consistently well-performing syntactic variation measure in two separate runs, and then adding together the two ranks achieved using the two methods for each item. The items are then reranked using the resulting sums. When this ranking is evaluated against the dictionaries it gives the scores recorded for "P,I,D & Freq" in tables 4.1 and 4.2. This gives better performance (over the combined dictionaries) than both methods for samples of 1000 and above and crucially a significantly higher precision than frequency over all evaluation items. The reason for this higher score seems to be that it identifies items that belong in the lexicon on the grounds of both frequency and syntax. An analysis shows that the top 1000 items includes 216 of the 1000 most frequent frequency list and 457 of the 1000 most syntactically fixed items. There are, then, 327 items in the combined top 1000 that would not appear using the two methods individually. These, we can assume, given the significantly higher score we obtain when we combine the ranks, are items which are not among the most frequent of the most syntactically fixed, but because they display considerable fixedness and they regularly occur are significantly marked items and consequently belong in the lexicon.

4.5 Comparison with collocation extraction

The last section described a method for identifying syntactically fixed units in corpora. We evaluated this method by measuring the extent to which it could be used to highlight items in two dictionaries. Since there is no work on the identification of syntactically fixed verb phrases to which we can compare our method, we chose to compare the method with frequency ordered lists, as a method of ranking that is available to lexicographers. In fact there is a large literature describing methods for identifying institutionalised phrases, which aims to provide a motivated alternative to raw frequency. As these methods are the current best performing techniques for identifying multiword phrases, we are going to compare our technique with these, looking at both the relative levels of performance and the degree of overlap in the phrases identified. In order to claim that our technique is useful, we will need to show that its performance is as good as or better than such methods, and that it is not duplicating information that is already available through the use of these techniques. Section 4.5.1 will describe the previous work on the identification of collocations. Section 4.6 will describe an experiment to examine how well these techniques perform on the task to which we put our method above.

4.5.1 Previous work on collocation extraction

The idea of collocation is intuitively simple. Collocations (or institutionalised phrases) are those forms that are characteristic of usage. What is not straightforward, however, is defining what exactly is meant by characteristic. The most obvious way to define it is in terms of frequency - a characteristic word combination is one that occurs a lot. If we take this definition, then the task of identifying collocations in a corpus is very simple. However, this is not satisfactory for all purposes. If we go ahead and use this technique to extract all multiword sequences from the British National Corpus as I discussed in chapter 1 we find that the top three two-word, three-word and four-word strings are *of the*, *one of the* and *the end of the* respectively. As Béjoint (1989) points out in the passage that we quoted at the beginning of this chapter, it is not always clear that these are the kind of units we want in our lexicon.

Much work, then, has been done on filtering extracted phrases to create lists of useful collocative units. A great deal of this has been concerned not with the creation of general lexical resources in mind, but rather the extraction of terminology for technical glossaries. However the work on general vocabulary and technical terminology share interests and tend to be discussed as a single literature, and so I will discuss them together here.

The largest body of work on extracting collocations has focused on finding better statistical methods than pure frequency for extracting significant units. The intuition behind this is as follows. The goal is to identify words which are characteristically seen together. Frequency seems a good way to identify these. Take a word combination like *of the*. This is very frequent string as we saw. However *the* is the second word in eight out of ten of the most frequently occurring two-word sequences in the British National Corpus. Taken independently, *the* is the most frequent word in the corpus, and *of* the second most frequent. In only 25% of its occurrences in the corpus is *of* followed by the word *the*. On the other hand, take an expression like *hermetically sealed*. This occurs 19 times in the written part of the British national corpus. This is not a very high frequency if taken in isolation. However the word *hermetically* occurs only 22 times in the corpus, meaning that in 86% of cases the word *hermetically* is followed by sealed. We might therefore say that the word *hermetically* is characteristically followed by the word *sealed*.

We are not interested in what is characteristic of a particular word, but rather in what is characteristic of the language as a whole, so perhaps a better way to think about this is to say that given the individual frequencies of *the* and *of*, they do not occur together any more frequently than we would predict they would according to chance. Chance, however, cannot account for the fact that *hermetically sealed* occurs as frequently as it does. Thinking of this cognitively, then, their frequent cooccurrence is not sufficient for us to posit that a phrase like *of the* is particularly salient for speakers, as the frequency of the combination can be accounted for simply in terms of the frequency of use of the component words, whereas in order to account for *hermetically sealed* we must posit that they have some special relationship. This is the intuition behind the work that has been done on finding statistically significant word cooccurrences from corpora. I will describe four methods that have been particularly widely used (for basic information on an impressively exhaustive collection of methods see Pecina, 2005).

The earliest publication to describe a method that has been widely used for collocation extraction was in fact aimed at the problem of identifying characteristic associates of words for the purpose of describing usage preferences for individual word dictionary entries. Nevertheless Church and Hanks (1990) has often been applied to phrase extraction and so will be described here. Mutual information is a measure from information theory of the reduction of uncertainty (entropy) about one random variable that results from knowing another (See chapter 2 of Cover and Thomas, 1991 for full details). This is a measure defined over whole distributions. Pointwise mutual information (which in the linguistics and computer science literature is often incorrectly referred to as simply "mutual information") is the reduction in uncertainty about one event that results from another event. If we treat two words x and y and their cooccurrence xy as events in the event space of the corpus, and calculate their probabilities using maximum likelihood estimation[2], we can calculate their pointwise mutual information $I(x;y)$ as follows:

$$I(x;y) = log_2 \frac{p(xy)}{p(x)p(y)} \tag{4.11}$$

This provides us with a measure of the information in bits that is gained about y by our knowing about x and vice versa. While this value in isolation might not be of a great deal of interest to the lexicographer, it allows them to rank word pairs according to their strength of association.

As I said before, the mutual information measure of association was suggested as a way to find characteristic associates for any given word. This has been its principal use by lexicographers, and it has been very successful. For the task of extracting significant phrases, however, while researchers continue to use it and some successful results have been claimed, it is less than perfect. As should be obvious from the equation, if the component words of the combination have a very low frequency, then even if the word combination occurs only once or twice, they are going to be given a high mutual information score. This means that the measure hugely overestimates low frequency

[2] This is calculated as follows:

$$p(x) = \frac{f(x)}{N} \tag{4.10}$$

where N is the size of a corpus and f(x) is the frequency of the word x in that corpus.

items. Because of the large number of very low occurrence words and phrases that we see in corpora, this is a significant problem.

The problem of distinguishing significant events from those that occur due to chance is the basic problem of hypothesis testing for which experimental statistics have provided us with many techniques. A simple and very widely used measure in this field is the t-test, and this has been adopted for use in collocation extraction. The t-test is a technique to evaluate whether a set of scores belongs to the same distribution or not (see chapter 6 of Hinton, 1995 for a straightforward introduction to the details). It has been used for collocation extraction then by using it to evaluate how likely it is that the count for a wordpair (the observed count) and the expected value belong to the same underlying distribution. Church *et al.* (1991) suggests a way to calculate this from a corpus as follows:

$$t = \frac{f(xy) - \frac{f(x)f(y)}{N}}{\sqrt{f(xy)}} \tag{4.12}$$

In principle the t-score gives us the difference between p(xy) and p(x)p(y) in the number of standard deviations. There is a problem here however. The t-test works on the assumption that the values are taken from a normal distribution, and of course we know that word frequencies do not follow a normal distribution. We cannot then take the t-score as meaningful in isolation. However as a method of producing ranked lists of collocations some relatively good results have been reported (see below).

The practical and theoretical problems with MI and t-score discussed above have continued to trouble researchers, and many other scores have been suggested. One test that is widely used in experimental statistics and which doesn't assume a normal distribution is the χ^2 test. This test is commonly used to examine frequency data, and decide whether an observed set of frequencies fits an expected pattern of frequencies. This is calculated for a given word pair as follows. We use the corpus to populate a 2x2 contingency table for a given word pair. We then calculate the following:

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{4.13}$$

where each O_{ij} is the observed count in the cell i, j, and each $E_{i,j}$ is the expected count for that cell, which we calculate by taking the product of its two marginals and dividing by the sum of all the cells.

A final related measure that continues to be widely used is the log-likehood ratio as suggested by Dunning (1994), who pointed out that the χ_2 test overestimated the significance of low frequency events. Intuitively this can be thought of as the ratio of two likelihood functions, one representing the likelihood of getting the counts we get under the hypothesis that the occurrence of two words is independent, and the other the likelihood of getting these scores under the hypothesis that they are not. Again it can be calculated from a contingency table, with expected counts estimated in the same way, as shown in equation 4.5.1:

$$-2log\lambda = \sum_{i,j} O_{i,j}\, log \frac{O_{i,j}}{E_{i,j}} \tag{4.14}$$

Thus far all of the justifications I have outlined for the various methods have been theoretical. What of the performance of the different methods? There is a significant problem in evaluating performance because the quality of institutionalisation is without question a matter of degree. The approach that has most frequently been taken is the comparison with lists created through human intuition, either by extracting items from existing lexical resources or by compiling new lists specifically for the evaluation. The most convincing and often cited example of this is Evert and Krenn (2001). They extract lists of adjective-noun pairs and preposition-noun-verb triples from German corpora of 800,000 and 8 million word corpora respectively, and have two annotators decide whether each items is a collocation. The four association measures discussed above are then calculated and ranked lists are created for each measure, as well as for item frequency. The precision and recall scores are then calculated for the top n of each list. A simple way to measure performance would be to simply choose a value for n and compare performance. However as the relative performance for each n varies greatly, they instead publish graphs of the precision and recall against n from 1 to the size of the corpus. This is then used to discuss the results.

For the adjective-noun combinations the best overall performance is achieved by the log-likelihood ratio, only narrowly beating the t-test which is followed in turn by frequency, χ_2 and mutual information. A χ_2 test is performed at various values of n to assess whether the difference between the scores is statistically significant. The difference between log-likelihood and the t-test is never significant, and log-likelihood only beats raw frequency at around $n = 1000$.

For the preposition-noun-verb combinations (with the association measures examining the relationship between the preposition and the verb) the t-test performs best, followed by frequency and then the log-likelihood ratio, then χ_2 and then mutual information. According to the χ_2 measure of significance, the t-test is significantly better than log-likelihood at various n, but is not significantly better than frequency. Somewhat surprisingly, contrary to the theoretical predictions, mutual information and chi-squared do not perform better for high frequency than low frequency items for either phrase variety.

4.6 Experiment six

The aim of this experiment is to evaluate the performance of the collocation measures described above on the task to which we put our syntactic-fixedness measure in experiment five, thereby enabling us to compare the usefulness of our methods with what are reported to be the best available techniques for identifying MWEs.

4.7 Materials

The materials for this experiment were the lists of verb plus noun object phrases extracted from two dictionaries that we used in experiment five, together with list of the same candidate items used in that experiment ranked using the four collocation measures described above.

4.8 Method

The procedure was as in experiment five.

4.9 Results

The precision scores achieved with lists of candidate items sorted according to the the four collocation measures, as well as lists sorted using frequency and the best performing syntactic variation measure and the combined rankeds of this method and frequency can be seen plotted in figure 4.3. The recall scores for the same can be seen in figure 4.4.

The best performing collocation extraction methods for both parsers are the t-score and the log-likelihood ratio, with MI and χ-squared performing very considerably

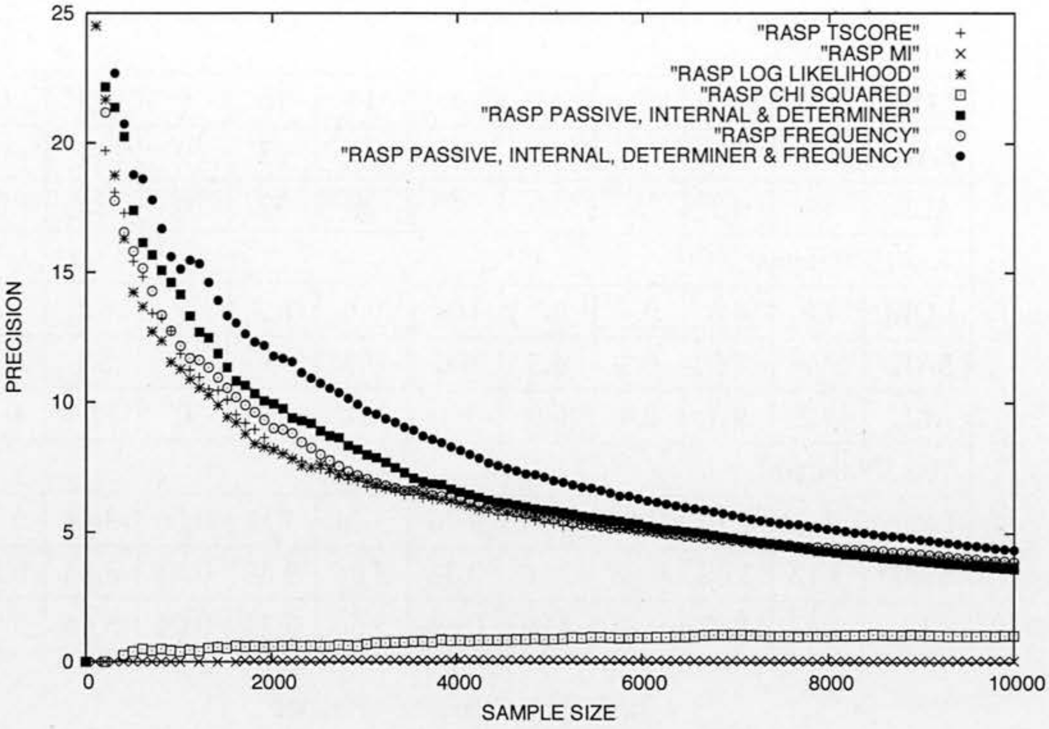
	MiniPar					Rasp				
	Freq	t	MI	LLR	χ^2	Freq	t	MI	LLR	χ^2
Top 100 items										
LDEI	8	8	2	11	4	14	16	0	13	0
SAID	10	12	1	11	3	21	23	0	17	0
ALL	13	16	2	16	4	28	32	0	25	0
Top 1000 items										
LDEI	4.8	4.8	0.4	4.7	0.9	6.6	6.3	0	6.5	0.3
SAID	7	7.2	0.3	6.3	0.6	9.1	9	0	8.1	0.2
ALL	9.2	9.4	0.4	8.8	1.1	12.2	12	0	11.4	0.4
Top 5000 items										
LDEI	2.5	2.76	0.12	2.94	0.46	3.24	3.12	0.06	3.44	0.58
SAID	3.12	3.08	0.08	3.26	0.32	3.86	3.68	0.04	3.86	0.54
ALL	4.42	4.62	0.14	4.82	0.62	5.56	5.34	0.04	5.66	0.88

Table 4.4: Collocation Results

worse. The best score for low values of n is t-score, with log-likelihood overtaking for larger values, and the best performing collocation measures often give a performance that is only equal to and often worse than raw frequency. This is consistent with the results reported by Evert and Krenn (2001). The scores for these measures as well as for frequency ranking at various sample sizes can be seen in table 4.4. None of the collocation scores perform significantly better than frequency. It has been reported that various collocation measures give unreasonably high scores to low frequency items and that it is useful to apply a frequency cutoff. We experimented with various cutoffs up to an occurrence rate of 5. We found that this led to no overall improvement in the performance of the best collocation method with each parser, and crucially did not lead to any significant difference from frequency.

The important thing to note from these plots is that the syntactic variation method performs better than all the collocation extraction techniques. The combination of the syntactic variation and the frequency ranks of course also performs better, and at a level that is statistically significantly higher than the best collocation extraction technique at sample sizes of 1000 and 5000.

While they produce ranks that are different from pure frequency, the collocation measures are still based on relative frequencies and are aiming at the identification of



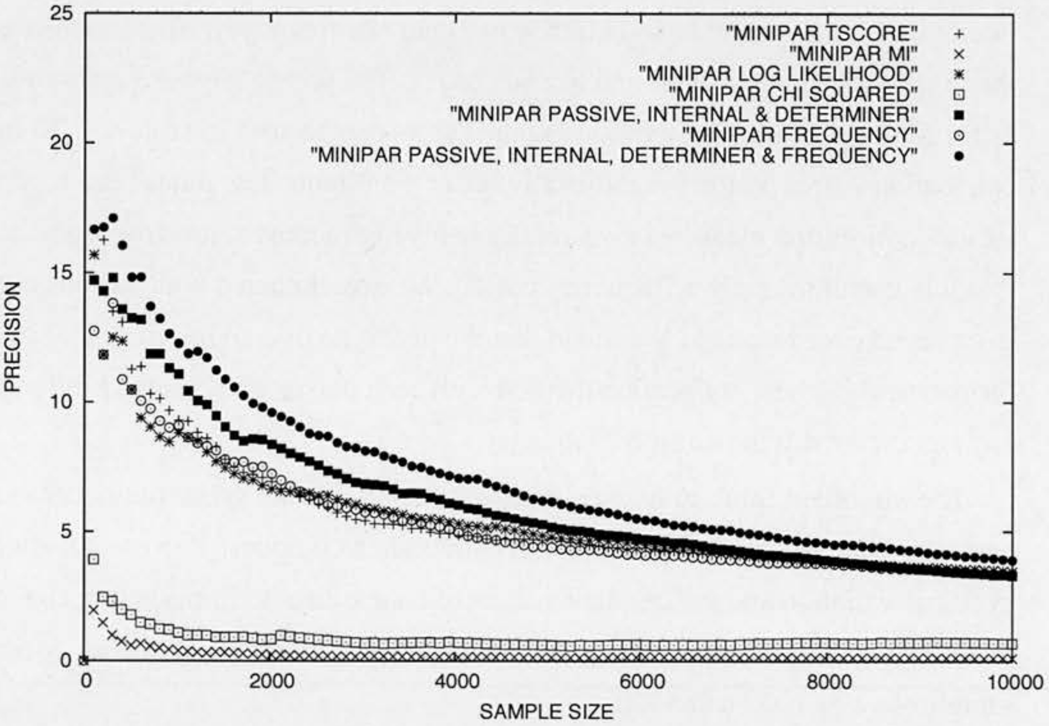
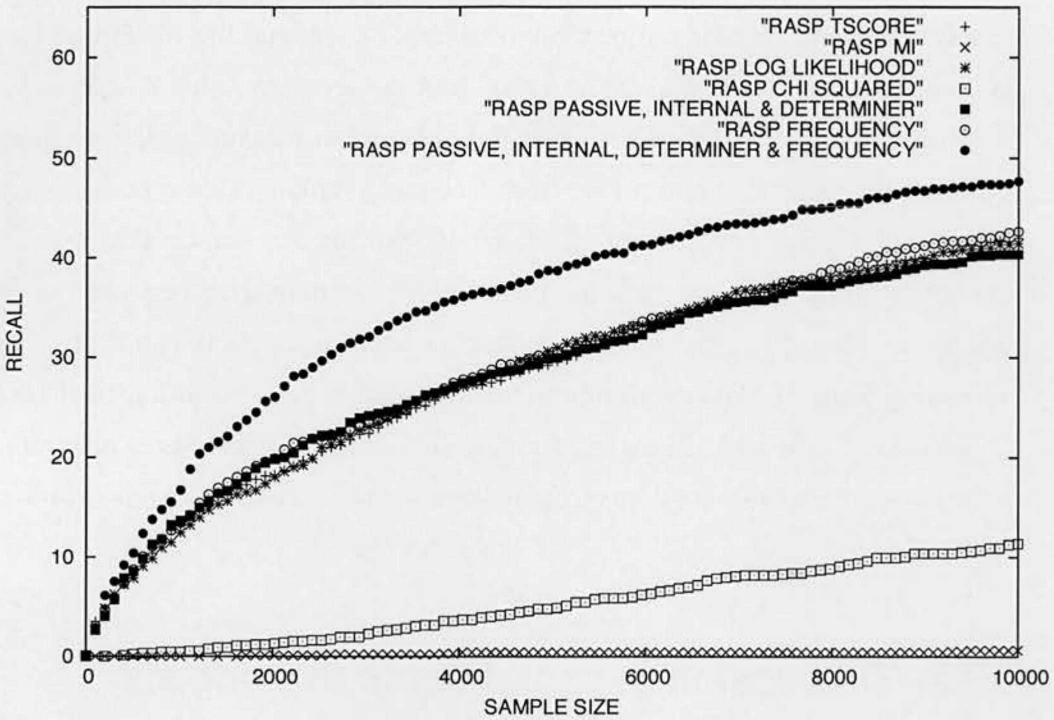


Figure 4.3: Precision score by sample size



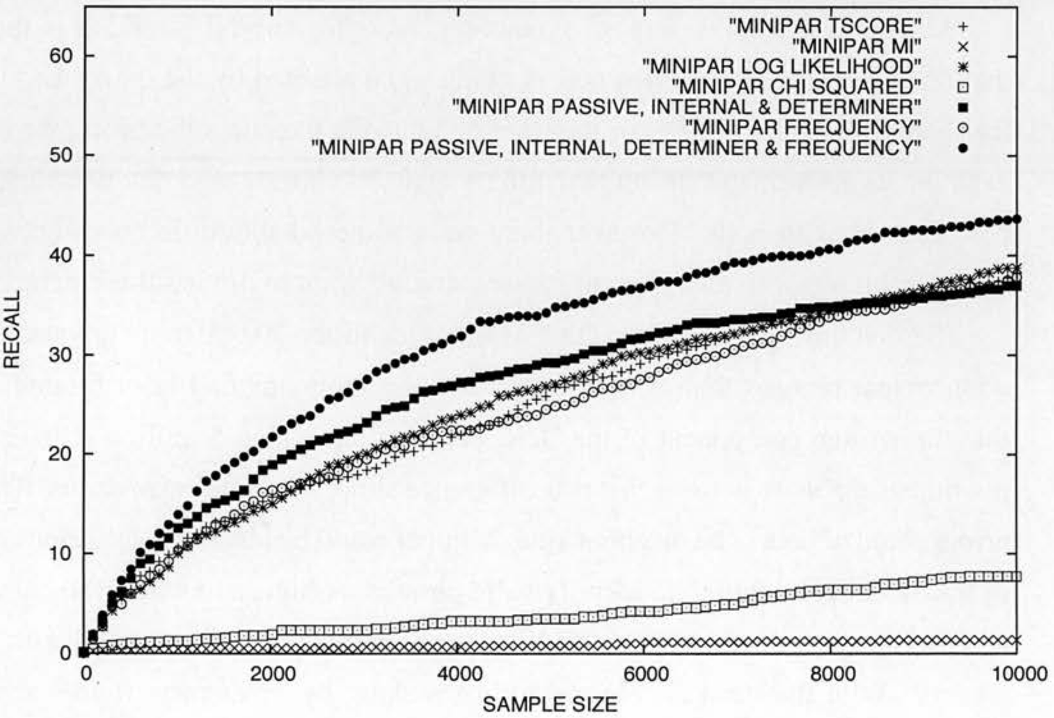


Figure 4.4: Recall score by sample size

institutionalised phrases. It will be interesting to investigate the extent to which the items they identify overlap with frequency and with our syntactic fixedness measure. The two high-performing collocation measures, t-score and log-likelihood have overlap with frequency of 795 and 624 out of 1000 respectively using RASP, and 552 and 414 using Minipar. This tells us that the collocation measures are significantly duplicating the information available from frequency ranking.This is compared with the overlap with frequency ranking of 98 and 105 for the two parsers that we noted with the three feature syntactic variation measure. The item overlap between t-score items and those extracted using the the syntactic variation measure is 116 using RASP, and 120 using Minipar. The overlap between syntactic variation and log-likelihood items is 108 using RASP and 125 using Minipar. This small overlap tells us that our measure is extracting very different items from the collocation measures, as it is from frequency ranking.

4.10 Discussion of experiments five and six

In this chapter we have introduced a measure of syntactic variability, and argued for its utility in identifying examples of one kind of MWE. The two experiments that we have described have provided evidence of its usefulness.

When using general purpose syntactic parsers for any NLP task, it is inevitable that one's performance on that task is going to be affected by the quirks and flaws of that parser. In order to test our measure of syntactic fixedness therefore, we chose to examine its performance using two different parsers. We showed that it achieved good performance with both. However there were some notable differences between the results achieved with the different parsers, and we want to discuss these here.

The first difference to note is that Minipar identified 30,000 more unique verb plus noun object phrases than RASP, even once the pronouns had been filtered. Given that the written component of the BNC contains more than 5 million sentences, it is not unreasonable to assume that this difference simply reflects parse errors. The parse errors could of course be on either side. Minipar could be identifying spurious phrases, or RASP could be failing to identify valid phrases. A hint as to where the balance lies can perhaps be gained by looking at the precision rates that are achieved with the two parsers. With the most simple method of ranking by Frequency, RASP achieved a higher rate of precision than Minipar. While this is of course not a measure of parse accuracy, if the parsers were identifying identical phrases the precision achieved would be the same. and one possible explanation for the difference could be parse errors from Minipar.

Given the different performance of the two parsers in identifying verb and noun object pairs, we would expect there to be some difference in the performance with all measures irrespective of how good the parsers were at identifying the relevant forms of syntactic variation. What we see in fact is that the difference in performance we achieve with the two parsers for the syntactic variation measure is even greater than in the frequency ranking. Perhaps the cleanest way to compare performance is to look at the scores achieved with each of the individual variations. When we add filters, as with the determiner variation, Minipar achieves the same level of performance as RASP. For all the other features, however, we see RASP performing better with up to an 8-point difference. The purpose of using two different parsers was not to compare their accuracy, but rather to check that our model is affective across different component technologies. That this is so should be clear from the fact that the measure outperforms frequency and all the collocational measures for both parsers.

In section 4.1.1 we provided the motivation for our choice of features. Although this was based on linguistic data, the idea that they would be useful in predicting lexicon membership was an unproven intuition. The experiments we performed with different combinations of features and with the weighting of those features affords us some insight into the usefulness of each feature. We found that in isolation and in combination with other features, event modification was of very limited use in deciding whether a phrase is lexicalised. When we used this feature in combination with the others, weighting of all features outperformed equal combination, because it allowed the contribution of this feature to be greatly reduced. Once we removed this weak feature we found that the weighting strategy in fact reduced the performance of the variation measure. There could of course be weaknesses in the implementation of weighting. It is possible, for example, that the differences between the two dictionaries reflects some significant linguistic distinction that means using one as a development set and the other for evaluation is problematic. It could also be that better results could be achieved with a different optimization algorithm. However the fact that we obtain better results when weighting over four features, but not when we remove the weak feature suggests that the weighting strategy is valid. That weighting with a development set fails to perform better than equal combination in all cases suggests that equal combination, while not always optimal, is in general a reasonable strategy.

Although the aim of these experiments was to evaluate the usefulness of our syntactic variation model for extracting one kind of MWE, they also provide wider insights into the overall strategy that should be taken to obtaining phrasal lexica from corpora. As should have been evident from our discussion here, much work has been done on the identification of significant collocations in corpora, with very little effort being directed at the extraction of other kinds of phrases. Only in the last five years have large scale evaluations such as that of Evert and Krenn (2001) been performed. Such work suggests that despite a decade and a half of intensive research there is no unassailable evidence that statistical tests of collocation are better than raw frequency at identifying MWEs. In section 4.5.1 we described how simply ranking with frequency prioritises arguably uninteresting sequences such as combinations of common function words (e.g. *of the*). However most of these are removed from consideration simply by employing some kind of syntactic processing. Justeson and Katz (1995) part-of-speech tag their data and define a series of tag sequences that cover phrase types that are of interest. Given a tagger of high accuracy, this approach is obviously very effective at ruling out non-constituent chunks. A number of researchers have used this approach to great effect in extracting phrases of particular interest (e.g. Baldwin and Villavicencio, 2002, Lapata and Lascarides, 2003b). The relatively high performance achieved in this chapter and in Evert and Krenn (2001) using raw frequency ranking supports the claim that this can be effective. One problem with this approach is that if the goal is exhaustive coverage of characteristic phrases it is simply not possible to predict what kinds of chunks are going to be of interest. However this problem is addressed by Dias (2003) who describes a technique for automatically extracting tag sequences of interest.

What all this suggests is that developing ever better methods of identifying significantly associated word pairs may not be of great help in the creation of multiword lexica. The techniques do not seem to work significantly better than raw frequency when evaluated on lexical acquisition tasks. Furthermore it has never been shown, for example, that strength of association is a better predictor of the kinds of chunks that people store than simple raw frequency. In fact there is some research (e.g. Lapata *et al.*, 1999) that suggests exactly the opposite. I am arguing in this thesis that a more diverse approach is necessary, taking a different approach to different kinds of MWE. The results described in experiment six suggest that raw frequency is as good a measure of institutionalisation as any of the measures of association that have been suggested in the literature (in this it is largely in agreement with other evaluations that have been performed). We have also described a linguistically motivated measure for approaching a different variety of MWE that not only pinpoints a larger number of the phrases in our evaluation set than any of the collocation extraction measures but also crucially has far less item crossover with frequency. This tells us that it is extracting information that is not available from raw frequency ordering. We therefore conclude that while efforts to create better measures of lexical association have until now dominated the field of MWE extraction, if the goal is to obtain new information, efforts would be more productively spent developing methods for identifying other varieties of MWE.

4.11 Chapter summary

In chapter 2 we introduced the phenomenon of syntactic fixedness as one of the dimensions along which phrases become lexicalised, and as defining one of the categories of phrases that need to be included in the lexicon in order to effectively account for linguistic performance. In this chapter we presented evidence that information about the syntactic fixedness of phrases can be obtained from corpora using an automatic syntactic parser, and that this information can be used to rank lists of candidate phrases for inclusion in lexicons in a way that captures more valid MWEs than either frequency ranking or various association measures that have been described in the literature.

In experiment five we used our measures to rank lists of verb and noun object phrases. We then evaluated by using a list of MWEs taken from two published dictionaries to measure the precision and recall of our measures for the top *n* items for various values of *n*. This evaluation was used because it usefully approximates the situation of the lexicographer needing to choose which items in a candidate list to prioritise when considering items for inclusion in a lexicon. We evaluated the performance of measures of fixedness on four different syntactic dimensions - the passivisation of the phrase, the addition of adjectival modifiers of the noun, the addition of adverbial modifiers of the verb and the addition, variation or dropping of determiners - and of various combinations of these features. We found that a combination of measurements of passivisation, adverbial modification and determiner variation produced the best results, which were equal to or better than frequency ranking for most values of *n* and when the measure was implemented using two different parsers. We found that when the ranks produced with our measures was combined with information from frequency ranking the results achieved were better than that achieved by either method in isolation.

In experiment six we evaluated the performance on the evaluation described in experiment five of various collocation measures that have been described in the literature as useful for MWE identification. We found that none of them performed significantly better than frequency ranking, and that they performed worse than our syntactic variation measure at a level that was often statistically significant. Furthermore, we found that the items extracted with such measures significantly overlapped with those items identified by frequency ranking, while our syntactic variation measure highlighted a set of phrases that had very little overlap. We presented this as evidence that our measure is providing novel information that is not available using existing techniques. Finallly we used this as supporting evidence for an argument that the study of MWE extraction would progress more successfully if it focused on identifying the full range of varieties of MWE rather than concentrating on collocation extraction as it has primarily done to date.

Chapter 5

Using lexical context to detect non-compositional units in corpora

The two preceding chapters have focused purely on linguistic form. In chapter 3 we discussed the extraction from corpora of information about the frequency of sequences of words. In chapter 4 we discussed the identification of phrases that have a single fixed form that is strongly preferred over all other forms. In this chapter, by contrast, we are interested in acquiring information not about linguistic form but about linguistic function.

In chapter 2 we introduced the phenomenon of the non-compositional phrase. We noted that a substantial part of human linguistic communication involves the use of phrases and sentences that the listener will not have heard in their entirety before, and that the hearer can recover information about the meaning of such novel utterances by combining the meanings of the component words and/or phrases. We then noted the existence of phrases in English that cannot be so analysed. For example while most speakers of English understand the meaning of the phrase *bite the bullet* to be "to accept something unpleasant and continue", the meaning of the phrase has nothing to do with the meaning of the component words *bite* or *bullet*. There is no way to interpret the sentence literally that would give this meaning. Furthermore it is not easy to see how the conventional meaning could be arrived at by any process of pragmatic or "figurative" interpretation. In order to explain the successful use of this phrase in human communication it is necessary to posit that speakers store some kind of form-meaning mapping for the whole phrase. Clearly any language description that is able to account for language use must put such phrases in the lexicon, or in some component of the grammar. This chapter will explore a technique for offering automatic support for the task of identifying which examples of a particular kind of phrase have a non-compositional meaning.

5.1 Verb-particle constructions

The phrase variety that we will focus on in this chapter is the verb-particle construction. Verb-particle constructions (hereafter referred to as VPCs) consist of a head verb and one or more obligatory particles. Examples of VPCs are *put up*, *finish off*, *gun down* and *make out* as used in the following sentences

(26) Peter put the picture up

(27) Susan finished her paper off

(28) Philip gunned down the intruder

(29) Barbara and Simon made out

Each of these examples includes a prepositional particle (although see O'Dowd, 1998 for a discussion of the complexities of this category). In fact there are some candidates for VPC status that include adjectives (*cut short*) or even verbs (*let go*) in the particle role. However, discussion of VPCs has been almost exclusively limited to the prepositional variety and I will follow this convention here.

The VPC is not the only phrase variety in which a verb is paired with a prepositional form. Other kind of phrases involving such a combination are prepositional verbs as represented by the phrase *refer to* in sentence 30, or simply free verb-preposition combinations as represented by *go into* in sentence 31:

(30) That song refers to the war

(31) It was time for me to go into the examination room

VPCs can be distinguished from such phrases on the basis of their syntax, using a number of tests. These are as follows:

1. In transitive usage VPCs can appear with the particle either before (*Peter put up the picture*) or after (*Peter put the picture up*) the object. This is not the case for other verb and preposition combinations (**That song refer the war to*; **It was time for me to go the examination room into*).

2. In transitive VPC usage, pronominal objects must occur between the verb and the particle (*Peter put it up* and not **Peter put up it*), whereas in other verb

and prepositional combinations, even pronominal objects always occurs after the preposition (*Peter put it on the table* and not *Peter put on it the table*).

3. In other kinds of verb and preposition combinations the preposition can occur at the beginning of sentences when forming questions (*To what did that song refer?*; *Into what room was it time for you to go?*), but not for VPCs (*Up what did Peter put?*; *Down who did Philip gun?*).

4. Unlike other verb and preposition combinations, VPCs do not allow manner adverbs to occur between the verb and the particle. So for example one would not say *Peter put the picture quickly up* or *Susan finished quickly up her paper*, while one would say *that song refers repeatedly to the war* or *it was time for me to go quietly into the examination room*. There seem to be a small number of non-manner adverbs that can in fact occur in this position. Two examples are *back* and *right*. We can say *Peter put the picture back up* or *Susan finished the paper right up*.

We are interested in VPCs here because they frequently have meanings that cannot be recovered through the simple composition of their independent parts. Compare, for example, sentences (26) and (29). In (26), the meaning seems to be that Peter *put* the picture somewhere and that as a consequence the picture was *up*. That is, the verb and the particle make independent semantic contributions to the sentence. A (partial) event-based semantic analysis of this might be as follows :

$$\mathrm{put}(e1, x, y) \wedge \mathrm{result}(e1, e2) \wedge \mathrm{up}(e2, y) \wedge \mathrm{peter}(x) \wedge \mathrm{picture}(y)$$

If we take (29) we see a rather different situation. Neither Barbara nor Simon can be said to have *made* or to be *out*. The semantic analysis we would want then might be something like the following:

$$\mathrm{make\_out}(e1, e2) \wedge \mathrm{and}(e2, x, y) \wedge \mathrm{barbara}(x) \wedge \mathrm{simon}(y)$$

How are we to identify whether the first or the second kind of analysis is appropriate for any given item? If we look at the other two sentences we can see that the problem is even more complicated. In (27) it is the case that the paper is finished, but it would be hard to claim that anything or anyone is off. Only the verb then seems to be contributing its simplex meaning, and the semantic analysis is:

$$\mathrm{finish}(e1,x,y) \wedge \mathrm{susan}(x) \wedge \mathrm{paper}(y)$$

In (28), by contrast, it is the particle that contributes its simplex meaning and not the verb. As a consequence of Philip's action the intruder is *down*, but since there is no simplex verb *to gun*, we would not say that anyone *gunned* or *was gunned*. The semantic analysis is consequently as follows:

$$\mathrm{gun\_down}(e1,x,y) \wedge \mathrm{down}(e2,y) \wedge \mathrm{result}(e1,e2) \wedge \mathrm{philip}(x) \wedge \mathrm{intruder}(y)$$

In chapter 2 we introduced the idea that a sequence of words is compositional if its meaning can be recovered from the meanings of its component words. Compositionality is assumed throughout this thesis to require that the component words have a meaning in the phrase that they can have freely across other utterances and phrases. With VPCs, however this assumption faces a potential challenge that it will be useful for me to address here for the sake of clarity. In the example of *finish off* above, I argued that the word *off* was not contributing an independent meaning. It is often argued, however, that the element is contributing some part of the phrase's meaning. Across multiple VPCs the particle *off* appears to contribute to meaning in terms of aktionsart. The addition of the particle *off* to the verb *finish* seems to indicate that the act of finishing was successfully completed. The same particle makes a similar contribution to a number of other VPCs such as *boil off* in which boiling is performed until the object has disappeared, and *cool off* in which cooling takes place until the entity is *cool* (or calming takes place until the person is completely calm).

While the particle *off* can contribute a similar completive meaning to multiple phrases, this sense is unique to VPCs[1] and consequently it would not be appropriate to have a separate entry for *off* in the lexicon that allowed it to freely combine. Of course one might argue that a speaker of English would be able to interpret the meaning of *cool off* using knowledge of *cool* plus knowledge (furnished from other

[1] It has been argued that this is not the case with the particle *up* in phrases such as *finish up* or *dry up* where *up* is said to have a similar completive meaning. This is because *up* seems to be able to take the form of an adjective with a completive meaning in the non-VPC context when combined with a copula. For example one can say that the game or the time is *up*. However the meaning of *up* in those cases is again specific to a particular construction, and cannot be freely used.

phrases) that within VPCs *off* contributes a completive meaning. However, it seems to be a reasonable requirement of compositionality that each meaningful component has a meaning that is independent of the presence of the other components, and this is not met here. Furthermore in order to offer a satisfactory description of the language in which *off_completive* is treated as a separable element, we would need to clarify exactly which verbs *off_completive* can combine with, for which no satisfactory account has yet been offered. For reasons of theoretical consistency and descriptive simplicity, therefore, we restrict our definition of compositionality to those phrases whose meaning can be derived from single word lexical entries, and treat cases such as *finish off* as non-compositional.

5.2 Identifying non-compositional VPCs

As with all phrase varieties, there are a variety of reasons why we might wish to put a VPC in the lexicon. One reason of course is simple frequency. According to Baldwin (2005) the most frequent VPC *set up* occurs at a rate of one per just over 14,000 words in the BNC. Many VPCs are simply an established way of communicating about a particular kind of event. To avoid using a VPC like *break down* in an utterance like *my car broke down*, or *take off* in an utterance like *the plane took off* would involve considerable circumlocation. And indeed there are many VPCs found in dictionaries that seem to have a fully compositional meaning and yet take their place in the lexicon because of their prominence in the language. Examples are *take away* and *fall down*. Such phrases can be identified in a corpus by their frequency.

Frequency is, however, not the only reason for including a VPC in lexicons. If a language description is to account for real language use, then non-compositional VPCs must be included regardless of their frequency. And indeed many non-compositional VPCs occur infrequently. Many VPCs that are non-compositional in some way such as *clock up*, *hit up*, *mull over* or *chill out* occur as few as 5 times in the BNC. Clearly such items cannot be identified on the basis of their frequency alone.

Baldwin (2005) identifies more than 7000 unique VPC items in the BNC using syntactic tags. Using a conservative model of VPC productivity created by employing existing linguistic resources and then validating the resulting items by searching for occurrences on the world wide web, Villavicencio (2005) posits the existence of 22,078 distinct VPC types. The majority of these will have compositional semantics, and we would not want to put them all in the lexicon. It would be very helpful then, to have some automatic technique for distinguishing which of the attested items have a non-compositional meaning and consequently need to be given a special lexical entry.

This chapter will explore one possible technique for doing this. The following sections will introduce this method. Section 5.2.1 will introduce the whole problem of computing meaning from corpora. Section 5.2.2 will look at how this might be usefully applied to the problem of detecting non-compositionality, and section 5.3 will describe the details of the model we are going to employ.

5.2.1 Learning about meaning from corpora

As I mentioned above, this chapter differs from the previous two in that it is concerned with acquiring information about linguistic function rather than linguistic form. Computers are very useful in the study of linguistic form. They are very good at identifying forms in large text collections, and of counting them and comparing them with others. Any aspect of language that is not directly apparent from its surface however, such as meaning, is less easy to acquire automatically from a corpus.

Fortunately linguistic tradition has provided us with a way of discovering aspects of meaning from surface forms in texts. The method I refer to is the so-called "distributional" analysis of words and phrases. The distributional method was crucial to the structural tradition in linguistics, and is central to key early works such as Bloomfield (1933). Harris (1964) sets out to describe how "each language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts (ultimately sounds) relative to other parts and how this description is complete without intrusion of other features such as history or meaning" (p.33). The distributional method was primarily used as an objective method for learning about the morphology or syntax of languages for the purposes of language description. However it was also noted that the distribution of forms could tell us about meaning. Harris (1964) writes that:

> "The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning differences. For it is not merely that different members of one class have different selections of members of the other class from which they are actually found. More than that: if we consider words or morphemes *A* and *B* to be more different in meaning than *A* and *C*, then we will often find that the distributions of *A* and *B* are more different than the distributions of *A* and *C*. In other words, difference of meaning correlates with difference of distribution." (p.43)

Meaning here then is something that is reflected in distributions. An early example of the rather more radical suggestion that the distribution of a word is not only a correlate of meaning but in fact one of the key dimensions of meaning was made by Firth (1957). He demonstrates this with reference to the word *ass*:

> "One of the meanings of the word *ass* is its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference. Even if you said 'An ass has been frightfully mauled at the Zoo', a possible retort would be, 'What on earth was he doing?' " (p.195)

He goes on to discuss other words, pointing out for example that "One of the meanings of night is its collocability with *dark*, and of *dark*, of course, collocation with *night*" (p.196). Firth argues that while distinct from the conceptual meaning of words collocation is nonetheless a crucial dimension of how words mean.

Analysing the collocation patterns of words over text collections is of course something that computers are very good at, and these ideas have been picked by contemporary researchers in corpus linguistics. Harris (1964) writes that the "the distribution of an element will be understood as the sum of all its environments. An environment of an elements *A* is an existing array of co-occurents, i.e. the other elements, each in a particular position, with which *A* occurs to yield an utterance". Exactly such distributions have found widespread use in modelling the meaning of words in computational linguistics. One example is in the task of word sense disambiguation. This is the task of distinguishing in which sense a particular occurrence of a polysemous word is being used. Some success has been achieved on this task using a distributional approach. The "distribution" (I will henceforce use this word in the sense of Harris, 1964) of particular meanings of words are learned from sense-labelled corpora, and the computer makes a decision as to which sense to assign to novel occurrences of this word by assessing which of the distributions the novel occurrence is closest to. A large range of methods have been used for modelling the distribution of senses and for deciding how to classify each novel occurrence (see Stevenson, 2003 for a survey of the field), however all the currently used approaches rely to some degree on distributional analysis. Another task in which the "distribution" of words have been used to model their meaning is automatic thesaurus construction. The task in this field is to use the context of words across large corpora to detect groups of words that have a similar meaning so as to automatically create thesauri (see Grefenstette, 1994 and Curran, 2003).

5.2.2 Compositionality and distributional semantics

We have seen then that the meaning of words is reflected in the lexical contexts in which they occur. We will now describe how this might be useful to us in detecting non-compositional units. Our hypothesis here is this: if a component word of a phrase is contributing a meaning that it can also have outside that phrase, then we would expect that the distribution of the phrase would be similar to the distribution of the word on the occasions when it occurs outside the phrase. If a phrase is non-compositional then we would expect it to have a very different distribution from its component words.

This chapter will examine whether this intuitive observation about compositionality and distribution can form the basis of a useful technique for automatically identifying non-compositional phrases in corpora. We saw above that we can usefully calculate the semantic difference between two words by creating a model of their distribution and calculating the difference between the two distributions. The technique described in this chapter will utilise such methods in order to compare the distribution of phrases with the distribution of their component words.

An early suggestion that the distribution of a phrasal verb[2] might be indicative of its compositionality came from Palmer (1965) who suggests that the collocational restrictions of a verb and particle might help to distinguish non-compositional "phrasal verbs". The topic is taken up by Berry-Rogghe (1974a). She notes that "an expression is said to be idiomatic when the meaning of the whole differs in meaning from the meaning of the parts separately" and suggests that "in collocational terms, this would be the case if the expression contracts a sufficiently different set of collocates from the set of each of its parts" (p.21). She explores this over a small text collection of 202,377 words by identifying the significant collocates of a single particle *in*, and of various phrasal verbs that include this particle (all of which are in fact prepositional verbs rather than VPCs), such as *interested in*, *live in* and *house in*. The collocation extraction is done using a technique described in Berry-Rogghe (1974b). Significant collocates are said to be those above a certain threshhold. She defines the degree of idiomaticity as the percentage of the number of collocates of the phrase that are collocates also of the particle *in*. She informally examines this and notes that, for example, the phrases *versed in*, *interested in* and *believe in* have no collocates in common with the particle, while *live in* shares the collocates *house*, *town*, *country*, *London*, *room*, *world* and *place*. While it is not true that any of the former phrases are non-compositional, the

[2] this term is commonly used to encompass all variety of regular verb plus preposition combination, and includes but is not restricted to VPCs

measure does seem to pick up on the fact that in the later phrases, the particle has its dominant locational sense. The study ends by concluding that larger corpora would be needed in order to really test the method.

A related work came 27 years later in Schone and Jurafsky (2001). This work described an attempt to improve upon various lexical association measures on the task of automatically identifying MWEs in a corpus. They took lists of multiword entries from a number of different lexicons and measured the performance of a variety of different lexical association measures on the task of extracting them. They then make a series of attempts to improve upon the performance of these measures. One of these measures works from a similar premise as that of Berry-Rogghe (1974a). They note that one important quality of MWEs is their non-compositionality, and hypothesize that since a compositional multiword unit should occur in similar lexical contexts to its component words, words which occur in dissimilar contexts are more likely to be a MWE. They quantify this similarity using Latent Semantic Analysis (LSA, see Deerwester *et al.*, 1990 for details). They measure the distance between the context of a candidate MWE and the weighted sum of the contexts of the component words, hypothesising that dissimilarity should indicate non-compositionality.

The arguments made for the technique are intuitively very sound. However they report that the method offers no improvement in extracting MWEs over existing techniques. While this result is not exactly encouraging, we do not consider the paper as evidence that the basic intuition is wrong. As we have repeated throughout this chapter, there are different varieties of multiword expression. Accordingly lexical resources will contain not only non-compositional phrases, but also many compositional items that are collocations or syntactically lexicalised. One might then not expect a measure that identifies non-compositional items to be better at predicting membership of such a resource than an association measure. Nonetheless one needs to be able to identify all varieties of MWE and the failure to beat the performance of lexical association does not negate the technique. For a proper assessment of whether the approach can predict non-compositionality, one would need to use evaluation materials that contained information about precisely that dimension.

5.3 The model

The aim in this chapter is to test the hypothesis that distribution is indicative of compositionality by using a large balanced corpus to extract the distribution of a number of VPCs and compare each in turn with the distribution of its component words. In order to do this we need to have a clear way to describe this context and to quantify similarity between two sets of contexts.

The work that we described above, particularly in automatic thesaurus extraction, provides us with a range of techniques for doing this. All techniques are alike in that they work by building arrays in which each entry is a measure of the words cooccurrence with a set of context words, either in the form of raw counts, association scores or a probability distribution. Regardless of the way this information is represented or manipulated, these descriptions of context are usually described as **context vectors**.

Our representation, then, is going to be a vector. First of all we need to decide how we are going to populate these vectors (what we are going to consider to be the context of our words and phrases). The context models employed in NLP can be loosely divided into those that define context over a word window, and those that utilise parser output and define context in terms of the syntactic dependencies of the target word. In this work we will employ a word window. This is motivated by the desire to have a rich context for the particles as well as the verbs. A parse of a sentence containing a VPC might provide us with useful dependencies for the verb. However particles are intransitive and will consequently not be assigned relations to any word other than the verb. It is also the case that in selecting context with a parser one is ruling out collocates that are indicative of meaning but not syntactically related to the word. As we are dealing with relatively rare events in our VPCs we want to make use of all the information we have available.

In our model, the context of a word token is said to be 5 words to the left and five words to the right of that word. For a VPC it is said to be 5 words to the left of the verb and the first five words to the right of the verb that are not the particle. This window is chosen because 5 words is a reasonable upper bound on the number of words that we will see appear between a verb and a particle in the split configuration. The context of a word type then is the sum of all of the context windows over the whole of a corpus. As with all experiments in this thesis, the counts were collected over the written component of the British National Corpus. Words can have multiple morphological variants, and computers are unable to make the connection between variants of the same form. This could potentially obscure connections between wordforms, and so we first lemmatised the corpus using Morph (Minnen *et al.* 2000).

We use these "context windows" to create vectors of cooccurrence counts for each phrasal verb and each component verb and particle. In order to remove uninformative dimensions from our vectors, the 50 most frequently occurring words in the corpus were excluded. This is based on the assumption that the most frequent words, such as determiners, will occur across all contexts and will not be characteristic of the usage of any of our VPCs. We also make the assumption that very infrequent words will provide uninformative dimensions in our vectors. There are in excess of 200,000 unique words (including numbers) observed over the context windows of our VPCs and their component words, and in order to make our model more computationally tractable we considered only the 2500 most frequent informative lemmas as context words (after we had excluded the 50 most frequent, i.e. we employed the 51st to 2550th most frequent lemmas overall). These were all words that occurred an average of at least once per document in the corpus (i.e. at least 3144 times over the whole corpus; see section 3.1.1 for further discussion of its make up). Our vectors recorded the number of times that each VPC or component word's context window contained each of this set of 2500 context words. [3] No further adjustment was made to the vectors.

Having built our vectors we next need a method for calculating the distance between them. Despite the considerable overlap, the task we are attempting here is different in various ways from those of word sense disambiguation and thesaurus extraction. Firstly, when one is comparing a phrase with its components, one is comparing two objects whose overall frequency is known to be very different. The phrases are guaranteed to have a lower frequency than the particle, and in cases where the verb is not unique to the phrase (i.e. not *gun*) they are guaranteed to have lower counts than the verb. This means that the phrase will have more sparsely populated vectors than the component words, either in terms of having more fewer observed collocates, or simply in having lower counts. As we want to acquire information about the meaning of the phrases rather than their frequency, we will not want the overall frequency to affect our measure of semantic similarity. We will therefore need a method that factors this frequency out of the comparison.

One further way in which our task is different from thesaurus extraction is that

[3] The number of unique words used to describe a word's context varies enormously across the literature from 500 words (McDonald 2000) through 1000 words (Schütze 1998), 70,000 words (Lund *et al.* 1995) to the full vocabulary (Pereira *et al.*, 1993; Curran, 2003). Patel *et al.* (1997) provide a plot of vector size against performance on a synonym detection task. They found that performance varied very little after the first 100 most frequent context words, rising very slightly to reach optimal performance at 2500 words. It is difficult for us to draw strong conclusions from any of this previous work as they are focused on different tasks. It is worth observing, however, that the distance measure we are going to employ is very little affected by the addition of low frequency words, and the addition or removal of any other than the most frequent context words produces only very small variation in our model. We therefore choose to test our hypothesis using the vectors described above.

in that task one is seeking synonyms so must only ever compare words in terms of their similarity to a single word. In the task we are attempting here however, we need all distance scores to be comparable and defined over a consistent space. It is desirable, for example, that any measure of the distance between two elements x and y, be symmetrical so that the distance of x from y be the same as the distance of y from x.

For these reasons we choose a measure that defines distance (or more accurately its converse, proximity) over a Euclidean space - the **cosine score**. While the requirements described about are not essential for thesaurus extraction they are crucial for the measurements of term and document distance used in information retrieval, and accordingly cosine is the most widely used measure of distance in that field.

The cosine score can be explained as follows (see chapter 5 of Widdows, 2004 for a detailed justification of this method). A standard measure of proximity in Euclidean space is the scalar product. This is simply the sum of the products of each pair of coordinates of the the two vectors. So if we have space with n dimensions, the scalar product will be calculated as:

$$a \cdot b = a_1 b_1 + a_2 b_2 + \ldots\ldots + a_n b_n \tag{5.1}$$

This would give us a measure of proximity. However, as we are simply adding together the products, vectors which both have high counts will obtain a higher overall score. We therefore want to factor this difference out. We can do this by normalising each vector by its length (its distance from the zero point). The length of a vector a can be calculated as follows:

$$\|a\| = \sqrt{\sum_{i=0}^{n} a_i \cdot a_i} \tag{5.2}$$

Rather than normalising each vector and then comparing them, we can in fact obtain the same result by calculating the scalar product of two vectors and then dividing by the sum of the lengths of the two vectors. The cosine score can thus be conveniently defined as follows:

$$cos(a,b) = \frac{a \cdot b}{\|a\| \|b\|} \tag{5.3}$$

This provides us with scores between 0 and 1 for each vector comparison that are symmetrical and comparable in all cases defined over the same set of context words (or *n*-dimensional space). We use this to calculate the similarity between each of our VPCs and each of their respective components.

5.4 Experiment seven

Having developed our model we are going to need a way to evaluate whether our model is predictive of VPC compositionality. In order to do this it is desirable to have a set of VPCs that have been coded for compositionality. One way to do this would be to ask a small number of linguists or lexicographers to make a judgement as to the compositionality of a set of items. However, we are claiming here that the compositional/noncompositional distinction is significant for ordinary learners and speakers of the language. In creating our resource, then, we are going to obtain judgements not from experts but from ordinary speakers. Our eventual aim in this is of course to create an evaluation resource which reflects the linguistic intuitions of this population. However, it will also initially allow us to confirm our claim that the distinction has reality for speakers, by measuring how consistent people are in making it. We do this at the level of each component word, so we have our subjects indicate whether each component word of each VPC is contributing compositionally to its meaning by answering simple and intuitive questions.

5.4.1 Subjects

A total of 121 adults participated in this experiments. They were all recruited by advertisements posted to newsgroups and mailing lists. They were split into 4 groups, which looked at sets of 40 VPC types each. All subjects declared themselves to be native speakers of English. For each set of 40 VPCs, then, there was a minimum of 28 subjects.

5.4.2 Materials

The principal material for this experiment was a set of VPCs. We wanted to be sure that these items are actually in common use and familiar to subjects, and so rather than using a dictionary to aqcuire the stimuli, we chose the items by randomly selecting 160 VPCs from a list of those VPCs found in the BNC with a minimum frequency of 50 by Baldwin (2005). We then extracted examples sentences including these VPCs from the BNC using the following automatic procedure, informed by the part of speech tagging provided with the BNC. The corpus was first lemmatised using Morph (Minnen *et al.* 2000) in order to allow us to recognise all morphological variants. We then searched for each head verb in turn. If the verb was found then we looked 5 words to the right to find the particle (5 words was the window searched within by Baldwin, 2005, based on a study of split VPCs reported in Baldwin and Villavicencio, 2002). A verb and particle token was accepted if there was no intervening verb, adverb or preposition. Five example sentences containing each VPC were randomly chosen to be presented as examples to our subjects.

5.4.3 Method

In order to discover whether our subjects interpret our set of VPCs compositionally, we made use of an entailment test first proposed for exploring the semantics of verb and preposition combinations by Hawkins (2000).

Entailment is conventionally defined for logical propositions, where a proposition *P* entails a proposition *Q* if and only if there is no conceivable state of affairs that could make *P* true and *Q* false. This can be generalised to refer to the relationship between two verb phrases V1 and V2 that holds when the sentence *Someone V1s* entails the sentence *Someone V2s* (see, e.g., the treatment of verbs in the WordNet hierarchy (Miller *et al.* 1990)). According to this generalisation we would then say that the verb *run* entails the verb *move* because the sentence *He runs* entails the sentence *He moves*. The same idea can be generalised to the relationship between simplex verbs (e.g. *walk*) and VPCs (e.g. *walk off*). For example, sentence (26) can be said to entail that *Peter put the picture somewhere* and so we can say that *put up* entails *put*. The same might be said of *finish off* and *finish* in (27). However, (28) and (29) produce a rather different result. (29) does not entail that *Simon and Barbara made*, and (28) cannot entail that *Philip gunned the intruder* because there is no simplex verb *to gun*. This is a very useful way of testing whether the simplex verb contributes to the meaning of the construction.

We can approach the relationship between VPCs and particles in this same way. For (26), while it is not true that *Peter is up*, it is true that *The picture was up*. We can therefore say that the VPC entails the particle here. For (27), it is not true that either Susan or her paper were off, and the VPC therefore does not entail the particle. In the case of (28), while it is not true that *Philip was down* it is true that *The intruder was down*, and the VPC therefore entails the particle. Finally, for (29), it is not true that

Barbara and Simon were out, and the VPC therefore does not entail the particle.

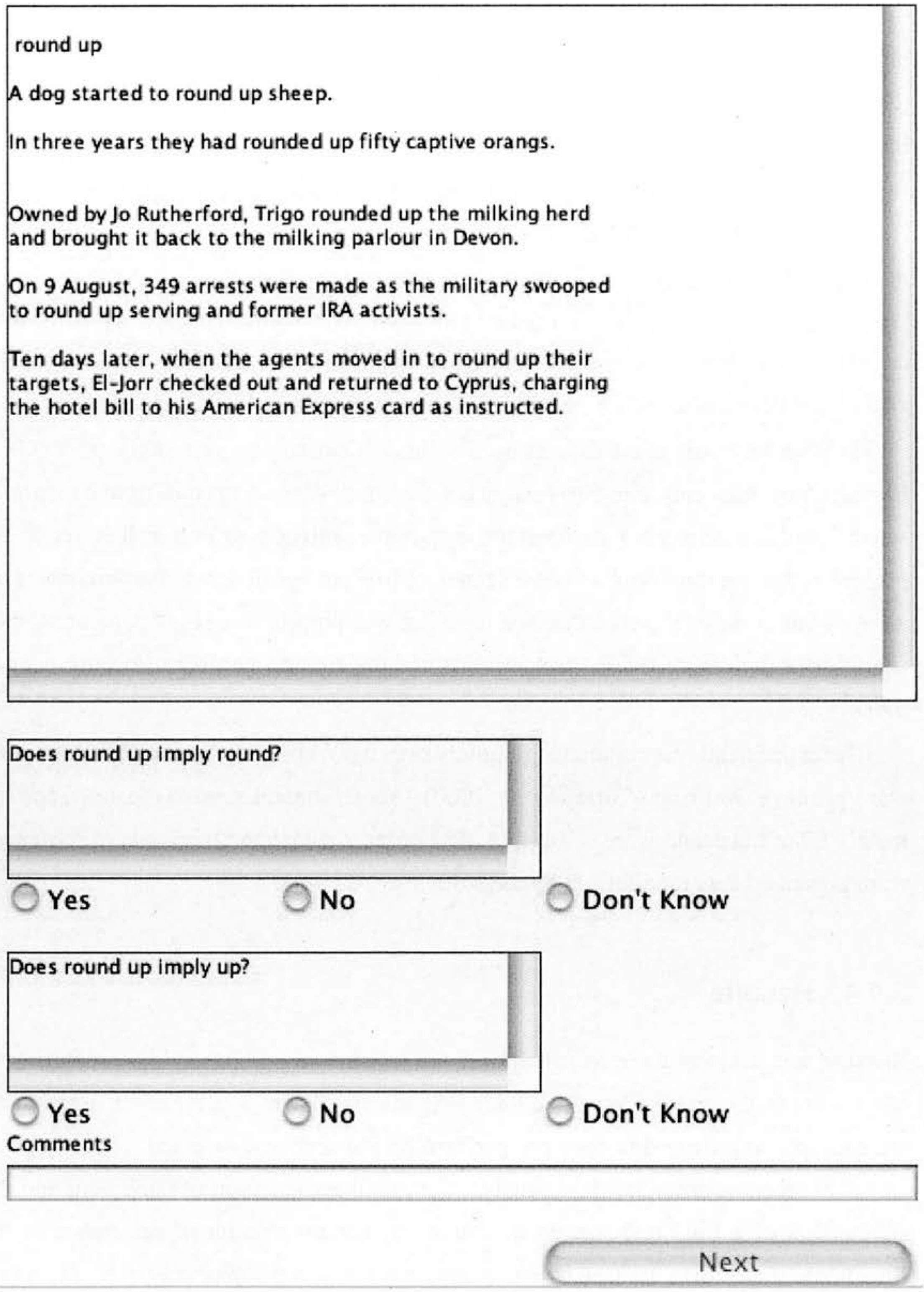
We used a version of this entailment test in order to elicit judgements from our non-experts. Each subject was presented with 40 sets of 5 sentences, where each of the five sentences contained a particular VPC. The VPC in question was indicated at the top of the screen, and they were asked two questions: (1) whether the VPC implies the verb, and (2) whether the VPC implies the particle. If the VPC was *round up*, for example, the subject would be asked "Does *round up* imply *round*?" and "Does *round up* imply *up*?", respectively. They were given the option of three responses: "Yes", "No" or "Don't Know". Once they had indicated their answer and pressed next, they advanced to the next VPC and set of 5 sentences. They were unable to move on until a choice had been indicated. An example of this interface can be seen in 5.1

As with any corpus-based approach to lexical semantics, our study of VPCs is hampered by polysemy, e.g. *carry out* in the *execute* and *transport out (from a location)* senses. Rather than intervene to customise example sentences to a prescribed sense, we presented the five randomly-selected sentences irrespective of sense. Participants were advised that if they felt more that one meaning was present in a set of sentences, they should base their decision on the sense that had the greatest number of occurrences in the set.

The experiment was conducted remotely over the Web, using the experimental software package WebExp (Corley *et al.* 2000). Experimental sessions lasted approximately 20 minutes and were self-paced. The order in which the forty sets of sentences were presented was randomised by the software.

5.4.4 Results

Because our subjects were recruited over the internet, we were unable to control the conditions under which they make their judgements. There are, however, some steps we can take to ensure that they are performing the task as instructed. The WebExp software records for us the time between the initial presentation of each item and the submission of a final response to that item. A certain amount of difference in the time taken to perform the task between items and between subjects is to be expected. However very short or very long average response times suggest that a subject is either not adequately considering their judgements or that their attention is distracted. In order to filter out subjects who are not committed to the task, we calculated the median response time for each, and discarded the top and bottom 5% of subjects. We have 121



round up

A dog started to round up sheep.

In three years they had rounded up fifty captive orangs.

Owned by Jo Rutherford, Trigo rounded up the milking herd and brought it back to the milking parlour in Devon.

On 9 August, 349 arrests were made as the military swooped to round up serving and former IRA activists.

Ten days later, when the agents moved in to round up their targets, El-Jorr checked out and returned to Cyprus, charging the hotel bill to his American Express card as instructed.

Does round up imply round?

Does round up imply up?

Figure 5.1: Experimental Interface

subjects, so we discard a total of 12 subjects in this way.

The judgements of the subjects for each item can be seen in Appendix G in tables G.1 to G.4. The judgements of one of the four subject groups (group four) as to whether each VPC implies its component verb can be seen in figure 5.2. The judgements of the same group as to whether each VPC implies its component particle can be seen in figure 5.3. Once subjects had been discarded on the basis of response time there were 24 subjects left in this group.

Informally we observe that the judgements are in line with our linguistic intuitions. The items for which all subjects are agreed in stating that the verb is implied are *sell off* (e.g. *The current thinking is to sell off freight services first*), *move out* (e.g. *Many of the townsfolk were waiting, ready to move out and make trouble for the enemy when the camp was roused*), *lift out* (e.g. *Lift out the leeks with a slotted spoon and set aside*) and *lie down* (*She went home to lie down*) . In all of these it seems that the verb has its dominant single word meaning. Four of the items that the largest group of subjects agree do not imply their component verbs are *carry out* (e.g. *Actually, the master criminal was carrying out his greatest coup, to murder and replace the world's most influential intelligence*), *wear on* (e.g. *As the night wore on, Lord Owen's team made their plans*), *carry away* (e.g. *Supposing one of them got carried away and hit the old boy too hard*) and *play down* (e.g. *Management at Monkton-hall last night played down the significance of Caledonian Mining's decision*). In none of these phrases does the verb seem to have a meaning that it can have outside of the context of the phrase.

For the subjects' judgements of the particles, we again see judgements that agree with our linguistic intuitions. Two of the four items that the subjects were in the greatest agreement implied its component particle were also among the top four that they judged to imply their component verbs. These are *lift out* an action which results in the object of the verb being out of whatever location or situation it has been lifted, and *lie down*, an action of which the consequence is that the subject of the verb is *down*.

The level of agreement can be observed in figures 5.2 and 5.3. If there was not agreement then the chart would be split in half with 50% saying "yes" and 50% saying "no" for each item. As can be seen this is the situation for only 3 of the 40 verb judgements, and 1 of the 40 particle judgements. We are going ultimately to derive a categorical judgement (component entailed/not entailed) for each item by taking the majority view. We are therefore interested in how reliably such a judgement would reflect the whole group. As we discussed above there is total agreement across all subjects about the contribution of the verb for 4 of the VPCs. We can see that in this

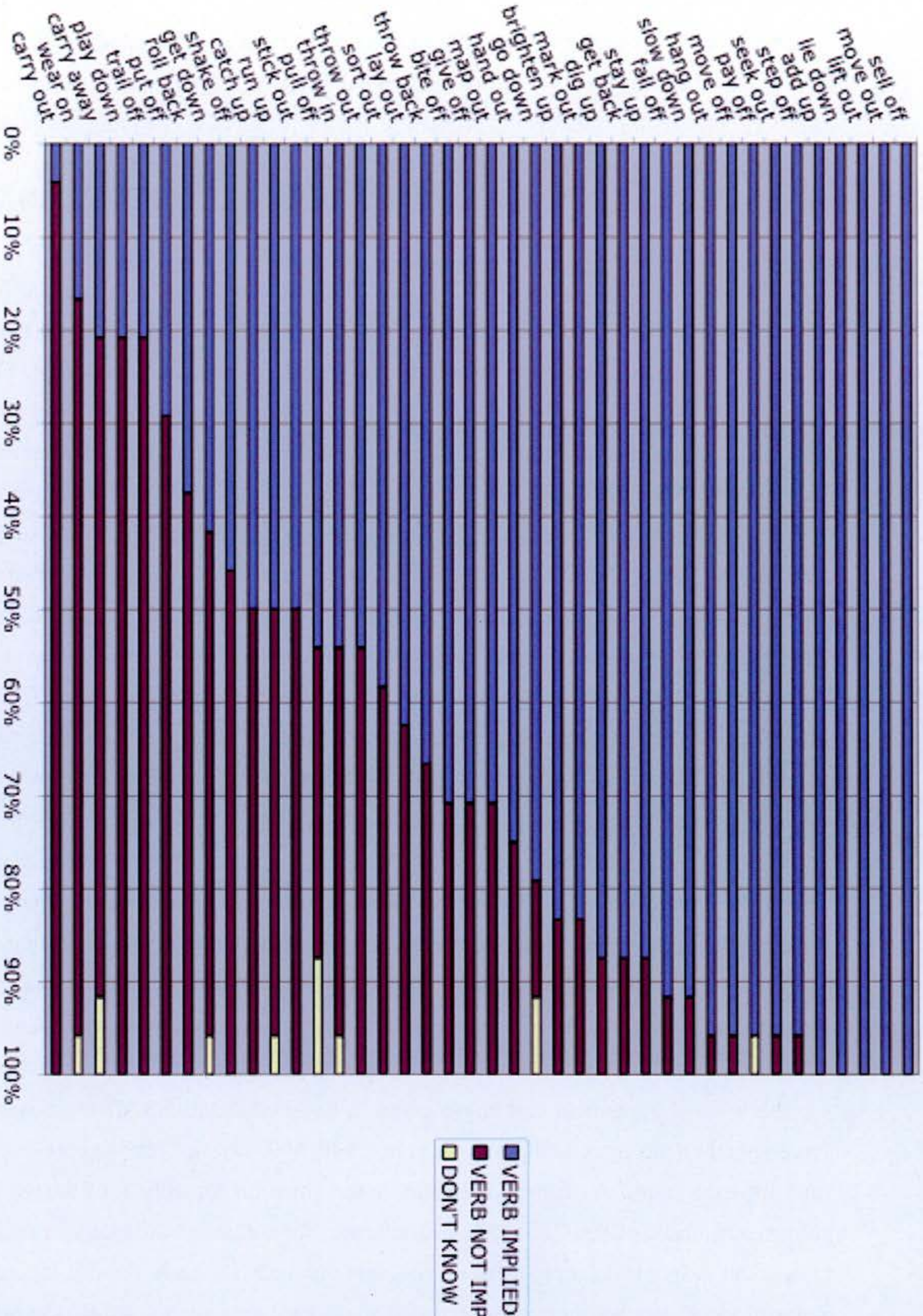


Figure 5.2: Example Verb Judgements

	Overall			Verb entailment			Particle entailment		
	P(A)	P(E)	κ	P(A)	P(E)	κ	P(A)	P(E)	κ
Group 1	0.65	0.49	0.32	0.68	0.62	0.16	0.63	0.49	0.26
Group 2	0.55	0.46	.016	0.59	0.50	0.15	0.51	0.43	0.14
Group 3	0.63	0.47	0.28	0.65	0.56	0.20	0.58	0.46	0.22
Group 4	0.65	0.49	0.31	0.69	0.55	0.31	0.61	0.45	0.28

Table 5.1: Agreement statistics for all groups

group 70% of subjects agree on judgement either way for 28 of the 40 items, and at least 60% agree for 31. For the particles, we again see a 70% decision for 28 items, and a 60% decision for 32 of the 40 items. Over all groups we see a 70% consensus on the verbs for 102 of the 160 VPCs and a 60% consensus for 127 items. For the particles we see a 70% agreement for 100 of the items and 60% consensus for 127.

These counts give us an impression of the reliability of the categorical judgements we are going to derive. However it will be useful to look more critically at the level of agreement. Another way to look at agreement is to calculate the total pairwise agreement between all pairs of subjects over all items. Calculated this way gives us a figure of 65% agreement for group four (the group seen in the chart above) and 0.65, 0.55, 0.63 and 0.65 for groups one, two and three respectively. The percentage of agreement calculated this way over judgements made for the verb and the particle separately and over all judgements combined can be seen for each group in table 5.1.

We have then a way of quantifying overall agreement. However, there is a problem with this figure, and this is that it is does not take into consideration the agreement that is produced by the simple preferences of the coders. We know that if the items were assigned randomly to the two groups with no preference for either one then we would expect there to be an agreement of 50% by chance. However, if there is an overall preference for one of the groups over the other, as there often is with linguistic judgements, and we can observe there is in our case (We can see from figure 5.2 that the subjects have an overall preference for saying that the component verb IS implied by the phrases. And in figure 5.3, we can see that there is a small preference in the opposite direction for particles). In such a situation we have a level of expected chance agreement that is actually higher than 50% for each group. A way to calculate the expected agreement and use this to calculate a more conservative estimate of agreement is provided by the κ coefficient (Cohen 1960). According to this measure in the

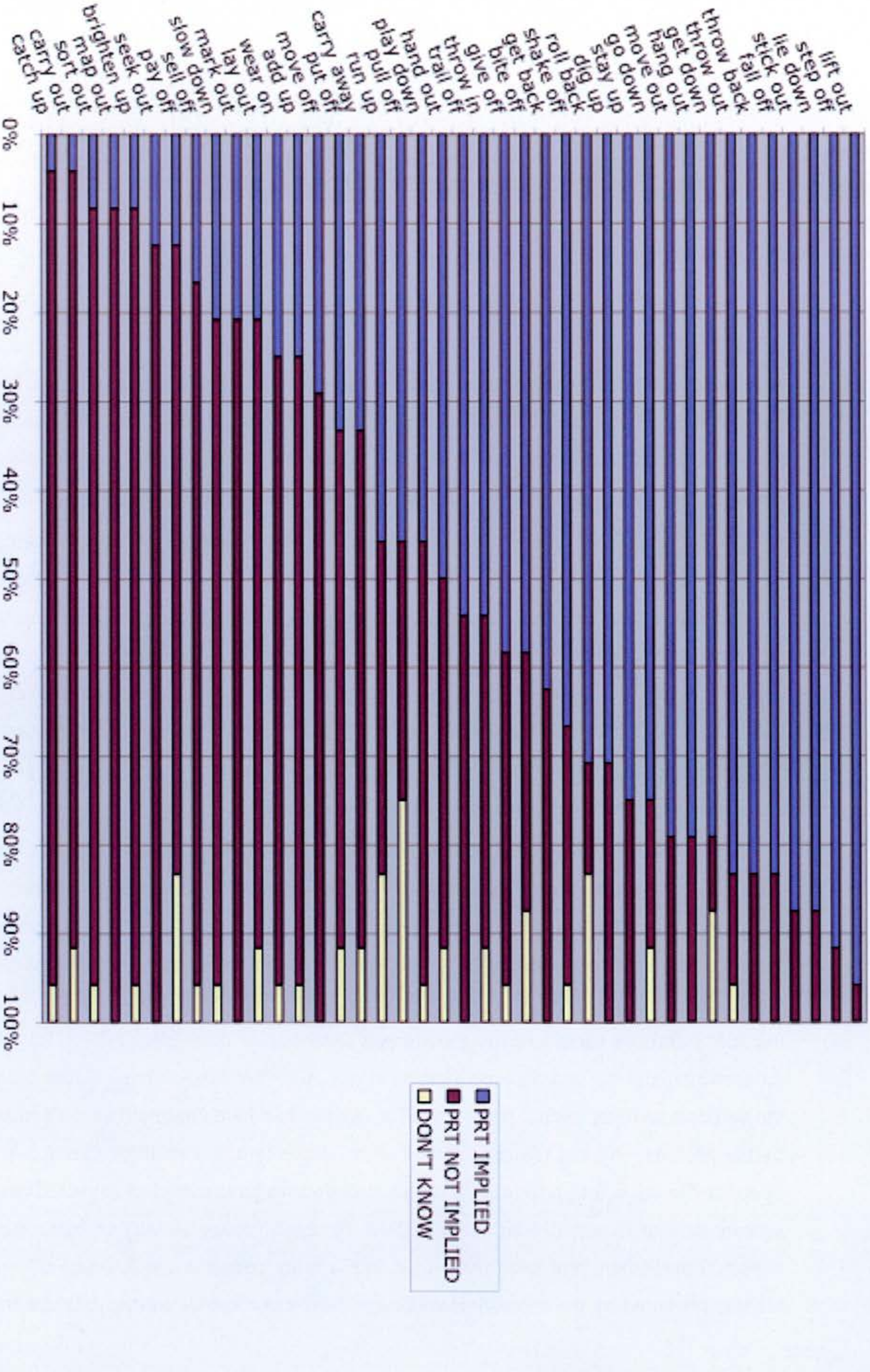


Figure 5.3: Example Particle Judgements

case of the verbs for group four we have an observed agreement of 68%, but we also have an expected agreement of 62%. The κ coefficent gives a conservative measure of agreement where observed agreement is adjusted for expected agreement. The expected agreements and all κ scores are listed in table 5.1. A z-test can be performed to test whether the value of kappa is significantly greater than 0 (whether agreement is significantly greater than chance). All kappa values for all our groups were found to be significant at a level of at least $p < 0.001$.

5.5 Discussion of experiment seven

This experiment had two purposes. The first was to check that the distinction we are making between compositional and non-compositional VPCs has reality for speakers of English. We found that subjects agree at a level significantly greater than chance on whether the component words of a large randomly selected group of VPCs entailed their component words. We take this as evidence that the distinction has reality for our judges. The second aim was to produce a useful evaluation resource. Judging success on this is less straightforward.

Interpreting agreement statistics is often difficult. Deciding what is acceptable must inevitably vary according to the task being performed. For example there is little agreement as to which levels of κ one should demand. A number of analyses have been offered in various fields. One scale was offered for Bioscientists in Landis and Kock (1977). This has a κ of 0 to 0.2 as slight agreement, 0.2-.39 as fair agreement, 0.4-0.59 as moderate agreement, 0.6-0.79 as substantial and 0.8-1 as almost perfect. Meanwhile for the field of content analysis Krippendorff (1980) suggest that the minimum acceptable level of κ should be 0.67. What we choose to do is to compare our agreement scores with agreement scores achieved and accepted for related tasks.

Between-subject agreement concerning matters of lexical semantics has been the subjects of considerable discussion in recent years, particularly among researchers concerned with the evaluation of word-sense disambiguation systems. An oft-cited study by Jorgensen (1990) looked at intersubject agreement concerning the meaning of 12 high frequency nouns. Sentences from the Brown corpus containing each of these words were typed on filing cards. Subjects were asked to sort the cards according to the sense in which the cards were being used. The mean pairwise agreement found over all words and subjects was 68%.

A number of subsequent studies have looked at the agreement among annotators of word senses creating training and testing resources for word sense disambiguation systems. Kilgarriff (2002) describes the creation of such a resource for the SENSEVAL-2 word sense disambiguation competition. He reports that the inter-annotator agreement for the first two taggings of each noun and adjective was 66.5%. Ng *et al.* (1999) perform a case study on the agreement between 2 annotators of two substantial corpora with word senses from WordNet. They report mean agreement between the two annotators of 56.7%. This gives a κ score of 0.317. While some concerns have been expressed that there is not more agreement between annotators concerning word sense, the research community have nonetheless concluded that the annotation procedures are producing useful resources.

Our study is related to these classification studies in various ways. As in these studies, our subjects were making a judgement about word sense. The entailment task involves stating whether in a particular context (as provided by the example sentences), the verb and/or the particle have a meaning that they can have in different linguistic contexts (not followed by the particle in the case of the verb, or not preceded by the verb in the case of the particle) [4] Our levels of agreement were equivalent to those found for word senses. Although the levels of between-subject agreement were not as high as would be demanded in some areas, the judgements we obtained reflected the clear significant preference of our subjects and were at the level we would expect for the task. We therefore consider the judgements a valid resource against which to evaluate our model.

5.6 Experiment eight

In experiment seven we described an experiment to elicit judgements concerning the compositionality of a set of VPC from a group of subjects. In this experiment we are going to use these judgements to evaluate the model that we described in section 5.3.

5.6.1 Materials

The main materials used for this experiment was a set of categorical compositional judgements derived from the judgements made in experiment seven. We have the judgements of a group of subjects as to whether the component verb and/or the com-

[4] In fact our task is arguably slightly more likely to result in disagreement that the word sense annotation as unlike in those experiments we provide no sense inventory as guidance, making the decision more dependent on individual intuition about what does or does not constitute a word sense.

ponent particle is entailed for 160 VPCs. We derive categorical judgements for these by taking the majority view. If a majority of the subject group say that a verb is entailed by a VPC then we derive a categorical judgement that it is. For 4 of the 160 verb judgements and 3 of the 160 particle judgements the subjects were tied in their entailment assessment. We therefore removed these judgements from consideration. As well as judgements concerning the VPC's entailment of the component words, we also derived a single overall compositional/ non-compositional judgement for the whole phrase, where a phrase was deemed compositional if both component words were entailed, and non-compositional otherwise.

In order to build our model we extracted from the written component of the BNC all sentences involving the use of all VPCs and all of their component words. The VPC sentences were extracted as described in section 5.4.2. The component words were also extracted by looking for their base forms in the lemmatised corpus. In order to ensure that only verb forms were extracted for each of the component verbs and not nominal graphonyms, we restricted ourselves to only those items that were tagged as a verb. As there is some variance of opinion as to the relation of particles to other word-forms (prepositions and adverbs; see O'Dowd, 1998), and such forms are notoriously difficult to automatically tag correctly (Toutanova and Manning 2000), we used all forms that matched the forms of our component particles, and did not use part-of-speech tags to distinguish between them.

5.6.2 Method

Firstly we used the corpus examples described above to build context vectors. We built the vectors as described in section 5.3. For each VPC, we obtained a context vector for the whole phrase as well as vectors representing a) the component verb when occurring outside of the VPC, b) the component particle when occurring outside the VPC and c) a combined vector created by adding together the two vectors for the two component words described under a) and b). The sum of a set of two or more vectors is created by simply adding together the entry for each coordinate for the two vectors so that for each coordinate i, $ab_i = a_i + b_i$.

We next calculated the cosine similarity between each VPC and the three related vectors described above. We then examined the relationship between these cosine scores and the compositionality judgements obtained from our subjects in various ways which we will explain below.

5.6.3 Results

Our principal hypothesis was that the cosine similarity between a VPC and a given component word will be higher when that component word is contributing compositionally to (is entailed by) the meaning of the whole phrase. We also tested the hypothesis that the cosine similarity between a phrase and the sum of the component vectors will be higher when that phrase is fully compositional. We first evaluated this by building a series of logistic regression models in which the binary classes of "compositional" and "non-compositional" were the outcome variables and the cosine scores were the predictor variables.

We first did an analysis for each component. We found that with verb entailment as the outcome variable and cos(VPC,Verb) as the predictor variable, we get a model with a β value of 2.703 (Exp(β) = 14.927). This tells us the amount of increase in the log odds of the compositional class that the model would predict given a 1 unit increase in our predictor variable. This indicates that there is a positive relationship between the cosine score and compositionality. It is conventional to perform a Wald χ^2 test in order to test whether this relationship is significant. This gives us a χ^2 value of 5.611, which is significant at a level of $p < 0.025$. A model with particle entailment as the outcome variable and cos(VPC,PRT) as the predictor has a β of 3.344 (exp(β) = 31.174) for which a Wald test give a χ_2 value of 4.99 which is significant at $p < 0.025$.

We next built a model with the overall compositionality of the VPC as the outcome (any item which entailed both items was said to be compositional and all other items to be non-compositional) and the cosine between the VPC and the sum of the vectors of its two component words as the predictor. This gave us a β of 4.278 (exp(β = 72.071)). A Wald test of this value gives a χ^2 value of 9.431 which is significant at $p < 0.0025$.

These tests allow us to reject the null hypothesis. Now we need to see how this relationship might translate into a useful tool for determining the compositionality of VPCs. The eventual goal is for the method to serve in a tool for lexicographers where it will help to direct attention to those items that are likely to be non-compositional. What we want to show then is that the measure can be used to rank items, where the items at the top of the list are more likely to be non-compositional and those at the bottom of the list are more likely to be compositional. In order to create such lists we inverse ranked our items using the cosine scores, so that the item with the lowest cosine (which we hypothesize will be more likely to be non-compositional) was at the top, and the item with the highest cosine (which we hypothesize will be more likely to be compositional) was at the bottom. We first examined these ranks using a Mann-Whitney U test. This told us whether the relationship we saw above in the logistic regression models translated into a significant difference in the ranks of the compositional and non-compositional items. We then went on to look at the ranks in terms of the precision and recall they gave when we look at the *n* top ranked items.

We introduced and defined the Mann Whitney test in section 3.2.3. The measure examines difference in the ranks that two different groups of entities (in our case VPC items) obtain in a sorted list of all entities. The U score is the difference between the sum of the ranks of the more highly ranked group and the maximum sum of ranks it could have obtained (where a high number indicates a high rank). This can be used to obtain a z-score that tells us whether the difference is significant.

Again we will first report results for the individual components. We inverse ranked the items using cos(VPC,Verb). The items that were judged to not entail their verbs had a mean rank of 98.88 out of 156, while the compositional items had a mean rank of 72.82. Remember that a high number indicates a high rank here, so, as we predicted, the non-compositional items had a notably higher mean rank. We use a Mann-Whitney test to examine whether this is significant. The Mann-Whitney U value is 1381, which gives us a Z-score of -2.975 which is significant at a level $p < 0.005$. We next inverse ranked the items using cos(VPC,Prt). The judgements that were judged not to entail their verbs had a mean rank of 87.53 and those that were judged to so entail had a mean rank of 69.93, both out of 157. This is again in line with our hypothesis. A Mann whitney test gave us U of 2389, which yielded a z-score of -2.420 which is significant at $p < 0.025$.

We next inverse ranked our item list using the cosine of the VPC vector and the sum of the vectors for the component words. Remember that we used the component entailment judgements to derive judgements concerning the compositionality of the whole phrase, with items for which both components were judged to be entailed being labelled compositional and all others non-compositional. In our ranked list the non-compositional phrases had a mean rank of 85.79 and the compositional phrases had a mean rank of 63.38. This is consistent with our hypothesis, and a Mann-Whitney test gave a U of 1972.5, with a z-score of -3.055 which is significant at $p < 0.005$.

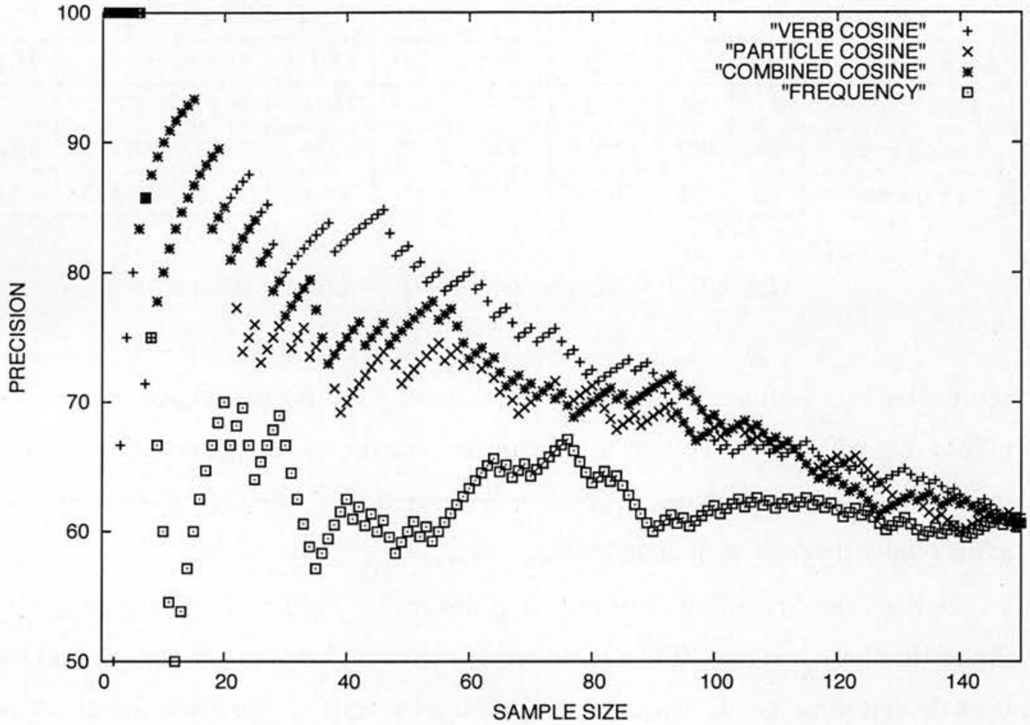
The results above tell us that our measure does give significantly higher ranking to non-compositional VPCs as we predicted. While this is a promising result, a more realistic measure of its usefulness is provided by precision and recall scores. What we want to do is maximise the number of non-compositional VPCs that a lexicographer would come across when reading down a ranked list of items. The evaluation method that we think most representative of this is calculating the precision that is achieved for different values of *n* as one reads down the list. The sample size we are dealing with here is of course small and so it would be straightforward to consider all the items. However we assume that our random sample is representative and that performance here will be representative of performance over larger samples.

The problem of lexicon development as we have discussed it in this thesis is in deciding what elements to put in the lexicon. We are ultimately interested then in making a judgement as to the compositionality of the whole item rather than its entailment of the components. We therefore evaluated our ranked lists for their precision and recall on the task of identifying items that were non-compositional as a whole unit (e.g. contained one or more elements that were judged to not be entailed by the phrase). The experiment is analogous to that reported in chapter 4 and for the reasons we gave in section 4.4.2, we report both precision and recall scores, but restrict our discussion to precision.

Figure 5.4 shows the precision and recall scores achieved for various lists ranked using cosine scores. The complete set of items for which we have a consensus judgement consists of 153 VPCs. Of this set 93 were deemed to be non-compositional. Following convention then we can identify the lower bound on performance at a precision score of 60.78 for all *n*. This value is plotted in figure 5.4. As can be seen our method beats this value for all plotted values of *n*. However, before we consider this a success it is necessary to check whether we are beating chance performance by a significant margin.

While 60.78 is a reasonable hypothetical lower bound, actual observed performance could potentially be considerably higher. A chance ordering could put the non-compositional items anywhere in the ranks. In order to measure whether our results are significantly better than chance then we need to calculate the probability that a random ordering would equal our performance, and show that this probability is below an acceptable level (which we will put at $p < 0.05$). Fortunately the situation is described by a known distribution. The situation of taking the top *n* items from a randomly ordered list is equivalent to that of sampling with replacement from some collection of entities, and is described by the hypergeometric distribution. We can use this distribution to calculate the probability that any number of noncompositional items we retrieve in any top *n* could have occurred by chance [5]. Table 5.2 reports the precision

[5] We use the phyper() function in the R statistical computing environment to calculate this p value.



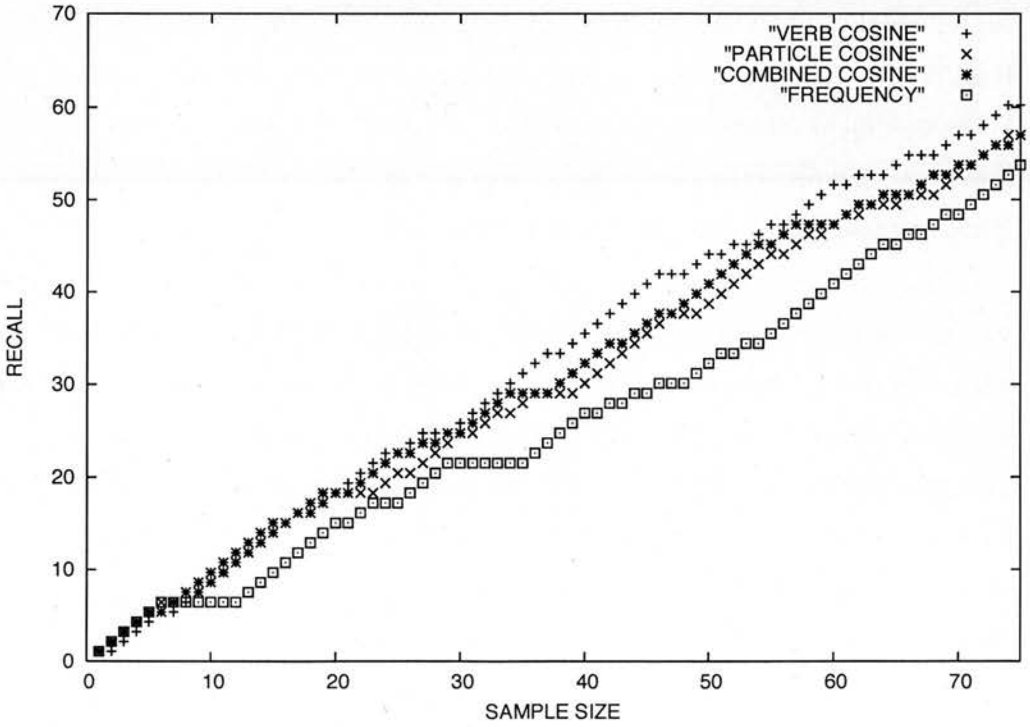


Figure 5.4: Precision and recall scores by sample size

n	10	20	30	40	50	60	70	80	90	100
cos(VPC,V)	**80**	**85***	**80***	**82.50***	**82***	**80***	**75.71***	**72.50***	**72.2***	**67***
cos(VPC,Prt)	**80**	**85***	**76.60**	70	**72**	**73.33***	**70**	**71.25***	**68.89***	**68***
cos(VPC,V+Prt)	**90***	**85***	**76.60**	**75***	**76***	**73.33***	**71.43***	**70**	**71.11***	**69***
Comb. ranks	**90***	**80**	**83.33***	**77.50***	**76***	**76.60***	**75.71***	**73.75***	**72.22***	**69***
Lowest comp.	**80**	**90***	**83.33***	**75***	**72**	**75***	**75.71***	**72.50***	**68.89***	**68***
Frequency	60	70	66.66	62.50	60	63.33	64.29	63.75	60	62

Table 5.2: Precision scores for different values of *n*

scores that our various ranking methods achieve for different values of *n*. Where there is less that a 0.05 chance that this result occurred due to chance, the value is shown in bold. Scores which have a less than 0.01 chance of having occurred by chance are additionally marked with an asterisk.

Perhaps the first thing to note is that the results achieved by inverse ranking with the cosine between the VPC and either of the two components in isolation markedly beat the baseline for all values of *n*. Ranking by verb cosine gives a performance that is statistically significant at at least $p < 0.05$ for all values of *n* and significant at a level of $p < 0.01$ for all values of *n* except for 10. The particle is significantly better than the baseline for all values of *n* except 40, and at a level of better than $p < 0.01$ for half of the values of *n* reported. The logistic regression models and Mann-Whitney tests reported above showed us the verb cosine to be marginally better at predicting verb entailment than the particle cosine is for predicting particle entailment. The fact that ranking with verb cosine should give a better precision value on the detection of non-compositional VPCs than ranking with particles is perhaps to be expected.

Having shown that ranking using the cosines of each of the component words in isolation significantly beats the baseline, we next looked at the performance achieved with various methods of combining them. First of all we looked at the results achieved using the cosine between the VPC vectors and the sum of the component word vectors. This was significant at all values of *n* and even achieved significance at the level of $p < 0.01$ for *n* equal to 10, with 9 of the items in the retrieved set being non-compositional. However it did not beat the components for all *n* and often scored with the lower of the two components.

The next way of combining the two component scores we tried was combining their ranks. The set of items was inverse ranked using the two scores in isolation. The ranks achieved for the two separate components was then added together to give a single value and the whole set of items reranked using this value. The intuition here is that items which have low cosines for both the components are the most likely to be noncompositional. This achieves good scores. However it only beats both components in isolation for 4 out of 10 values of *n*.

That the sum of the component vectors or of their ranks does not always beat the components in isolation is not surprising. The compositionality of the whole phrase can result from the non-entailment of either of the component parts, and in many cases one element will be entailed but the other not. Therefore by combining information about them we are potentially removing the distinguishing power of the component cosines. We therefore try an alternative approach where instead of combining the contribution of the two component vectors or cosine scores for each item we instead take the lowest scoring of the two methods. The motivation for this is that the non-compositionality of the whole will be indicated by the existence of a single component word which has a low cosine irrespective of the cosine of the other component word. This approach gives a different distribution of scores to the component combinations described above. However, it does not surpass the performance of the verb cosine ranking to anything approaching a significant degree.

The final ranking method that is reported in table 5.2 and is also plotted in figure 5.4 is the frequency of the VPC form. Unlike in experiments five and six, we are evaluating here against human compositionality judgements rather than a general purpose lexicon and we would not expect frequency to be a useful way of ranking the forms. Nonetheless it is the most basic method available for sorting, and is most likely to be the first resort of the corpus lexicographer. We therefore evaluate what performance a frequency ranked list would give us. As can be seen in figure 5.4 its precision score does not vary much from the baseline value, and we can see in table 5.2 that its precision value is never significantly greater than that achieved by chance.

5.7 Discussion of experiment eight

Having performed a quantitative evaluation, and shown that our model beats the baseline by a significant margin, it is interesting to perform a qualitative study of performance and look at examples of the items we are recovering. Table 5.3 shows the 25 items with the lowest and highest cosines respectively. Items that were judged to be non-compositional overall (to have at least one component that is not entailed by the phrase) are marked with an asterisk. Items for which the component relevant to the column (i.e. in columns where the items are sorted by the verb cosine those items where the verb was judged not be entailed by the whole VPC) are in bold.

Looking at the distributions (the context vectors) that produce the results, we see that the method reflects our initial intuitions. Two of the ten lowest scoring items for the verb cosine are *carry out* and *carry away*, both of which are non-compositional in what the subjects deemed to be their dominant sense. The VPC *carry out* has as its top three most frequent collocates the words *work*, *research* and *study*. All of these are indicative of the phrase meaning "to conduct" (e.g. *She carried out the work/research/study effectively*). Similarly the phrase *carry away* has as its top three collocates *get*, *too* and *so*, all of which are indicative of the phrase meaning "to move or excite greatly" (e.g. *You mustn't get too/so carried away*). Neither of these phrases have a meaning that overlaps with the meaning of the verb *to carry*, and none of these most frequent collocates for either item are found in the top 100 items for that verb. The verb *carry* has among its most frequent collocations the words *passenger*, *bag* and *weight*, all of which are are indicative of the word's meaning, and none of which overlap with the phrases *carry out* or *carry away*. A similar observation can be made for the particle cosine scores. The VPC *trail off* has the second lowest cosine relative to its component particle. For example, the three most frequent collocates for the top ranked non-entailed VPC *trail off* are *voice*, *her* and *sentence*. Neither *voice* nor *sentence* are in the top 1000 collocates for *off*.

We see a similar confirmation of our intuitions at the other end of the ranks. A phrase that is in the highest scoring 25 items for both verbs and particles is *live in*. Of the top 20 most frequent collocates for this phrase, 18 are in the top 100 collocations for the component verb *live*, including the intuitively semantically related forms *life*, *people*, *house*, *family* and *home*. Nine of the top 20 most frequent collocates of *live in* are in the top 100 collocates for *in* including *time*, *country* and *area*.

All of these examples are consistent with our hypothesis that the frequent collocations of a phrase reflect its meaning and that compositional items should show an overlap in these with their component words that non-compositional items don't. Of course it should be evident from the list, as it was from the quantitative results, that the performance is not perfect, and it is also interesting to look at where the technique fails.

The most significant problem as in all computational approaches to meaning seems to be polysemy. One of the items in the ten lowest scoring VPCs that is compositional is *flush out*. The erroneous suggestion of our method that this item is non-

	Lowest		Highest	
	cos(VPC,V)	cos(VPC,P)	cos(VPC,V)	cos(VPC,P)
1	eke out*	**close up***	go in	back off
2	fend off	trail off*	**get in***	**back up***
3	**carry away***	**eke out***	live in	**start off***
4	mop up*	hold out*	do up*	take back
5	bail out*	**stretch out***	back up*	go in
6	**carry out***	**drown out***	bring in	get down*
7	flush out	flush out	**back off***	**work up***
8	brighten up*	**whip up***	hand out	go down
9	close up*	blow out	move out	**help out***
10	**stamp out***	**speed up***	stay on*	get in*
11	patch up*	**brighten up***	**get down***	**head off***
12	shut down*	**mix up***	come back	**run down***
13	figure out*	**patch up***	go out	**finish off***
14	**tick off***	**do up***	go down	come back
15	**stir up***	**sell off***	help out*	pull back
16	**set out***	**stamp out***	throw in	bring in
17	mix up*	**bail out***	get back	**stay on***
18	read out*	**stir up***	start off*	go out
19	**whip up***	**curl up***	slide down	stand up
20	trickle down	jump up	move off*	get back
21	**shake off***	shake out	pull down	move out
22	**trail off***	throw back	pull back	**come about***
23	**draw up***	pay back	**wear on***	**pack up***
24	speed up	**pay off***	take over*	live in
25	**throw back***	**draw up***	**let off***	sit down

Table 5.3: The 25 highest and lowest scoring items for cosine measure

n	10	20	30	40	50	60	70	80	90	100
Inv f(V)	70	65	60	60	60	58.33	55.71	56.25	56.66	58
Inv f(Prt)	20	40	50	50	48	46.67	50	52.5	51.11	52
t-score	60	70	60	60	64	66.66	65.71	63.75	62.22	60.61
log likelihood	**80**	75	60	60	64	63.33	64.29	61.25	65.55	63
MI	70	75	66.66	62.5	68	63.33	64.29	62.5	64.44	64
χ^2	70	70	73.33	72.5	68	63.33	64.28	61.72	63.33	64

Table 5.4: Collocational measure precision scores for different values of *n*

compositional comes from the fact that the sense in which the word is used in the phrase is the less frequent of multiple senses of the word. The phrase has the meaning "to force out into the open", as in the sentence *The hounds are sent in to flush out the fox from its haven*. The verb on its own can have the meaning "to force from some location" as in the sentence *it was as if the game birds had been flushed by beaters and driven towards a central point*. However two more commonly attested meaning are the related "to be emptied by a flow of water" as in the sentence *As if on cue, a cistern flushed and the door of one of the WC cubicles opened to reveal the emerging figure of George Prendergast, the Personnel Director* or "to glow" as in the sentence *Roger's face had flushed, and she knew he was thinking about that wallet*. Indeed it is these two meanings that are reflected in the most frequent collocates for the verb *flush*. The top twenty most frequent collocates includes the words *toilet* and *water* which reflect the first of the attested meanings, and *face*, *cheek*, *hot* and *feel* which reflect the latter. None of these words are in the most frequency collocates for the VPC, and consequently a comparison of the phrase and the verb receives a low cosine score, and the phrase appears near the top of our list of candidate non-compositional VPCs.

Looking at the top and the bottom of the ranked lists, one apparent tendency is for the items with the very lowest cosine scores to have less frequent component verbs, such as *eke* or *fend*. It seems possible then that sorting the list by the inverse frequency of the component words will be an effective way of detecting non-compositional items. In order to check whether this more straightforward method could produce equivalent results, we created a list of items inversely sorted by the frequency of the verb and a list inversely sorted by the frequency of the particle and evaluated them against our set of compositionality judgements. We calculated the precision scores obtained for various values of *n* as we did above for our cosine scores. These are reported in table 5.4. Neither of these sorted lists performed at better than chance.

We showed in section 5.6.3 that sorting by frequency did not produce significant results. We have now also showed that sorting by the inverse frequency of the components does not produce significant results. As a final check that frequency information cannot provide a performance that rivals cosine, we calculate a series of measures that combine all of these frequency values – the lexical association measure that we introduced in section 4.5.1. All of these provide a measure of whether the frequency of the whole phrase is at a level predictable given the component words. Phrases that are common but whose component words are infrequent will achieve high values. The precision scores achieved by each of these scores is presented in table 5.4. We see a single significant result for the collocation measures, achieved using the log-likelihood ratio and taking the top 10 items. No other significant result is achieved for this or any other measure. We therefore conclude that collocation measures do not provide a useful alternative to our cosine ranking, and crucially that the model proposed in this chapter gives a performance on the identification of the non-compositional phrases in our random selected sample of items that surpasses that available through frequency-based methods for candidate ranking.

5.8 Other related work

This section will describe other work on the detection of non-compositional units that employ methodologies unrelated to ours.

Melamed (1997) proposes a technique for identifying what he calls "non-compositional compounds" (NCCs). His basic intuition is that a string of words that constitutes a non-compositional compound will be translated as a unit in any pair of parallel texts. He attempts to discover when this occurs by inducing two different translation models for a language pair, using slightly different training corpora. In the corpus used to train one translation model the candidate NCC is left as separate words, while in the other it is concatenated to form a single unit. Mutual information is then used to discover how well the two models predict the distribution of words in the target language given the source language. The idea is that if MI is higher for the translation model in which the candidate NCC is concatenated then the NCC is valid. Candidate NCCs can be identified using only the base translation model by making the assumption that for any given series of adjacent words xy, if xy is a NCC then at most one of x and y will be linked to a word in the target language.

While this work could conceivably provide information that is useful for MT (where collocational information can be very important in outputting the correct string), as a technique for identifying non-compositional compounds it is flawed. The paper provides 50 example NCCs, of which less than a quarter appear to me to actually be non-compositional. The largest group on the list are compound nouns of which the majority, such as *video tape* (translates as *vidéo*), *machine gun* (translates as *mitrailleuse*), and *swimming pool* (translates as *piscine*) are compositional. Compound nouns are often a problem for NLP, but we still would not want to regard all of them as non-compositional. There are also a great number of compositional examples of other kinds of constructions included, such as *understand the motivation* (translated as *saisir le motif*), and *we lag behind* (translated as *nous trainions de la patte*). The problem is that the assumption on which the technique for identifying NCCs is based is incorrect. It simply isn't the case that a compositional expression in one language will translate into the same number of words in another language. There are many compositional constructions that repeatedly translate as a single word in another language.

Lin (1999) addressed the same problem, this time based on distributions in a monolingual corpus. Lin's method is based on the premise that non-compositional items should have a markedly different degree of association to expressions derived through synonym substitution over the original word combination. The items to be tested were taken from a collocation database, the construction of which is described in Lin (1998b). For each collocation, he substituted each of the component words with a word with a similar meaning. The list of similar meanings was obtained by taking the 10 most similar words according to a corpus-derived thesaurus, the construction of which is described in Lin (1998a). The mutual information value was then found for each item produced by this substitution [6] A phrase **a** was then said to be non-compositional iff there exists no phrase **b** where: (a) **b** can be produced by substitution of the components of **a** as described above, and (b) there is an overlap between the 95% confidence interval of the mutual information values of **a** and **b**. The evaluation offered is a comparison of Lin's system results with the contents of a dictionary of idioms. If an item is in the dictionary then it is said to be non-compositional. This produces scores of

[6] We described the use of the mutual information score for measuring collocation in section 4.5.1. In fact Lin uses a slightly different measure, designed to take into consideration the prior probability of occurrence of the dependency relation between the words. He takes a collocation to consist of three events: the type of dependency relationship (A), the head lexical item (B), and the modifier (C), and calculates MI as follows:

$$I(A,B,C) \quad = \quad log_2 \frac{P(A,B,C)}{P(B|A)P(C|A)P(A)} \tag{5.4}$$

15.7% for precision and 13.7% for recall.

There are problems with the underlying assumptions of Lin's method. The theoretical basis of the technique is that compositional items should have a similar distribution to items formed by replacing components words with semantically similar ones. The idea presumably is that if an item is the result of the free combination of words, or a fully productive lexical rule, then word-substituted variants should be distributed similarly. This seems a reasonable basis for modelling productivity but not compositionality, as Lin claims. There are many examples in natural language of phrases that are not at all productive but are still compositional (e.g. while we frequently encounter the phrase *frying pan* but rarely hear mention of *steaming pots* the meaning of both phrases can be recovered from their parts as well as for any compound nominal; this is an institutionalised rather than a non-compositional phrase). This is very similar work to that carried out by Pearce (2001), for the purpose of identifying collocations, a task to which it is much more suited. Since Lin evaluates against a general list of MWEs rather than a list of items judged to be non-compositional, it does not tell us a great deal about its success on the task it attempts. His results will reflect the retrieval of institutionalised as well as non-compositional expressions. It is therefore very difficult to assess their significance.

5.9 Chapter summary

In chapter 2 we introduced non-compositionality as one of the dimensions along which phrases become lexicalised, and as defining one of the categories of phrase that need to be included in the lexicon. In this chapter we introduced the specific problem of non-compositionality in a particular variety of syntactic phrase - the verb particle construction. Having acknowledged that learning about the semantic compositionality of phrases from corpora was vastly more difficult than obtaining information about their frequency or syntactic fixedness, we went on to describe how the lexical collocates of a given VPC could be used to assess its compositionality.

It has long been acknowledged in linguistics that the words with which a word cooccurs are indicative of its meaning. This observation has more recently been harnessed for a range of NLP tasks, where such "distributions" have been used to find synonymous words and to distinguish between sets of homonyms. We introduced a suggestion, made originally in the linguistic literature, that compositional VPCs should have similar distributions to their component words, and that by identifying items that have a very different distribution from their components, we might identify non-compositional phrases. We then introduced a way of quantifying the similarity of two such distributions to give a value between 0 (indicating complete disjunction) and 1 (indicating identity). We proposed that inversely sorting a list of VPCs by this score would result in a list in which non-compositional items are more likely to be at the top.

In experiment seven, we set about testing whether the distinction between compositional and non-compositional phrases has reality for speakers of the language, by obtaining judgements from a set of 121 native-speakers using a test based on the idea of lexical entailment. This test has been performed by linguistic experts before, but it has not used to examine the intuitions of a more diverse subject set, and no study of between annotator agreement on the task has been performed. We obtained judgements for 160 items at the type level. We found levels of agreement that were statistically significant and equivalent to that found for related tasks.

In experiment eight we used the judgements obtained in experiment seven in order to evaluate the model we introduced earlier. We first of all built a series of logistic regression models with the assessment of distributional similarity as the predictor variable and a label of either "compositional" or "non-compositional" as the outcome. We found that the distributional similarity of a VPC to its component verb or particle was a significant predictor of the entailment of that item by the phrase, with a higher similarity score predicting compositionality. We further found that the similarity of the distribution of the VPCs and a combination of the distribution of the two component words was a significant predictor of the overall compositionality of the VPC.

We next examined whether the scores could be effectively used to rank VPC items. We used the judgements as to the overall compositionality of the phrases to calculate precision and recall values for the top *n* items of the list for all *n* where the list was inversely sorted using our measure of the similarity of each VPC to its component words. We showed that taking the components either together or in isolation this ranking method performed significantly above the random baseline. We went on to show that ordering the items by the frequency of the VPCs or their components, or using various measures of lexical association, failed to beat chance performance. We therefore conclude that our measure provides a more effective way of ranking candidate VPCs according to their compositionality than available frequency-based measures.

Chapter 6

Conclusions

This chapter will summarize the main findings of this thesis and discuss some issues to be explored in further research.

6.1 Main findings

The main claim of this thesis is that we can acquire information from corpora about recurrent multiword sequences that could be useful in the creation of lexical resources. In chapter 2 we provided a distinction between institutionalised and lexicalised phrases, and argued that in order to provide an adequate description of the phrasal lexicon, it is necessary to extract information about both varieties. The main finding of this thesis was that we could obtain information from corpora about both kinds of phrase. We will discuss each of these in turn.

6.1.1 Institutionalised phrases

An institutionalised phrase (or collocation) is a word sequence that is conventional in the language. Whether we define conventionality in terms of raw frequency or of some measure of the association between the words, the basic requirement of an institutionalised phrase must be that it is recurrent. While the extraction of recurrent two word sequences and their processing by speakers have been studied extensively (see section 4.5.1 and 2.3.1.1 respectively for discussion of this work), there has been very little work on the recurrence of longer strings [1]. And while work on the extraction of two

[1] The most extensive discussion of the occurence of longer repeated strings in corpora has been in the applied fields of n-gram language modelling and phrase-based statistical machine translation (this has mostly concerned three word strings, although see Callison-Burch *et al.*, 2005 for work exploring

word sequences has found application in the construction of terminology databases Daille (1996), many specialists in the study of MWEs and formulaic language have questioned whether corpora can provide information about the frequency of MWEs that is reliable (Moon 1998) or informative (Wray 2002). In chapter 3 we explored the adequacy of frequency information extracted from corpora about multiword sequences of up to 7 words in length along both of these dimensions.

We first looked at the issue of reliability. Corpora have long been used to obtain estimates of the frequency of individual words that are taken to be representative of the language as a whole. It might seem reasonable to attempt a similar description for phrasal lexical items. However, it has been claimed that MWEs are too specific to particular genres and topics for even balanced-design corpora to be able to provide generalisable information about their frequency in the language. We set out to examine whether this was the case.

We defined the notion of burstiness and introduced a number of ways of quantifying the stability of frequency counts. We then used these measures to study the stability of two different sets of multiword sequences – first a set of recognised MWEs randomly selected from an existing lexical resource (reported in experiment one), and then a random selection of arbitrary recurrent multiword strings acquired directly from a corpus (reported in experiment two). We ensured that both sets contained an equal number of phrases for each length of between 2 and 7 words and that the phrases covered the full frequency range. We matched each phrase to a single word of equivalent frequency. We took the written component of the BNC and split it into segments of 28,000 words, the average document size in the corpus. We then estimated the stability of the words and phrases over these segments. We found that for both the random sample of dictionary MWEs and the randomly selected arbitrary substrings, the counts were at least as stable as those for individual words. We therefore conclude that, contrary to previous claims, the counts obtained from the BNC for phrases are at least as reliable as counts for individual words.

We next looked at the informativity of these counts. We wanted to show that these frequencies are not merely a fact about the distribution of linguistic objects but are in fact reflected in speakers' processing of the language. We began by extracting all repeated strings of between four and seven words from the written component of the BNC. For experiment three we selected 12 pairs of phrases, where each pair was identical except for the final word. For experiment four we selected 12 pairs of phrases

the use of phrases of up to a length of 10 words in SMT).

where each pair was matched for syntactic form. In both experiments one of each pair was of high frequency and the other was of low frequency.

Example sentences for each sequence were then presented to subjects in a self-paced reading experiment. Each sentence was split into chunks of between 3 and 8 words which were presented in sequence, with the subject being instructed to press a button to progress to the next chunk. The time between initial presentation and the pressing of the button was recorded for each chunk. We then compared the reading times for the frequent sequences with the reading time for the infrequent sequences. We found that in both experiments the frequent phrases were read more quickly than the infrequent (controlling for length in characters, component word frequency and phrase-internal transitional probabilities). We therefore concluded that the frequency of multiword sequences as found in the BNC are reflected in speakers' processing of the language.

6.1.2 Lexicalised phrases

Antilla (1989:151) writes that "[w]henever a linguistic form falls outside the productive rules of grammar it becomes lexicalised". As we discussed in chapter 2, Boguraev and Briscoe (1989:4-5) state that the job of the lexicon is to provide the "information not predictable from the rules" of the language. Institutionalised phrases are usually thought not to be predictable from any set of "rules", as they concern the frequency of combination of particular lexical forms. It is necessary therefore to include them in the language description somehow. Lexicalised phrases, by contrast, are those items that it is necessary to specify in the language description because they are not consistent with the productive, compositional communication system at the level of syntax or of meaning.

In chapter 2, we introduced two different kinds of lexicalised phrase. The first variety we discussed was syntactically lexicalised, and consisted of word combinations for which a single form is strongly preferred, and which do not allow the full range of syntactic variation that we would expect for phrases of their syntactic type. The second variety was semantically lexicalised, consisting of phrases whose meaning is not predictable from the meaning of their component words. In this thesis we showed that it is possible to distinguish to a significant degree which phrases are lexicalised in both these ways.

In chapter 4 we showed that it is possible to quantify the degree of syntactic fixedness of verb-plus-noun-object phrases using a corpora, and that this measure is a useful predictor of which phrases will be found in existing multi-word lexical resources. In order to calculate our measure, we use the linguistic literature to explore which kinds of variation we might expect to see for this phrase type, and which have been reported to be subject to restriction in lexicalised phrases. We used an automatically parsed corpus to calculate the probability of seeing each variation for all verb phrases found in our corpus. We argued (and later empirically showed) that these probabilities are not adequate in themselves and that we need instead to calculate how much they deviate from what we would expect given the component words of the phrase. We described how to model this deviation using an information theoretic measure and how to combine the individual variation types to give a single overall assessment of fixedness.

Having calculated a measure of fixedness for all phrases, we used this to rank all verb and object combinations. What we want is a measure that will sort items so that those that are fixed (and hence need to go in the lexicon) are at the top of the list, thus reducing the time that a lexicographer would need to spend in order to find the items that most need to go in the lexicon. We evaluated the utility of our measure by using verb and object pairs found in two different published lexicons. We calculated the precision and recall scores obtained for the top n items for various values of n. We showed that over the selected values of n, ranking with our fixedness score provided a higher precision than sorting by either raw frequency or by a set of four popular association measures, to a degree that was often statistically significant. Furthermore we showed that while raw frequency and the association measures identify very similar sets of phrases, our method identifies a largely disjoint set of items. We therefore concluded that measuring syntactic fixedness in this way provides a way of identifying MWEs that are not available from methods that focus upon institutionalisation. This confirms our hypothesis that in order to create adequate phrasal lexicons we need to focus upon lexicalised as well as institutionalised phrases.

In chapter 5, we explored a method for identifying which phrases have a meaning that cannot be predicted from the meaning of their component words. We focused upon one variety of phrase - the verb particle construction. We applied an insight from the linguistic literature that compositional phrases should occur in very similar lexical contexts to their component words. The intuition was that by inversely ranking VPCs according to their similarity to their component words, we should create a list of items in which non-compositional items are more likely to be at the top. We described a way to quantify this, based upon work on quantifying the distributional similarity of words in word sense disambiguation and automatic thesaurus extraction.

In order to perform an evaluation of this method we obtained judgements from 121 subjects as to the compositionality of 160 VPC items. We did this for each component word, eliciting separate judgements as to whether the verb and the particle contribute compositionally to the whole phrase. We subsequently used these to calculate an overall judgement as to whether the phrase was compositional or not (with items in which both components are contributing compositionally being considered as compositional, and all others as non-compositional). We showed that there is significant agreement between subjects, and then set about using the judgements to evaluate our model. We first of all showed in a series of logistic regression models that the contextual similarity of each VPC to the context of both components in isolation is a significant predictor of whether that component word is contributing compositionally to the whole, and that the similarity between the contexts of the VPC and the combined contexts of the two component words is a significant predictor of the the overall compositionality of the phrase. We next used these same scores to rank the VPCs, and perform a top *n* evaluation of precision as in chapter 4. We showed that ranking with our measure provided a precision value that was significantly above chance, and was better than that achieved by ranking according to frequency or lexical association. We therefore concluded that lexical context can be used to highlight non-compositional items.

6.2 Issues for further research

6.2.1 Lexical acquisition

6.2.1.1 Extending to other lexicalised phrase types

In both chapters 4 and 5 we focused on the lexicalisation of particular phrase types. It would be valuable to examine how the techniques might extend to other varieties of syntactic phrase.

In chapter 4 we described a measure of the syntactic fixedness of verb phrases. The basic linguistic intuition (that the syntactic fixedness of any phrase can be defined probabilistically and should be assessed relative to the syntactic flexibility of the component words) and the formalisation (estimating this by assessing the ratio of the probability of seeing a particular syntactic variation for the phrase and the probability of seeing it for the relevant component word) are extendable to any phrase type.

The single element that was phrase specific in this work was the selection of features. We defined four kinds of variation that a verb phrase could undergo - passivisation, internal modification, event modification and determiner variation/addition/dropping, and calculated overall variation by combining these. Clearly these variations would not be relevant for other phrases types. However all that would be needed as input to extend the model would be the different phrase type to be explored and its set of relevant relations.

One of the kinds of syntactic variations that has been discussed in the literature on lexicalisation is the non-appearance of the predicative or nominative form for lexicalised adjective noun phrases. For a freely productive adjective noun combination (e.g. *green forest*) we might expect to see one or both of these forms (e.g. *the forest is green*; *the greenness of the forest*). A parser could be used to identify these. If we were working with the RASP parser then we would see the following relations:

```
(|ncsubj| |be+s:11_VBZ| |valley:10_NN1| _)
(|xcomp| _ |be+s:11_VBZ| |green:12_JJ|)
```

indicating the appearance of the predicative form. And in the case of the nominative we would look for the relation:

```
(|ncmod| |of:3_IO| |greenness:2_NN1| |forest:5_NN1|)
```

We could then calculate the probability of seeing each of the relations for a given adjective and noun pair and for each of the component words and then proceed to calculate variation in the same fashion as for verbs and objects.

If the syntactic variation measure is to be of maximum use in the construction of lexicons, then it is going to be necessary to extend it to all phrase types. Selecting features by hand for multiple phrase types is of course possible. However, it would be ideal if the feature set could be extracted for each phrase type without human intervention. It would be possible to automatically mine parsed corpora for the grammatical relations that are distinct for each phrase type and to calculate variation over these. An initial experiment then, might be to see how well such an approach performs on the extraction of VPs relative to our work with a linguistically precise feature set.

In chapter 5 we described a method for distinguishing non-compositional VPCs that relied upon very little linguistic information. We compared the lexical contexts of VPCs and their component words, where lexical context was defined over a simple window around the item of interest. Extending this to other phrase types would be very simple. We showed that there was a difference in the ability to distinguish non-compositionality for verb and for the particle. We might therefore expect to find that there would be a different performance for different phrase type composed of words of different syntactic category. One might want to use a different window size for different phrase varieties.

6.2.1.2 Extending to larger lexicalised constituents

As we discussed above, experimental work on the storage of multiword units has focused almost exclusively on sequences of two words, and the same is true for the measures of lexical association we introduced in section 4.5.1. In chapter 2 we provided numerous examples of MWEs that consisted of more than two words. And in chapter 3 we provided experimental evidence that frequent sequences of more that two words are processed more quickly than matching infrequent phrases. However when we came to discuss the identification of lexicalised phrases we looked exclusively at two word combinations. In order to cover the full range of lexicalised phrases in English, it would be necessary to extend this work to longer phrases.

The work that we described in chapter 4 involved an automatic assessment of the syntactic productivity of a particular variety of two word phrase. Because in our measure the different kinds of variation are simply added together as shown in equation 4.6, it would be simple to extend it to longer word sequences. For example if one was interested in assessing the syntactic variation of a subject-object-verb combination *abc* then one could calculate it as

$$SynVar(abc) = \sum_{i}^{n} I(aVar_i; bc|a) + \sum_{j}^{n} I(bVar_j; ac|b) + \sum_{k}^{n} I(cVar_k; ab|c) \quad (6.1)$$

Experimentation would of course be necessary to determine if this would work in practice.

The work that we described in chapter 5 involved assessing the similarity between the lexical collocates of a particular kind of two word combination (the verb-particle construction) and the lexical collocates of its component words. These similarities for the component words were combined to provide a single score for the whole phrase by either adding together the context vectors of the words or combining the similarity scores they received. Extending this to phrases of more than two words would theoretically be straightforward. It might, however, result in some loss of accuracy and, again, it would be necessary to perform an experiment to observe what effect this might have.

6.2.1.3 Combining factors in building phrasal lexicons

Throughout this thesis we have argued that the construction of any adequate phrasal lexicon must consider the separate factors of frequency, syntax and meaning. We have shown that information about these factors can be acquired from corpora, and that these can be used to help identify items that need to be included in a lexicon. While these factors do need to be considered separately, however, they will be in most cases be contributing to the construction of a single lexicon, and the insights from each will need to be combined at some point during the development process. A necessary line of continuing research will therefore be ways to best combine the information.

The most straightforward way to proceed would be to consider each different kind of information in turn, working with lists sorted by each criteria in succession, thereby making it likely that one would consider the items that most need to be included. However, doing this this way ignores the fact that there is some overlap between the varieties of MWE. In chapter 2 we presented the argument that institutionalisation and lexicalisation are different stages in the history of a lexical item. And as we would expect given this, we showed in chapter 4 that a sorted list that combined information about frequency and syntactic fixedness outperforms lists sorted by either kind of information in isolation. It seems then that the ranking method that would give quickest access to those items that have the strongest claim on a place in a lexicon (those items that belong in there on the ground of more than one dimension), might be to combine information. In our continuing work, therefore, we intend to acquire information about frequency, syntactic fixedness and semantics for a single phrase type, and to explore whether all three kinds of information can be effectively combined.

6.2.2 Other areas

6.2.2.1 Human language processing

Chapter 3 described two self-paced reading experiments in which we presented subjects with matched pairs of word sequences (in sentence contexts), one of which was frequent in the BNC, and the other of which was infrequent. In both experiments we found that subjects took less time to read the frequent sequences than the infrequent ones, controlling for the frequency of the final word and for the component transitional probabilities of the sequence.

This result demonstrates that the frequency of the sequences is reflected in human language processing. And most importantly for us it demonstrates that the frequency of phrases as found in a corpus are indicative of their real world frequencies. This conclusion was adequate for our purposes in this thesis. In order to provide a more detailed explanation for the result, however, it will be necessary to conduct further research.

There are multiple possible explanations for the sequence frequency effect. One possible explanation is that the reader stores whole form-meaning mapping for frequent phrases and the reduced reading time reflects the ease of retrieval for this single stored representation, relative to composing a meaning online for the infrequent one. This would be similar to the explanation given by Swinney and Cutler (1979) for the reduced reading time for non-compositional idioms relative to compositional word combinations.

Another explanation is that the overall effect is due to readers recognising each individual word more easily because they are more predictable in context. We of course held the individual word frequencies and the bigram transitional probabilities constant. The predictability therefore would need to come from higher order lexical dependence. This would be similar to the explanation for the effect of transitional probabilities on reading time offered by McDonald and Shillcock (2003), simply extended to longer n-grams. In the case of experiment three, where the frequent and infrequent phrases were identical except from the last word, the effect would occur due to the reader finding it easier to recognise the final word given the rest of the phrase. The claim then would be that the cooccurence frequencies of the words (of a higher order than bigrams) are stored by the reader and are utilised by her/his lexical retrieval mechanisms.

An additional factor might be at play in these experiments. We saw a considerably stronger effect in that experiment, where the sequences were identical except for the final word, than we did in experiment four where the sequences were matched for syntactic form but were lexically different. This might suggest that in experiment three the similarity of the infrequent phrase to a frequent (and therefore more familiar) phrase is interfering with processing. As the reader progresses through the phrase s/he could be expecting to see the frequent phrase, only to have this expectation broken on reaching the last word and thus be forced to reanalyse.

It is not possible, based on the current results, to distinguish between these explanations. One way forward would be to repeat the experiment using eyetracking rather than self-paced reading to measure reading time. This would provide us with valuable detailed data concerning the location of each subject's gaze, which could help us to further understand the effect. If either of the first explanations above is responsible then we would expect to see a decrease in gaze duration as the reader progressed through the sequence. In experiment three we would simply expect a difference in the gaze duration on the penultimate and final words. If the third explanation is true, however, we might expect to see some backtracking, indicating reanalysis.

6.2.2.2 Lexical change and development

Although our focus throughout this thesis has been on synchronic language description, some of our most basic distinctions have been rooted in diachronic language change. The distinction between institutionalisation and lexicalisation that we made in chapter 2 actually come from the study of language development where they refer to various stages in the historical development of lexical items. We think that it would be interesting to use the techniques discussed in this chapter to provide a broad empirical basis for examining assumptions in this area.

The statistics of usage are often claimed to lie behind lexical development (e.g. Bybee, 2003; Bybee, 2005). We discussed in chapter 4 how it has been claimed that syntactic fixedness comes about through repetition of the form. It is also been argued that repetition lies behind the existence of non-compositional phrases. It is usually assumed that through repeated usage an originally analysable phrase comes, over time to assume a function that cannot be related to its compositional meaning. Haiman (1994) explains this through analogues in other species. One example come from the mating rituals of the dancing or balloon fly. Prior to copulation the male fly presents the physically superior female with a gift of an empty balloon of silk. This ritual is explained as follows:

> "...originally, the male dancing fly distracted the predaceous female with a distracting gift of a dead insect...the male partially wrapped his tiny prey up in silk exuded from his anal glands, probably in order to subdue it: the silk, like the insect had an instrumental function, and its similarity to "wrapping" was incidental. Finally, however, the male achieved his original "purpose" by giving the female the elaborated wrapping alone, and it is the wrapping which serves as the mating signal" (p.4)

Through repetition the originally instrumental act has taken on a special symbolic function. In the same way, Haiman argues, through repetition words and phrases become “emancipated” from their instrumental function. So take for example the frequently repeated phrase *how are you?* used as a greeting in shops and restaurants across the English speaking world, and particularly in the USA. It is meant to indicate that you as a customer have the attention of the employee, who would be very surprised were you to actually burden them with details of your mental or physical state. The original purpose of ascertaining the well-being of the addressee has through repetition come to carry the symbolic function of indicating an intention to provide service. It has taken on a meaning that is different from its compositional one.

As we saw in chapter 2, in line with these claims about the role of repetition in the development of lexical items, Bauer (1983) views the history of a word or phrase as beginning with nonce formation, and progressing through repetition to institutionalisation and onto lexicalisation. According to Bauer’s account, items are required to have a certain frequency in the language at some point of their development, but once lexicalised an item can become less frequent again and retain its lexicalised status. However this pattern has been observed based on isolated examples. We have shown in this thesis that it is possible to separately assess the degree of institutionalisation and lexicalisation of all phrases of a given type. It would be interesting to use our technique to locate phrases on this line of development, and to see whether this empirical assessment agrees with more traditional data concerning the age of the item in the language. For example it would be interesting to see if we found relatively new coinages that were already lexicalised, with or without being frequent.

Appendix A

Phrases, frequencies and burstiness scores from experiment one

The tables in this appendix contain the burstiness scores for SAID idioms and their matching words as calculated in experiment two. Each table of idioms is followed by a table of single words that are matched to these phrases for frequency. The headings refer to different values as follows:

TF = Term frequency
DF = Document frequency
Var = Variance
IDF = Inverse document frequency (SparckJones 1972)
Burst = Burtiness (Katz 1996)
Entropy = Entropy (Church and Gale 1994)

Phrase	TF	DF	Var	IDF	Burst	Entropy
get out	1063	706	0.6437	2.1373	1.5057	1.0826
get into	1285	886	0.6315	1.8097	1.4503	1.2310
look after	1840	1030	1.3509	1.5924	1.7864	1.4948
set to	2264	1135	1.9568	1.4524	1.9947	1.6635
fall into	902	708	0.3928	2.1332	1.2740	0.9871
every time	1569	995	0.9377	1.6423	1.5769	1.3768
lost in	1089	804	0.5013	1.9498	1.3545	1.1126
and all	7091	2566	4.6234	0.2755	2.7634	2.8005
go with	1172	819	0.5775	1.9231	1.4310	1.1635
go home	1084	636	0.8673	2.2880	1.7044	1.0637
lit up	342	261	0.1682	3.5729	1.3103	0.5020
wreathed in	61	56	0.0225	5.7935	1.0893	0.1381
even up	26	26	0.0083	6.9004	1.0000	0.0698
succumb to	121	113	0.0426	4.7807	1.0708	0.2389
not bad	289	241	0.1230	3.6880	1.1992	0.4530
pull on	102	86	0.0472	5.1746	1.1860	0.2000
clean up	375	280	0.2021	3.4716	1.3393	0.5332
stick by	30	28	0.0109	6.7935	1.0714	0.0775
thank for	56	54	0.0190	5.8460	1.0370	0.1305
take sides	31	29	0.0118	6.7429	1.0690	0.0784
empty into	4	4	0.0013	9.6008	1.0000	0.0142
rev up	6	6	0.0019	9.0159	1.0000	0.0202
swamp with	3	3	0.0010	10.0159	1.0000	0.0111
suck up	14	14	0.0045	7.7935	1.0000	0.0416
send down	13	13	0.0042	7.9004	1.0000	0.0391
bolster up	11	11	0.0035	8.1414	1.0000	0.0339
fuck about	5	5	0.0016	9.2789	1.0000	0.0173
pop out	23	22	0.0080	7.1414	1.0455	0.0627
be oneself	6	6	0.0019	9.0159	1.0000	0.0202
dab on	2	2	0.0006	10.6008	1.0000	0.0078

Table A.1: Frequencies and variability measures for selected two-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
parade	1063	522	1.7602	2.5729	2.0364	0.9571
auction	1285	433	3.7120	2.8426	2.9677	0.9175
compete	1841	988	1.4616	1.6525	1.8634	1.4844
cotton	2264	850	4.7913	1.8695	2.6635	1.4955
retaining	902	683	0.4437	2.1851	1.3206	0.9842
employ	1565	962	1.0465	1.6909	1.6268	1.3656
coventry	1089	421	2.3127	2.8832	2.5867	0.8749
justice	7094	1557	32.6013	0.9963	4.5562	2.6038
jail	1172	467	1.8319	2.7336	2.5096	0.9612
offset	1084	583	1.1692	2.4135	1.8593	1.0193
alkaline	342	123	0.6820	4.6583	2.7805	0.3337
secede	61	35	0.0894	6.4716	1.7429	0.1039
brag	26	24	0.0096	7.0159	1.0833	0.0685
animate	121	77	0.1295	5.3341	1.5714	0.1939
eerie	289	211	0.1720	3.8797	1.3697	0.4332
irreconcilable	102	86	0.0491	5.1746	1.1860	0.2010
steak	375	214	0.3760	3.8594	1.7523	0.4719
southland	30	14	0.0360	7.7935	2.1429	0.0499
remarry	56	39	0.0377	6.3154	1.4359	0.1112
flagellation	31	24	0.0183	7.0159	1.2917	0.0714
jee	4	4	0.0013	9.6008	1.0000	0.0142
congeniality	6	6	0.0019	9.0159	1.0000	0.0202
montero	3	2	0.0016	10.6008	1.5000	0.0084
diddle	14	8	0.0103	8.6008	1.7500	0.0295
uppercut	13	10	0.0061	8.2789	1.3000	0.0341
cherubim	11	11	0.0035	8.1414	1.0000	0.0339
geodetic	5	5	0.0016	9.2789	1.0000	0.0173
impute	23	20	0.0093	7.2789	1.1500	0.0601
pursuant	6	5	0.0026	9.2789	1.2000	0.0184
ramification	2	2	0.0006	10.6008	1.0000	0.0078

Table A.2: Frequencies and variability measures for matching words for selected two-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
look forward to	902	563	0.8081	2.4639	1.6021	0.9412
as long as	999	744	0.7359	2.0617	1.3427	1.0303
take advantage of	1023	714	0.5274	2.1211	1.4328	1.0657
on the whole	1120	795	0.5577	1.9660	1.4088	1.1314
at the time	7447	2408	7.7791	0.3672	3.0926	2.9283
in time to	578	470	0.2429	2.7243	1.2298	0.7412
at the last	825	623	0.3696	2.3178	1.3242	0.9347
come up with	985	703	0.4844	2.1435	1.4011	1.0434
something of a	1155	834	0.5182	1.8969	1.3849	1.1557
as a whole	3474	1424	3.9995	1.1251	2.4396	2.0683
behind the scenes	269	240	0.0991	3.6940	1.1208	0.4350
rise and fall	189	153	0.0977	4.3435	1.2353	0.3229
any and every	29	27	0.0105	6.8460	1.0741	0.0753
with a difference	105	92	0.0449	5.0773	1.1413	0.2087
not think of	165	150	0.0606	4.3720	1.1000	0.3024
given over to	136	124	0.0509	4.6466	1.0968	0.2604
a long haul	21	21	0.0067	7.2085	1.0000	0.0585
of a kind	313	256	0.1454	3.6008	1.2227	0.4768
in bad faith	20	18	0.0077	7.4309	1.1111	0.0543
take control of	104	99	0.0356	4.9715	1.0505	0.2129
chew the cud	3	3	0.0010	10.0159	1.0000	0.0111
all too brief	10	10	0.0032	8.2789	1.0000	0.0313
lay emphasis on	5	5	0.0016	9.2789	1.0000	0.0173
at cross purposes	12	12	0.0038	8.0159	1.0000	0.0365
a fast buck	8	8	0.0026	8.6008	1.0000	0.0259
fly the flag	15	9	0.0125	8.4309	1.6667	0.0322
a double chin	8	8	0.0026	8.6008	1.0000	0.0259
carry back to	5	5	0.0016	9.2789	1.0000	0.0173
be in service	12	12	0.0038	8.0159	1.0000	0.0365
catch a cold	16	16	0.0051	7.6008	1.0000	0.0466

Table A.3: Frequencies and variability measures for selected three-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
topped	902	613	0.7309	2.3411	1.4715	0.9625
surgeon	999	457	1.6567	2.7648	2.1860	0.8962
invention	1023	547	3.1468	2.5054	1.8702	0.9406
reproduction	1120	438	5.7669	2.8261	2.5571	0.8700
truth	7462	2032	27.8458	0.6122	3.6722	2.8755
unsure	578	469	0.2648	2.7274	1.2324	0.7372
refugee	825	337	3.6774	3.2042	2.4481	0.7084
miracle	978	566	2.4875	2.4562	1.7279	0.9518
kindly	1155	717	0.7455	2.1150	1.6109	1.1311
emotional	3466	1232	7.2306	1.3341	2.8133	1.9464
flurry	269	240	0.1010	3.6940	1.1208	0.4349
pageant	189	76	1.0578	5.3529	2.4868	0.2020
superimpose	29	26	0.0112	6.9004	1.1154	0.0741
perk	105	63	0.1672	5.6236	1.6667	0.1636
rehearse	165	125	0.0838	4.6351	1.3200	0.2856
flatten	136	120	0.0612	4.6940	1.1333	0.2550
interlining	21	4	0.1052	9.6008	5.2500	0.0153
signify	313	217	0.4918	3.8393	1.4424	0.4307
fluorine	20	13	0.0173	7.9004	1.5385	0.0432
brazen	104	91	0.0491	5.0930	1.1429	0.2049
confabulation	3	3	0.0010	10.0159	1.0000	0.0111
pandanus	10	6	0.0064	9.0159	1.6667	0.0230
mandrel	5	4	0.0023	9.6008	1.2500	0.0153
wop	12	11	0.0045	8.1414	1.0909	0.0355
meld	8	8	0.0026	8.6008	1.0000	0.0259
lope	15	12	0.0074	8.0159	1.2500	0.0397
absoluteness	8	6	0.0045	9.0159	1.3333	0.0215
cacao	5	3	0.0035	10.0159	1.6667	0.0120
beatification	12	8	0.0084	8.6008	1.5000	0.0286
sextet	16	13	0.0077	7.9004	1.2308	0.0423

Table A.4: Frequencies and variability measures for matching words for selected three-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
as we have seen	1003	472	1.1040	2.7182	2.1250	0.9279
on the one hand	1353	791	1.0102	1.9733	1.7105	1.2358
in the form of	2717	1396	1.9602	1.1538	1.9463	1.8597
in the hope of	423	365	0.1653	3.0891	1.1589	0.5996
in the long run	469	359	0.2615	3.1130	1.3064	0.6294
at the expense of	1067	784	0.4934	1.9861	1.3610	1.0997
at the same time	6466	2396	4.6237	0.3744	2.6987	2.7448
in the way of	781	610	0.3556	2.3482	1.2803	0.9017
in the event of	1030	548	1.4058	2.5028	1.8796	0.9716
for the first time	5295	2158	3.3349	0.5254	2.4537	2.5430
for better or worse	39	38	0.0130	6.3529	1.0263	0.0974
in the last analysis	49	44	0.0187	6.1414	1.1136	0.1145
the next best thing	61	57	0.0225	5.7680	1.0702	0.1384
on the understanding that	84	71	0.0360	5.4511	1.1831	0.1738
have an effect on	84	71	0.0450	5.4511	1.1831	0.1704
live happily ever after	26	25	0.0089	6.9570	1.0400	0.0695
like the look of	57	53	0.0206	5.8729	1.0755	0.1312
get a grip on	34	34	0.0108	6.5134	1.0000	0.0870
the cost of living	107	85	0.0809	5.1915	1.2588	0.1985
for the love of	65	57	0.0256	5.7680	1.1404	0.1428
money for old rope	8	8	0.0026	8.6008	1.0000	0.0259
now and again then	2	2	0.0006	10.6008	1.0000	0.0078
come on the scene	11	10	0.0042	8.2789	1.1000	0.0328
be just right for	3	3	0.0010	10.0159	1.0000	0.0111
as hard as nails	5	5	0.0016	9.2789	1.0000	0.0173
backward in coming forward	3	3	0.0010	10.0159	1.0000	0.0111
peace in our time	17	14	0.0074	7.7935	1.2143	0.0450
x marks the spot	6	6	0.0019	9.0159	1.0000	0.0202
a freak of nature	3	3	0.0010	10.0159	1.0000	0.0111
take a chance on	16	15	0.0058	7.6940	1.0667	0.0458

Table A.5: Frequencies and variability measures for selected four-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
courtesy	1003	676	0.6314	2.2000	1.4837	1.0383
super	1353	615	1.6953	2.3364	2.2000	1.1305
architecture	2717	695	12.8726	2.1600	3.9094	1.3661
barrage	423	256	0.6057	3.6008	1.6523	0.5096
dynasty	469	253	0.7109	3.6178	1.8538	0.5359
psychiatric	1067	317	8.1521	3.2925	3.3659	0.6958
drink	6476	1717	11.2033	0.8552	3.7717	2.7149
ample	781	582	0.5617	2.4160	1.3419	0.8824
descent	1030	484	3.5777	2.6820	2.1281	0.8833
picked	5285	1811	5.9653	0.7783	2.9183	2.5506
dowel	39	19	0.0684	7.3529	2.0526	0.0625
unscrew	49	27	0.0554	6.8460	1.8148	0.0804
affectation	61	53	0.0244	5.8729	1.1509	0.1351
woefully	84	77	0.0308	5.3341	1.0909	0.1784
subtraction	84	55	0.0695	5.8195	1.5273	0.1499
reposed	26	25	0.0089	6.9570	1.0400	0.0695
indiscreet	57	54	0.0199	5.8460	1.0556	0.1319
curie	34	28	0.0160	6.7935	1.2143	0.0808
affable	107	93	0.0481	5.0617	1.1505	0.2107
zoned	65	40	0.0984	6.2789	1.6250	0.1131
architectonic	8	8	0.0026	8.6008	1.0000	0.0259
illusive	2	2	0.0006	10.6008	1.0000	0.0078
skit	11	10	0.0042	8.2789	1.1000	0.0328
rustler	3	3	0.0010	10.0159	1.0000	0.0111
calcification	5	5	0.0016	9.2789	1.0000	0.0173
doltish	3	3	0.0010	10.0159	1.0000	0.0111
regulative	17	11	0.0100	8.1414	1.5455	0.0386
agleam	6	6	0.0019	9.0159	1.0000	0.0202
unblushing	3	3	0.0010	10.0159	1.0000	0.0111
paunchy	16	16	0.0051	7.6008	1.0000	0.0466

Table A.6: Frequencies and variability measures for matching words for selected four-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
get to the bottom of	60	54	0.0241	5.8460	1.1111	0.1344
to name but a few	79	72	0.0293	5.4309	1.0972	0.1696
be that as it may	82	71	0.0367	5.4511	1.1549	0.1698
the face of the earth	61	54	0.0244	5.8460	1.1296	0.1365
as the case may be	162	82	0.2336	5.2433	1.9756	0.2194
as a matter of fact	327	223	0.1985	3.7999	1.4664	0.4661
the pros and cons of	95	88	0.0342	5.1414	1.0795	0.1973
a breath of fresh air	68	62	0.0272	5.6466	1.0968	0.1492
in this day and age	65	60	0.0237	5.6940	1.0833	0.1456
the end of the world	147	129	0.0618	4.5896	1.1395	0.2727
at a snail s pace	18	18	0.0058	7.4309	1.0000	0.0514
be one of the boys	9	9	0.0029	8.4309	1.0000	0.0286
by hook or by crook	14	14	0.0045	7.7935	1.0000	0.0416
nothing more or less than	7	7	0.0022	8.7935	1.0000	0.0231
put a sock in it	8	7	0.0032	8.7935	1.1429	0.0244
a nasty piece of work	9	9	0.0029	8.4309	1.0000	0.0286
the sky s the limit	10	10	0.0032	8.2789	1.0000	0.0313
it is about time that	14	14	0.0045	7.7935	1.0000	0.0416
the survival of the fittest	31	27	0.0131	6.8460	1.1481	0.0773
for days at a time	13	13	0.0042	7.9004	1.0000	0.0391
live to fight another day	6	6	0.0019	9.0159	1.0000	0.0202
come out in the wash	3	1	0.0029	11.6008	3.0000	0.0042
the law of diminishing returns	5	5	0.0016	9.2789	1.0000	0.0173
the warp and woof of	2	2	0.0006	10.6008	1.0000	0.0078
set the world to rights	3	3	0.0010	10.0159	1.0000	0.0111
a gift from the gods	3	3	0.0010	10.0159	1.0000	0.0111
as free as a bird	6	5	0.0026	9.2789	1.2000	0.0184
as bright as a button	5	5	0.0016	9.2789	1.0000	0.0173
a power in the land	3	3	0.0010	10.0159	1.0000	0.0111
a name to conjure with	3	3	0.0010	10.0159	1.0000	0.0111

Table A.7: Frequencies and variability measures for selected five-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
interlocutor	60	34	0.0717	6.5134	1.7647	0.1029
anticipatory	79	56	0.1214	5.7935	1.4107	0.1395
hellenic	82	36	0.1654	6.4309	2.2778	0.1110
sniping	61	44	0.0630	6.1414	1.3864	0.1154
summarize	162	114	0.1170	4.7680	1.4211	0.2683
sparse	327	281	0.1399	3.4664	1.1637	0.4944
shifty	95	81	0.0400	5.2610	1.1728	0.1926
thermos	68	44	0.0440	6.1414	1.5455	0.1269
sleight	65	39	0.1493	6.3154	1.6667	0.1037
tripping	147	134	0.0580	4.5348	1.0970	0.2747
bossed	18	17	0.0064	7.5134	1.0588	0.0508
vintner	9	7	0.0042	8.7935	1.2857	0.0250
browbeaten	14	13	0.0051	7.9004	1.0769	0.0407
floe	7	7	0.0022	8.7935	1.0000	0.0231
photomicrograph	8	3	0.0122	10.0159	2.6667	0.0120
sniffle	9	9	0.0029	8.4309	1.0000	0.0286
imbibe	10	9	0.0039	8.4309	1.1111	0.0301
virginian	14	13	0.0051	7.9004	1.0769	0.0407
rightist	31	22	0.0189	7.1414	1.4091	0.0687
compote	13	9	0.0087	8.4309	1.4444	0.0315
unimpeachably	6	6	0.0019	9.0159	1.0000	0.0202
sublunary	3	3	0.0010	10.0159	1.0000	0.0111
acidulous	5	5	0.0016	9.2789	1.0000	0.0173
maggoty	2	2	0.0006	10.6008	1.0000	0.0078
viator	3	2	0.0016	10.6008	1.5000	0.0084
matriculate	3	3	0.0010	10.0159	1.0000	0.0111
undrinkable	6	6	0.0019	9.0159	1.0000	0.0202
syndic	5	4	0.0023	9.6008	1.2500	0.0153
hereunto	3	2	0.0016	10.6008	1.5000	0.0084
varmint	3	1	0.0029	11.6008	3.0000	0.0042

Table A.8: Frequencies and variability measures for matching words for selected five-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
at the drop of a hat	28	27	0.0096	6.8460	1.0370	0.0740
a force to be reckoned with	39	38	0.0130	6.3529	1.0263	0.0974
on the spur of the moment	66	64	0.0221	5.6008	1.0313	0.1490
if you know what i mean	45	41	0.0175	6.2433	1.0976	0.1072
at one and the same time	81	72	0.0312	5.4309	1.1250	0.1715
by the scruff of the neck	30	29	0.0102	6.7429	1.0345	0.0784
the fact of the matter is	22	21	0.0077	7.2085	1.0476	0.0603
only the tip of the iceberg	39	38	0.0130	6.3529	1.0263	0.0974
that is not to say that	67	58	0.0282	5.7429	1.1552	0.1458
have a long way to go	41	41	0.0130	6.2433	1.0000	0.1013
all s well that ends well	13	11	0.0055	8.1414	1.1818	0.0364
at the end of the rainbow	5	5	0.0016	9.2789	1.0000	0.0173
live to a ripe old age	5	5	0.0016	9.2789	1.0000	0.0173
have nothing better to do than	8	8	0.0026	8.6008	1.0000	0.0259
a hard day at the office	5	4	0.0023	9.6008	1.2500	0.0153
as poor as a church mouse	5	5	0.0016	9.2789	1.0000	0.0173
for all the world to see	5	5	0.0016	9.2789	1.0000	0.0173
in the twinkling of an eye	14	14	0.0045	7.7935	1.0000	0.0416
at all hours of the day	15	14	0.0054	7.7935	1.0714	0.0433
every cloud has a silver lining	7	6	0.0029	9.0159	1.1667	0.0215
make hay while the sun shines	4	4	0.0013	9.6008	1.0000	0.0142
live under the same roof as	2	2	0.0006	10.6008	1.0000	0.0078
as mad as a march hare	2	2	0.0006	10.6008	1.0000	0.0078
at all hours of the night	3	3	0.0010	10.0159	1.0000	0.0111
be a sight for sore eyes	2	2	0.0006	10.6008	1.0000	0.0078
all the world and his wife	2	2	0.0006	10.6008	1.0000	0.0078
have a hard row to hoe	2	2	0.0006	10.6008	1.0000	0.0078
full of the joys of spring	3	3	0.0010	10.0159	1.0000	0.0111
this year next year sometime never	2	2	0.0006	10.6008	1.0000	0.0078
needs must when the devil drives	2	2	0.0006	10.6008	1.0000	0.0078

Table A.9: Frequencies and variability measures for selected six word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
fastens	28	26	0.0109	6.9004	1.0769	0.0718
camelot	39	32	0.0201	6.6008	1.2188	0.0883
interstellar	66	39	0.0498	6.3154	1.6923	0.1174
providential	45	30	0.0375	6.6940	1.5000	0.0883
uncommitted	81	61	0.1007	5.6701	1.3279	0.1482
foretell	30	27	0.0134	6.8460	1.1111	0.0740
gainful	22	18	0.0115	7.4309	1.2222	0.0550
reedy	39	32	0.0176	6.6008	1.2188	0.0912
stewardess	67	50	0.0346	5.9570	1.3400	0.1360
chamois	41	32	0.0246	6.6008	1.2813	0.0903
tindal	13	9	0.0080	8.4309	1.4444	0.0308
underhanded	5	5	0.0016	9.2789	1.0000	0.0173
screechy	5	5	0.0016	9.2789	1.0000	0.0173
unguided	8	8	0.0026	8.6008	1.0000	0.0259
outfielder	5	4	0.0023	9.6008	1.2500	0.0153
jager	5	5	0.0016	9.2789	1.0000	0.0173
endogamy	5	5	0.0016	9.2789	1.0000	0.0173
manumission	14	4	0.0399	9.6008	3.5000	0.0153
strophe	15	3	0.0364	10.0159	5.0000	0.0126
geocentric	7	4	0.0061	9.6008	1.7500	0.0153
hetman	4	3	0.0019	10.0159	1.3333	0.0120
capercailzie	2	1	0.0013	11.6008	2.0000	0.0042
mennonite	2	2	0.0006	10.6008	1.0000	0.0078
combinable	3	3	0.0010	10.0159	1.0000	0.0111
omelet	2	1	0.0013	11.6008	2.0000	0.0042
supernaturalism	2	2	0.0006	10.6008	1.0000	0.0078
kaddish	2	2	0.0006	10.6008	1.0000	0.0078
allegoric	3	3	0.0010	10.0159	1.0000	0.0111
unuttered	2	2	0.0006	10.6008	1.0000	0.0078
secant	2	1	0.0013	11.6008	2.0000	0.0042

Table A.10: Frequencies and variability measures for matching words for selected six word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
now i come to think of it	16	15	0.0058	7.6940	1.0667	0.0458
with the best will in the world	24	23	0.0083	7.0773	1.0435	0.0650
what do you think you re doing	34	30	0.0134	6.6940	1.1333	0.0840
if the worst comes to the worst	20	19	0.0070	7.3529	1.0526	0.0556
be glad to see the back of	8	7	0.0032	8.7935	1.1429	0.0244
like a bull in a china shop	4	4	0.0013	9.6008	1.0000	0.0142
like a bear with a sore head	5	5	0.0016	9.2789	1.0000	0.0173
like a red rag to a bull	8	7	0.0032	8.7935	1.1429	0.0244
a land flowing with milk and honey	6	4	0.0039	9.6008	1.5000	0.0153
the long and the short of it	6	6	0.0019	9.0159	1.0000	0.0202
the valley of the shadow of death	6	5	0.0026	9.2789	1.2000	0.0184
be the thin end of the wedge	7	6	0.0029	9.0159	1.1667	0.0215
the slings and arrows of outrageous fortune	4	4	0.0013	9.6008	1.0000	0.0142
nice work if you can get it	4	4	0.0013	9.6008	1.0000	0.0142
make the best of a bad job	4	4	0.0013	9.6008	1.0000	0.0142
the best things in life are free	5	5	0.0016	9.2789	1.0000	0.0173
it is just one of those things	5	5	0.0016	9.2789	1.0000	0.0173
do as you would be done by	4	4	0.0013	9.6008	1.0000	0.0142
how many times do i have to	10	10	0.0032	8.2789	1.0000	0.0313
the life and soul of the party	10	10	0.0032	8.2789	1.0000	0.0313
those whom the gods love die young	2	2	0.0006	10.6008	1.0000	0.0078
things that go bump in the night	3	3	0.0010	10.0159	1.0000	0.0111
did he fall or was he pushed	2	2	0.0006	10.6008	1.0000	0.0078
with an eye for the main chance	2	2	0.0006	10.6008	1.0000	0.0078
spare the rod and spoil the child	3	3	0.0010	10.0159	1.0000	0.0111
the world the flesh and the devil	2	2	0.0006	10.6008	1.0000	0.0078
not for all the tea in china	2	2	0.0006	10.6008	1.0000	0.0078
a square peg in a round hole	2	2	0.0006	10.6008	1.0000	0.0078
tell the truth and shame the devil	2	2	0.0006	10.6008	1.0000	0.0078
have a lot to put up with	2	2	0.0006	10.6008	1.0000	0.0078

Table A.11: Frequencies and variability measures for selected seven-word SAID idioms

Phrase	TF	DF	Var	IDF	Burst	Entropy
flapper	16	12	0.0116	8.0159	1.3333	0.0381
syrupy	24	24	0.0077	7.0159	1.0000	0.0653
polka	34	32	0.0121	6.6008	1.0625	0.0863
imperturbable	20	20	0.0064	7.2789	1.0000	0.0561
penurious	8	8	0.0026	8.6008	1.0000	0.0259
avocation	4	4	0.0013	9.6008	1.0000	0.0142
foothill	5	4	0.0023	9.6008	1.2500	0.0153
clannish	8	8	0.0026	8.6008	1.0000	0.0259
costive	6	6	0.0019	9.0159	1.0000	0.0202
mucilage	6	4	0.0039	9.6008	1.5000	0.0153
moccasin	6	5	0.0026	9.2789	1.2000	0.0184
splintery	7	6	0.0029	9.0159	1.1667	0.0215
capsicum	4	3	0.0019	10.0159	1.3333	0.0120
soapsuds	4	4	0.0013	9.6008	1.0000	0.0142
cyclorama	4	3	0.0019	10.0159	1.3333	0.0120
malediction	5	5	0.0016	9.2789	1.0000	0.0173
inapt	5	5	0.0016	9.2789	1.0000	0.0173
pampa	4	4	0.0013	9.6008	1.0000	0.0142
methuselah	10	10	0.0032	8.2789	1.0000	0.0313
carnality	10	10	0.0032	8.2789	1.0000	0.0313
crewel	2	2	0.0006	10.6008	1.0000	0.0078
dogberry	3	3	0.0010	10.0159	1.0000	0.0111
seminole	2	2	0.0006	10.6008	1.0000	0.0078
bunt	2	2	0.0006	10.6008	1.0000	0.0078
clonic	3	3	0.0010	10.0159	1.0000	0.0111
unappeasable	2	2	0.0006	10.6008	1.0000	0.0078
ogress	2	2	0.0006	10.6008	1.0000	0.0078
artillerist	2	2	0.0006	10.6008	1.0000	0.0078
envenomed	2	2	0.0006	10.6008	1.0000	0.0078
necromantic	2	2	0.0006	10.6008	1.0000	0.0078

Table A.12: Frequencies and variability measures for matching words for selected seven-word SAID idioms

Appendix B

Phrases, frequencies and burstiness scores from experiment two

The tables in this appendix contain the burstiness scores for multiword sequences and their matching words as calculated in experiment two. Each table of multiword items is followed by a table of single words that are matched to these phrases for frequency. The headings refer to different values as follows:

TF = Term frequency
DF = Document frequency
Var = Variance
IDF = Inverse document frequency (SparckJones 1972)
Burst = Burtiness (Katz 1996)
Entropy = Entropy (Church and Gale 1994)

Phrase	TF	DF	Var	IDF	Burst	Entropy
she thought	3996	774	17.4073	2.0047	5.1628	1.6617
a person	4590	1355	17.2926	1.1968	3.3875	2.1387
a day	6520	2215	5.9309	0.4878	2.9436	2.7928
at this	8548	2639	6.9411	0.2351	3.2391	3.0468
an important	5843	2082	5.0010	0.5771	2.8064	2.6784
into his	3963	1404	5.1277	1.1455	2.8226	2.1741
such a	18447	3013	20.7655	0.0439	6.1225	3.9180
of this	32816	3075	69.6681	0.0145	10.6719	4.7010
about the	32781	3069	43.5323	0.0173	10.6813	4.5956
prepared to	5140	2117	3.7276	0.5530	2.4280	2.5108
woman that	147	130	0.0605	4.5785	1.1308	0.2736
the memorial	202	131	0.1831	4.5674	1.5420	0.3034
to regard	669	511	0.3261	2.6037	1.3092	0.8112
well received	200	169	0.0854	4.2000	1.1834	0.3441
and saturday	142	115	0.0752	4.7554	1.2348	0.2554
in art	419	209	0.8063	3.8935	2.0048	0.4696
ground at	231	200	0.0894	3.9570	1.1550	0.3859
sitting down	314	271	0.1340	3.5187	1.1587	0.4818
coins in	72	58	0.0548	5.7429	1.2414	0.1441
been studying	100	95	0.0344	5.0310	1.0526	0.2064
based directly	9	9	0.0029	8.4309	1.0000	0.0286
that midnight	5	5	0.0016	9.2789	1.0000	0.0173
horizon is	57	37	0.0560	6.3914	1.5405	0.1067
was protected	65	62	0.0224	5.6466	1.0484	0.1468
were straight	14	12	0.0058	8.0159	1.1667	0.0390
on supplying	10	10	0.0032	8.2789	1.0000	0.0313
or smaller	47	44	0.0175	6.1414	1.0682	0.1117
extending elements	2	2	0.0006	10.6008	1.0000	0.0078
of allegation	2	2	0.0006	10.6008	1.0000	0.0078
this controversy	22	22	0.0070	7.1414	1.0000	0.0608

Table B.1: Frequencies and variability measures for selected two-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
border	3997	1220	8.8247	1.3482	3.2762	2.0453
criticism	4591	1588	15.0138	0.9678	2.8911	2.2783
technical	6521	1772	13.9756	0.8097	3.6800	2.7122
teaching	8544	1395	89.7660	1.1548	6.1247	2.4674
football	5846	1115	46.9101	1.4780	5.2430	2.0862
funding	3968	1035	11.8425	1.5854	3.8338	1.9296
meeting	18433	2529	94.8106	0.2965	7.2887	3.9991
large	33035	3031	94.6977	0.0353	10.8990	4.7719
seen	32564	3080	52.2543	0.0121	10.5727	4.5745
moral	5145	1300	32.3459	1.2565	3.9577	2.1399
overwhelm	147	132	0.0593	4.5564	1.1136	0.2743
consequential	202	137	0.1838	4.5028	1.4745	0.3116
lateral	669	244	4.2186	3.6701	2.7418	0.5462
pierce	200	144	0.1311	4.4309	1.3889	0.3226
nestling	142	122	0.0623	4.6701	1.1639	0.2638
slam	419	244	0.5256	3.6701	1.7172	0.5135
mop	231	171	0.1912	4.1830	1.3509	0.3588
parasite	314	109	0.9138	4.8327	2.8807	0.2893
hunk	72	55	0.0387	5.8195	1.3091	0.1458
tenable	100	93	0.0357	5.0617	1.0753	0.2056
chomp	9	9	0.0029	8.4309	1.0000	0.0286
backstitch	5	4	0.0023	9.6008	1.2500	0.0153
interstitial	57	33	0.0463	6.5564	1.7273	0.1018
siesta	65	41	0.0533	6.2433	1.5854	0.1184
nob	14	12	0.0058	8.0159	1.1667	0.0390
metaphysic	10	3	0.0212	10.0159	3.3333	0.0120
pyramidal	47	41	0.0201	6.2433	1.1463	0.1087
expurgation	2	2	0.0006	10.6008	1.0000	0.0078
artillerist	2	2	0.0006	10.6008	1.0000	0.0078
puerile	22	22	0.0070	7.1414	1.0000	0.0608

Table B.2: Frequencies and variability measures for matching words for selected two-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
that it has	1264	823	0.7372	1.9161	1.5358	1.2094
not mean that	732	532	0.3830	2.5456	1.3759	0.8572
at the moment	2939	1416	2.1224	1.1332	2.0756	1.9373
they were not	1468	1019	0.6820	1.6079	1.4406	1.3313
and many of	588	498	0.2192	2.6408	1.1807	0.7512
the power of	1975	1049	1.8027	1.5660	1.8827	1.5382
to use the	2389	1416	1.4976	1.1332	1.6871	1.7302
has been a	3012	1484	2.0892	1.0656	2.0296	1.9650
the impression that	786	614	0.3513	2.3387	1.2801	0.9061
for a moment	3451	952	7.4181	1.7060	3.6250	1.8039
that was once	55	53	0.0187	5.8729	1.0377	0.1286
to his knees	147	114	0.0754	4.7680	1.2895	0.2627
a strong feeling	60	57	0.0215	5.7680	1.0526	0.1368
in the classical	125	82	0.1062	5.2433	1.5244	0.2083
act did not	34	31	0.0128	6.6466	1.0968	0.0852
already been set	28	28	0.0089	6.7935	1.0000	0.0742
is the easiest	48	44	0.0184	6.1414	1.0909	0.1133
and so one	28	28	0.0089	6.7935	1.0000	0.0742
would have ended	28	28	0.0089	6.7935	1.0000	0.0742
to be top	21	21	0.0067	7.2085	1.0000	0.0585
change may have	6	6	0.0019	9.0159	1.0000	0.0202
and very exciting	2	2	0.0006	10.6008	1.0000	0.0078
of overtaking it	2	2	0.0006	10.6008	1.0000	0.0078
which he directed	6	6	0.0019	9.0159	1.0000	0.0202
finish his drink	3	3	0.0010	10.0159	1.0000	0.0111
if someone could	9	9	0.0029	8.4309	1.0000	0.0286
we put formula	5	3	0.0035	10.0159	1.6667	0.0120
indulge their taste	2	1	0.0013	11.6008	2.0000	0.0042
as closer to	5	5	0.0016	9.2789	1.0000	0.0173
our request we	3	3	0.0010	10.0159	1.0000	0.0111

Table B.3: Frequencies and variability measures for selected three-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
embassy	1264	450	2.8312	2.7871	2.8089	0.9496
consultative	732	326	1.2497	3.2521	2.2454	0.7003
truly	2940	1562	1.8729	0.9917	1.8822	1.9365
grasp	1468	955	0.8474	1.7015	1.5372	1.3243
finn	588	95	14.4445	5.0310	6.1895	0.2557
proceed	1971	1056	2.0490	1.5564	1.8665	1.5258
sensible	2391	1360	1.4837	1.1915	1.7581	1.7417
statistics	3012	1085	9.8529	1.5174	2.7760	1.7495
fuss	786	586	0.3661	2.4061	1.3413	0.9041
global	3452	873	27.6515	1.8310	3.9542	1.6322
misshapen	55	52	0.0193	5.9004	1.0577	0.1281
petite	147	104	0.1056	4.9004	1.4135	0.2480
reprehensible	60	56	0.0222	5.7935	1.0714	0.1366
fluoride	125	31	1.6078	6.6466	4.0323	0.0963
rascal	34	32	0.0121	6.6008	1.0625	0.0863
dervish	28	18	0.0205	7.4309	1.5556	0.0588
dissect	48	42	0.0210	6.2085	1.1429	0.1105
levity	28	24	0.0115	7.0159	1.1667	0.0703
metier	28	26	0.0109	6.9004	1.0769	0.0718
calculable	21	19	0.0080	7.3529	1.1053	0.0567
congeniality	6	6	0.0019	9.0159	1.0000	0.0202
keelson	2	2	0.0006	10.6008	1.0000	0.0078
flatland	2	2	0.0006	10.6008	1.0000	0.0078
kiang	6	4	0.0039	9.6008	1.5000	0.0153
surtout	3	3	0.0010	10.0159	1.0000	0.0111
chowder	9	9	0.0029	8.4309	1.0000	0.0286
endogamy	5	5	0.0016	9.2789	1.0000	0.0173
luster	2	2	0.0006	10.6008	1.0000	0.0078
acidulous	5	5	0.0016	9.2789	1.0000	0.0173
hereunto	3	2	0.0016	10.6008	1.5000	0.0084

Table B.4: Frequencies and variability measures for matching words for selected three-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
for the rest of	1277	918	0.5608	1.7585	1.3911	1.2277
at the top of	1445	951	0.8798	1.7075	1.5195	1.3078
little more than a	432	382	0.1584	3.0234	1.1309	0.6092
had nothing to do	304	279	0.1070	3.4767	1.0896	0.4752
in an effort to	683	497	0.3873	2.6437	1.3742	0.8138
in which it is	350	291	0.1509	3.4160	1.2027	0.5211
to the fact that	970	739	0.4170	2.0714	1.3126	1.0385
on the side of	555	469	0.2195	2.7274	1.1834	0.7211
to get on with	350	298	0.1431	3.3817	1.1745	0.5229
it is possible that	750	461	0.5528	2.7522	1.6269	0.8320
i will not be	64	61	0.0221	5.6701	1.0492	0.1450
with the choice of	35	34	0.0118	6.5134	1.0294	0.0891
this has to be	100	95	0.0344	5.0310	1.0526	0.2064
the fact that even	30	30	0.0096	6.6940	1.0000	0.0785
to fall in love	92	77	0.0416	5.3341	1.1948	0.1861
in the publication of	31	30	0.0105	6.6940	1.0333	0.0806
that the board of	25	19	0.0138	7.3529	1.3158	0.0594
the ability to produce	26	22	0.0115	7.1414	1.1818	0.0657
was to be allowed	21	21	0.0067	7.2085	1.0000	0.0585
you do not wish	30	29	0.0102	6.7429	1.0345	0.0784
one hand across the	4	4	0.0013	9.6008	1.0000	0.0142
are absent altogether and	2	2	0.0006	10.6008	1.0000	0.0078
could perhaps best be	2	2	0.0006	10.6008	1.0000	0.0078
was a major step	10	10	0.0032	8.2789	1.0000	0.0313
into the mugs and	2	1	0.0013	11.6008	2.0000	0.0042
level of enthusiasm for	2	2	0.0006	10.6008	1.0000	0.0078
away with another man	3	3	0.0010	10.0159	1.0000	0.0111
central theme of this	5	5	0.0016	9.2789	1.0000	0.0173
will be reflected by	4	4	0.0013	9.6008	1.0000	0.0142
is being sponsored by	17	16	0.0061	7.6008	1.0625	0.0483

Table B.5: Frequencies and variability measures for selected four-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
stamp	1277	720	2.4868	2.1090	1.7736	1.1287
regardless	1445	911	0.9159	1.7695	1.5862	1.3048
novice	432	253	0.4494	3.6178	1.7075	0.5316
scatter	304	206	0.3008	3.9143	1.4757	0.4254
spoon	683	326	1.1271	3.2521	2.0951	0.6768
haze	350	261	0.1837	3.5729	1.3410	0.5076
rebellion	970	502	1.0074	2.6293	1.9323	0.9364
spider	555	265	0.8950	3.5510	2.0943	0.5823
certification	350	150	1.2049	4.3720	2.3333	0.3668
fairy	750	410	1.6919	2.9214	1.8293	0.7494
indigestible	64	54	0.0363	5.8460	1.1852	0.1354
flaky	35	33	0.0124	6.5564	1.0606	0.0884
boniface	100	31	0.4806	6.6466	3.2258	0.0944
occidental	30	25	0.0128	6.9570	1.2000	0.0734
pout	92	59	0.1595	5.7182	1.5593	0.1523
galena	31	13	0.0640	7.9004	2.3846	0.0439
pontifical	25	20	0.0131	7.2789	1.2500	0.0609
bandstand	26	22	0.0128	7.1414	1.1818	0.0645
inboard	21	18	0.0086	7.4309	1.1667	0.0552
elk	30	26	0.0128	6.9004	1.1538	0.0750
huckster	4	3	0.0019	10.0159	1.3333	0.0120
tillet	2	2	0.0006	10.6008	1.0000	0.0078
scarify	2	1	0.0013	11.6008	2.0000	0.0042
carnality	10	10	0.0032	8.2789	1.0000	0.0313
omelet	2	1	0.0013	11.6008	2.0000	0.0042
oxalate	2	2	0.0006	10.6008	1.0000	0.0078
pestilent	3	3	0.0010	10.0159	1.0000	0.0111
sentimentalize	5	5	0.0016	9.2789	1.0000	0.0173
tetragonal	4	3	0.0019	10.0159	1.3333	0.0120
abjection	17	8	0.0215	8.6008	2.1250	0.0299

Table B.6: Frequencies and variability measures for matching words for selected four-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
with the rest of the	352	312	0.1307	3.3154	1.1282	0.5281
the far end of the	278	207	0.1433	3.9074	1.3430	0.4257
person per night single room	225	3	6.4059	10.0159	75.0000	0.0126
the end of the first	235	208	0.0938	3.9004	1.1298	0.3901
if he will make a	226	18	1.0630	7.4309	12.5556	0.0721
in the course of the	405	301	0.2222	3.3672	1.3455	0.5618
the other side of the	1095	746	0.6198	2.0578	1.4678	1.1074
in the same way as	638	476	0.3693	2.7060	1.3403	0.7746
at the back of the	586	436	0.3295	2.8327	1.3440	0.7342
in other parts of the	232	203	0.0923	3.9355	1.1429	0.3884
early hours of the morning	116	102	0.0475	4.9284	1.1373	0.2277
that at the beginning of	18	18	0.0058	7.4309	1.0000	0.0514
was one of the largest	24	23	0.0083	7.0773	1.0435	0.0650
all she could think of	28	25	0.0109	6.9570	1.1200	0.0718
other side of the road	67	60	0.0256	5.6940	1.1167	0.1476
is a direct result of	16	15	0.0058	7.6940	1.0667	0.0458
can i do for you	77	70	0.0287	5.4716	1.1000	0.1660
in the territory of the	24	9	0.0341	8.4309	2.6667	0.0346
to walk all the way	15	15	0.0048	7.6940	1.0000	0.0441
at the end of their	105	89	0.0797	5.1251	1.1798	0.1971
there is enough to suggest	2	2	0.0006	10.6008	1.0000	0.0078
better off than they are	2	2	0.0006	10.6008	1.0000	0.0078
it gave him time to	5	5	0.0016	9.2789	1.0000	0.0173
did want to go to	3	3	0.0010	10.0159	1.0000	0.0111
that a couple of years	4	4	0.0013	9.6008	1.0000	0.0142
was a dream come true	8	8	0.0026	8.6008	1.0000	0.0259
more of the same in	8	8	0.0026	8.6008	1.0000	0.0259
have been looking forward to	13	13	0.0042	7.9004	1.0000	0.0391
which had been formed in	12	11	0.0045	8.1414	1.0909	0.0355
keep you on your toes	4	4	0.0013	9.6008	1.0000	0.0142

Table B.7: Frequencies and variability measures for selected five-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
adhesive	352	136	0.8835	4.5134	2.5882	0.3426
heaving	278	215	0.1427	3.8526	1.2930	0.4304
depict	225	163	0.1445	4.2521	1.3804	0.3543
whence	235	176	0.1446	4.1414	1.3352	0.3709
itinerary	226	141	0.1988	4.4613	1.6028	0.3282
dial	405	262	0.3182	3.5674	1.5458	0.5323
occurrence	1095	533	1.6082	2.5429	2.0544	0.9777
ninety	638	391	0.6487	2.9898	1.6317	0.7226
alley	586	286	1.0700	3.4410	2.0490	0.6008
secession	232	92	0.6358	5.0773	2.5217	0.2538
crafty	116	104	0.0456	4.9004	1.1154	0.2284
soothsayer	18	12	0.0122	8.0159	1.5000	0.0412
bearish	24	18	0.0180	7.4309	1.3333	0.0550
bludgeon	28	27	0.0096	6.8460	1.0370	0.0740
meteoric	67	58	0.0340	5.7429	1.1552	0.1426
cerulean	16	9	0.0167	8.4309	1.7778	0.0315
thermodynamic	77	18	0.2547	7.4309	4.2778	0.0666
unalloyed	24	23	0.0083	7.0773	1.0435	0.0650
recusant	15	13	0.0061	7.9004	1.1538	0.0417
tasteless	105	93	0.0410	5.0617	1.1290	0.2114
bassinet	2	2	0.0006	10.6008	1.0000	0.0078
slanderer	2	2	0.0006	10.6008	1.0000	0.0078
sclerotic	5	5	0.0016	9.2789	1.0000	0.0173
outclass	3	2	0.0016	10.6008	1.5000	0.0084
bogy	4	4	0.0013	9.6008	1.0000	0.0142
punchbowl	8	5	0.0045	9.2789	1.6000	0.0188
mistrial	8	6	0.0039	9.0159	1.3333	0.0220
laotian	13	11	0.0055	8.1414	1.1818	0.0364
chive	12	8	0.0071	8.6008	1.5000	0.0292
capsicum	4	3	0.0019	10.0159	1.3333	0.0120

Table B.8: Frequencies and variability measures for matching words for selected five-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
last time on turnover that rose	111	18	0.2991	7.4309	6.1667	0.0686
at the end of the day	432	372	0.1635	3.0617	1.1613	0.6090
to the other side of the	104	94	0.0394	5.0463	1.1064	0.2110
from the point of view of	461	342	0.2378	3.1830	1.3480	0.6207
at the end of the first	116	102	0.0469	4.9284	1.1373	0.2276
in the early hours of the	95	89	0.0335	5.1251	1.0674	0.1978
in the second half of the	267	214	0.1230	3.8594	1.2477	0.4230
in such a way as to	342	279	0.1637	3.4767	1.2258	0.5078
on security and operation in europe	100	59	0.0737	5.7182	1.6949	0.1651
if he will make a statement	213	18	0.9583	7.4309	11.8333	0.0721
list his official engagements for tuesday	93	13	0.2525	7.9004	7.1538	0.0505
all the circumstances of the case	32	21	0.0218	7.2085	1.5238	0.0673
for the rest of her life	71	67	0.0249	5.5348	1.0597	0.1572
to meet the needs of the	72	66	0.0265	5.5564	1.0909	0.1577
have something to do with the	40	40	0.0127	6.2789	1.0000	0.0993
is not the end of the	37	35	0.0131	6.4716	1.0571	0.0927
child in the family of four	11	9	0.0048	8.4309	1.2222	0.0308
came at the end of a	11	11	0.0035	8.1414	1.0000	0.0339
the afternoon and the period between	14	2	0.0315	10.6008	7.0000	0.0078
in the family of three sons	38	18	0.0327	7.4309	2.1111	0.0624
is more than i can say	7	7	0.0022	8.7935	1.0000	0.0231
in the case of the earth	8	3	0.0122	10.0159	2.6667	0.0120
first time that i had ever	2	2	0.0006	10.6008	1.0000	0.0078
the terms upon which it was	2	1	0.0013	11.6008	2.0000	0.0042
to the point at which it	6	6	0.0019	9.0159	1.0000	0.0202
and made my way back to	3	3	0.0010	10.0159	1.0000	0.0111
the smallest pin could be heard	2	1	0.0013	11.6008	2.0000	0.0042
made it clear that it is	4	3	0.0019	10.0159	1.3333	0.0120
which was at the end of	2	2	0.0006	10.6008	1.0000	0.0078
to be more than just a	10	10	0.0032	8.2789	1.0000	0.0313

Table B.9: Frequencies and variability measures for selected six word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
patter	111	99	0.0435	4.9715	1.1212	0.2211
fancied	432	330	0.2273	3.2345	1.3091	0.5935
disrespect	104	96	0.0394	5.0159	1.0833	0.2106
screwed	461	303	0.4348	3.3577	1.5215	0.5848
poplar	116	74	0.1036	5.3914	1.5676	0.1941
wily	95	84	0.0393	5.2085	1.1310	0.1944
pulp	267	174	0.2125	4.1579	1.5345	0.3854
avail	342	292	0.1443	3.4110	1.1712	0.5129
overlying	100	53	0.1355	5.8729	1.8868	0.1511
aspire	213	186	0.0851	4.0617	1.1452	0.3643
apathetic	93	86	0.0336	5.1746	1.0814	0.1939
melange	32	32	0.0102	6.6008	1.0000	0.0828
rabid	71	58	0.0359	5.7429	1.2241	0.1481
repent	72	61	0.0310	5.6701	1.1803	0.1535
augustan	40	27	0.0282	6.8460	1.4815	0.0824
tarry	37	26	0.0246	6.9004	1.4231	0.0787
bisque	11	8	0.0055	8.6008	1.3750	0.0283
misbegotten	11	10	0.0042	8.2789	1.1000	0.0328
statesmanlike	14	14	0.0045	7.7935	1.0000	0.0416
effectual	38	32	0.0179	6.6008	1.1875	0.0894
splintery	7	6	0.0029	9.0159	1.1667	0.0215
unfitting	8	8	0.0026	8.6008	1.0000	0.0259
capercailzie	2	1	0.0013	11.6008	2.0000	0.0042
mantic	2	1	0.0013	11.6008	2.0000	0.0042
burro	6	4	0.0032	9.6008	1.5000	0.0155
incise	3	2	0.0016	10.6008	1.5000	0.0084
malapropism	2	2	0.0006	10.6008	1.0000	0.0078
glissade	4	3	0.0019	10.0159	1.3333	0.0120
envenomed	2	2	0.0006	10.6008	1.0000	0.0078
lascar	10	5	0.0103	9.2789	2.0000	0.0195

Table B.10: Frequencies and variability measures for matching words for selected six word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
ask the prime minister if he will	179	16	0.7549	7.6008	11.1875	0.0627
example of this option is shown in	88	4	1.0224	9.6008	22.0000	0.0168
conference on security and operation in europe	100	59	0.0737	5.7182	1.6949	0.1651
will list his official engagements for tuesday	92	13	0.2490	7.9004	7.0769	0.0509
reported net profits for the year to	10	9	0.0039	8.4309	1.1111	0.0301
to give a true and fair view	29	17	0.0253	7.5134	1.7059	0.0562
the sale or supply of alcoholic liquor	12	2	0.0232	10.6008	6.0000	0.0078
the largest of its kind in the	10	10	0.0032	8.2789	1.0000	0.0313
the edge of the cake drum securing	13	2	0.0273	10.6008	6.5000	0.0084
is created in the directory specified in	22	1	0.1558	11.6008	22.0000	0.0042
in the afternoon and the period between	14	2	0.0315	10.6008	7.0000	0.0078
with the terms of this contract the	9	2	0.0171	10.6008	4.5000	0.0084
the chairman of the revolutionary command council	12	3	0.0212	10.0159	4.0000	0.0126
ask the secretary of state for the	74	8	0.3510	8.6008	9.2500	0.0336
and have an extra week free on	10	2	0.0167	10.6008	5.0000	0.0084
the last decade of the nineteenth century	10	10	0.0032	8.2789	1.0000	0.0313
following cases are referred to in the	59	20	0.0663	7.2789	2.9500	0.0707
it was only a matter of time	55	51	0.0200	5.9284	1.0784	0.1273
of the association of east asian nations	14	13	0.0051	7.9004	1.0769	0.0407
protocols of the learned elders of zion	5	2	0.0042	10.6008	2.5000	0.0084
deaf and dumb of which he was	2	1	0.0013	11.6008	2.0000	0.0042
of between one day and a year	2	1	0.0013	11.6008	2.0000	0.0042
the date of such letting or permission	2	1	0.0013	11.6008	2.0000	0.0042
be released from custody on the ground	2	1	0.0013	11.6008	2.0000	0.0042
only four known examples of the form	2	1	0.0013	11.6008	2.0000	0.0042
see me as a bit of a	2	2	0.0006	10.6008	1.0000	0.0078
were reported to have been killed and	4	4	0.0013	9.6008	1.0000	0.0142
a line then the cursor will be	2	1	0.0013	11.6008	2.0000	0.0042
was expecting them to unite behind a	2	1	0.0013	11.6008	2.0000	0.0042

Table B.11: Frequencies and variability measures for selected seven-word sequences

Phrase	TF	DF	Var	IDF	Burst	Entropy
epitome	179	149	0.1039	4.3817	1.2013	0.3078
premonition	88	73	0.0507	5.4110	1.2055	0.1748
nab	100	26	0.6815	6.9004	3.8462	0.0855
saucy	92	74	0.0436	5.3914	1.2432	0.1833
samovar	10	7	0.0071	8.7935	1.4286	0.0244
warty	29	15	0.0685	7.6940	1.9333	0.0475
misconstruction	12	7	0.0135	8.7935	1.7143	0.0244
milligram	10	10	0.0032	8.2789	1.0000	0.0313
bedpost	13	8	0.0138	8.6008	1.6250	0.0273
yogi	22	17	0.0115	7.5134	1.2941	0.0536
chide	14	14	0.0045	7.7935	1.0000	0.0416
washboard	9	8	0.0035	8.6008	1.1250	0.0273
trig	12	8	0.0071	8.6008	1.5000	0.0292
gadget	74	61	0.0374	5.6701	1.2131	0.1517
nary	10	9	0.0039	8.4309	1.1111	0.0301
carob	10	5	0.0129	9.2789	2.0000	0.0184
restive	59	54	0.0219	5.8460	1.0926	0.1342
minuet	55	18	0.2730	7.4309	3.0556	0.0602
unwed	14	12	0.0058	8.0159	1.1667	0.0390
ambidextrous	5	5	0.0016	9.2789	1.0000	0.0173
superlunary	2	1	0.0013	11.6008	2.0000	0.0042
bombus	2	2	0.0006	10.6008	1.0000	0.0078
censorial	2	2	0.0006	10.6008	1.0000	0.0078
kaddish	2	2	0.0006	10.6008	1.0000	0.0078
cremate	2	2	0.0006	10.6008	1.0000	0.0078
punster	2	2	0.0006	10.6008	1.0000	0.0078
narcosis	4	4	0.0013	9.6008	1.0000	0.0142
letch	2	2	0.0006	10.6008	1.0000	0.0078
catechize	2	2	0.0006	10.6008	1.0000	0.0078

Table B.12: Frequencies and variability measures for matching words for selected seven-word sequences

Appendix C

Stimuli sentences from experiment three

C.1 Group one

1. When the firemen went
   on strike in 1977
   a state of emergency
   was called by
   the Callaghan government
   and the army was employed.

2. The official exchange rate
   at the time of writing
   is around 50 Kcs to the pound.

3. The hotel is situated
   near Central Station
   in the heart of the city
   by the famous Dam Square.

4. The Attlee administration
   and Ernest Bevin in particular
   believed it was
   only a matter of time
   before the British economy would recover.

5. Weeks went by and
months went by and in
the course of time
they had a little son.

6. The loss or senseless alteration of
the many ordinary pubs that serve these functions
can deal a serious blow to
the quality of life
of their users and the
being of the wider community.

7. And you've got to
be able to enjoy something with
a bit of speed
excitement and controlled danger.

8. And indeed this
must be done not
as a matter of sympathy
with Ireland but as a demand
made in the interest of the English proletariat.

9. There is a decline in attendance
in the early hours of the afternoon
in most subjects.

10. It was usually I who
pinned down occasions with
that sort of fact
so I was surprised that
he should sound so definite.

11. We reckon that both of them are in
the right place at the right price
and will no doubt become
our favourite places to stay in Bodrum.

12. Remember that in Great Britain
there is no such force
and the gap has been
filled in recent years by
a militarization of the police force.

C.2 Group two

1. The females spend most of their life in
a state of pregnancy
as the livebearers breed continuously.

2. The Government will support them
if they decide
at the time of review
not to grant fresh permissions
on environmental and social grounds.

3. The little clearing
in the heart of the woods
looked different at night.

4. My courtship of your daughter is then
only a matter of form
he said.

5. From being expensive to produce
they suffer wear in
the course of use
and as a result are excluded.

6. There were certain problems common to
all nineteen denominational colleges such as
the quality of men
who entered the ministry
and from whom students were recruited.

7. There's fresh water and I
   found some fishing tackle so with
   a bit of luck
   we could be eating in style tonight.

8. Yes I do know that it's
   bad for my health and
   as a matter of fact
   that's why I like it.

9. Day stated that
   in the early hours of the morning
   he had been awakened by
   Henry and Francis Tidbury.

10. They had a fund for dealing with
    that sort of thing
    and that's why you saw
    many of those cases there.

11. They did so by being in
    the right place at the right time
    for the first and last time
    in their history.

12. You must always remember that
    there is no such thing
    as bad food
    only bad diets.

Appendix D

Stimuli sentences from experiment four

D.1 Group one

1. (a) Prices start at 50 pounds
   per person per night
   making our holidays fantastic value for money.

   (b) The compounds were passed on
   from prey to predator
   by their accumulation in fatty tissue
   involving a metabolic process
   which led to higher concentrations
   as the insecticide was passed along the chain.

2. (a) I am building my pond and I
   would be grateful if
   you could tell me if
   it is possible to keep Koi
   without pumps and filters.

   (b) Unfortunately even this threat
   may be ineffective if
   auditors are as constrained
   by their competitive environment
   as some believe them to be.

3. (a) We come from different backgrounds
from the point of view
of our upbringing
and our education
and approach our racing in different ways.

(b) There is also growing concern
about the role of taste
in aesthetic choice
and about the fate
of analytical criticism.

4. (a) They believed that
in the fullness of time
as scientific thinking comes
to supersede religious thinking
the whole category of the supernatural
will come to be recognized as illusory.

(b) If we should look
at the nature of work
as developed through the application
of scientific management principles
we find that workers are only engaged
in part tasks rather than whole ones.

5. (a) Both laws under discussion
are derived from observations and
there can be no doubt
about their empirical credentials.

(b) They might not gain
total victory but at least
there would be a result
and we could all
get on with the game.

6. (a) I can not deny
that in this context

a settled public demand ought to
be taken into account
or that at a certain point
it would have to prevail.

(b) Described as an innovation were
little packs or growbags
of planting soil that could
be used under gravel
to improve your results.

D.2 Group two

1. (a) A few points need to
be borne in mind
when trying out this sequence.

(b) It is expected
that the aircraft will
be moved by road
to the Weeks Museum in Florida
where it will be put
into stock flying condition.

2. (a) He thinks it is
a waste of time
for foreigners to try to
help change things in Central America.

(b) If innkeepers did not have
a right of sale
they would be left
with the property of guests
which they could not realise
in order to satisfy the debt.

3. (a) Nippon Telegraph and Telephone Corp has reported
net profit for the year

to March 31 down at
the equivalent of 516m
on turnover that rose at 823m.

(b) He found Woodruffe with an
armed man in a raincoat
and a bowler hat.

4. (a) Farr's garden was
on the right hand side
at the end of the footpath
and must have been
almost an acre in extent.

(b) It was apparently a group of
laborers seeking stone
for a new engine house
whose opening of a barrow
was witnessed in his youth
by Mr. J. Harris of Liskeard.

5. (a) She was born before
the turn of the century
so it is likely that
her parents had been
born into slavery.

(b) One of the pellet samples
included spines and
the skull of a hedgehog
showing that this predator
can manage to kill and eat
even well protected prey animals.

6. (a) She stood for a moment
on the other side of
the room sizing me up
and me sizing her up

and then she came over
to speak to me.

(b) A doomsday watch has been kept
on the great slab of
rock nearly the size
of a football pitch
ever since workers discovered a fault
in the face of the mountainside.

Appendix E

Reading time data from experiment three

The first table in this appendix reports the mean reading time over all subjects for each phrase. "Group One" and "group two" refer to the subject groups. The columns labelled "Raw" contain the mean unadjusted reading time in milliseconds, those labelled "Length" report the mean reading time in milliseconds adjusted for length, and the columns labelled "All" report the mean Reading time in milliseconds adjusted for length, position in sentence, word frequency, and transitional probabilities.The item numbers are the identifying number for each stimuli items given in appendix C. Those prefixed with an F are in the frequent condition and those marked with an I are in the infrequent condition.

The second table reports the mean reading time over all items for each subject. The numbers each refer to a unique subject ID with subjects 1-15 being in group 1, and 16-30 being in group 2. The columns are labelled "Raw", "Length" or "All" as in the first table.

Group One				Group Two			
	Raw	Length	All		Raw	Length	All
F1	1089.43	-168.17	-142.63	I1	1175.62	-357.37	-313.32
F2	1179.54	-177.68	-119.36	I2	1708.90	132.81	276.99
F3	1055.77	-401.06	-203.74	I3	1280.03	-468.48	-467.90
F4	928.25	-379.15	-217.56	I4	1332.08	-244.01	-200.92
F5	979.83	-178.15	-47.47	I5	1452.82	49.15	158.24
F6	963.74	-249.18	-251.43	I6	1264.89	-193.11	24.43
I7	1161.74	202.99	241.83	F7	1010.29	-220.97	-181.46
I8	1201.79	-205.24	-167.83	F8	1091.11	-398.77	-224.97
I9	1559.68	-445.04	-274.15	F9	1611.28	-482.06	-442.73
I10	1168.13	59.95	95.39	F10	1273.44	-173.34	-181.45
I11	1778.95	-175.96	-17.76	F11	1535.62	-557.73	-431.87
I12	1229.36	-127.86	-149.17	F12	1195.41	-423.79	-515.42

Table E.1: Reading times averaged over subjects for 24 items

	Raw		Length		All	
	Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
1	1234.01	2384.57	-357.97	395.88	-476.55	405.94
2	1394.80	1900.13	-458.74	-134.56	-339.09	-74.89
3	872.28	1080.64	-277.94	-161.53	-127.73	-60.91
4	1256.34	1561.26	-338.37	-194.36	-287.32	-158.30
5	912.60	1271.22	-273.92	-49.83	-83.51	90.26
6	1062.45	1441.85	-408.64	-194.86	-317.60	-74.84
7	849.51	1027.42	-173.13	-86.10	-119.89	-47.66
8	928.26	1024.21	-87.26	-35.08	23.36	33.29
9	945.45	972.76	-125.93	-186.33	70.01	-105.79
10	867.46	1301.71	-253.89	-5.73	-122.26	65.34
11	2028.30	1984.65	-380.47	-823.87	-192.77	-751.13
12	844.48	975.37	-189.03	-133.18	-79.11	-69.21
13	912.92	1563.09	-172.23	197.26	-217.08	249.67
14	633.05	950.95	-279.14	-117.02	-140.60	-60.56
15	749.53	809.31	-106.82	-198.57	-45.35	-120.46
16	1918.42	2078.02	-417.93	-156.02	-258.50	70.65
17	1121.62	1239.58	-533.14	-279.89	-509.46	-328.00
18	708.43	711.73	-119.01	-50.91	-107.38	3.16
19	1288.96	1405.27	-788.74	-514.04	-965.74	-660.47
20	989.44	1034.21	-125.71	-44.47	-97.22	-7.43
21	1566.21	1644.01	-300.91	-112.01	-225.84	23.86
22	1374.51	1220.03	-245.95	-315.13	-201.70	-180.80
23	1044.49	1036.48	-389.33	-283.47	-287.23	-43.91
24	1719.40	1425.62	-411.56	-588.34	-370.78	-562.93
25	1208.15	1115.78	-61.83	-52.86	-62.01	-74.92
26	1755.84	2126.23	-741.31	-173.44	-637.64	155.92
27	1479.67	1641.83	-372.77	-28.09	-271.46	93.99
28	615.26	1069.30	-503.20	33.62	-305.44	138.06
29	1316.43	1437.85	-287.94	-52.09	-274.34	-9.36
30	1186.07	1349.94	-342.31	-85.30	-370.02	75.94

Table E.2: Reading times averaged over items for 30 subjects

Appendix F

Reading time data from experiment four

The first table in this appendix reports the mean reading time over all subjects for each phrase. "Group One" and "group two" refer to the subject groups. The columns labelled "Raw" contain the mean unadjusted reading time in milliseconds, those labelled "Length" report the mean reading time in milliseconds adjusted for length, and the columns labelled "All" report the mean Reading time in milliseconds adjusted for length, position in sentence, word frequency, and transitional probabilities.The item numbers are the identifying number for each stimuli items given in appendix D.

The second table reports the mean reading time over all items for each subject. The numbers each refer to a unique subject ID with subjects 1-15 being in group 1, and 16-30 being in group 2. The columns are labelled "Raw", "Length" or "All" as in the first table.

Group one							
Frequent				Infrequent			
	Raw	Length	All		Raw	Length	All
1(a)	1179.60	-91.81	-158.13	1(a)	1354.73	55.42	-97.59
2(a)	1349.92	78.51	102.72	2(b)	1365.52	62.95	-16.53
3(a)	1061.62	-272.11	-244.01	3(b)	1387.10	22.21	15.72
4(a)	1123.81	-241.08	-273.84	4(b)	993.98	-308.58	-272.24
5(a)	1381.13	78.56	121.01	5(b)	1161.52	-203.37	-112.75
6(a)	1117.02	-185.55	-45.86	6(b)	1271.31	-.09	73.23
Group two							
Frequent				Infrequent			
	Raw	Length	All		Raw	Length	All
1(a)	1555.96	126.87	9.93	1(b)	1692.66	263.57	194.96
2(a)	1233.84	-147.10	-155.26	2(b)	1617.55	236.61	209.85
3(a)	1682.61	-83.51	-68.31	3(b)	1927.38	161.26	-53.86
4(a)	1778.80	60.83	41.75	4(b)	1792.07	74.11	112.20
5(a)	1239.07	-527.05	-327.73	5(b)	2196.06	429.95	377.33
6(a)	1052.32	-569.36	-396.78	6(b)	1610.90	-10.78	24.87

Table F.1: Reading times averaged over subjects for 24 items

	Raw		Length		All	
	Frequent	Infrequent	Frequent	Infrequent	Frequent	Infrequent
1	2468.75	2114.48	16.63	-5.47	33.01	-2.05
2	1554.06	2619.89	-379.08	-90.53	-318.13	-91.80
3	1537.49	2283.32	-595.03	123.35	-562.93	61.67
4	851.70	1010.82	-100.77	-117.51	-61.50	-145.15
5	1019.18	1101.14	70.43	-135.35	143.08	-125.76
6	1621.84	3155.59	-120.52	-134.03	-119.09	-116.69
7	831.15	1043.86	-144.80	-230.16	-158.18	-225.10
8	1102.93	1183.43	89.20	-47.83	93.79	-50.45
9	1013.38	1251.66	-31.78	-33.40	37.71	-32.35
10	1733.65	2065.33	-88.36	40.35	-61.92	44.65
11	1620.34	2555.97	-200.70	-4.62	-126.61	-11.85
12	1640.88	2119.97	65.60	9.90	-56.64	-3.70
13	1495.97	1170.2‘9	-162.17	-113.27	-92.29	-155.21
14	1524.06	1726.26	-9.34	-90.10	-47.49	-63.29
15	1341.10	1689.54	7.00	-100.00	51.91	-108.30
16	1894.12	1883.51	431.44	77.17	342.57	25.43
17	1385.52	1689.85	-248.78	817.04	-173.51	750.12
18	1160.55	1894.53	-468.81	277.03	-431.88	181.79
19	1052.79	1043.65	-109.47	49.65	-45.94	65.37
20	1606.70	1422.15	-13.81	68.15	-4.83	17.79
21	669.98	662.07	-1097.91	435.84	-1044.11	168.95
22	1037.22	954.42	-207.64	5.07	-95.67	36.49
23	1388.16	1262.47	-156.02	-75.52	-61.79	-60.47
24	946.30	952.66	-131.13	107.15	-101.13	46.70
25	867.26	1003.30	-116.41	215.27	7.97	173.58
26	820.03	1021.29	-627.78	307.85	-561.27	244.68
27	2182.97	2141.41	-67.67	411.42	-28.51	371.34
28	800.12	857.60	274.47	-51.20	69.51	-197.41
29	1368.22	1294.12	-150.05	52.15	-79.47	62.89
30	852.80	752.42	-158.74	189.71	-32.91	276.11

Table F.2: Reading times averaged over items for 30 subjects

Appendix G

Entailment judgements for verb-particle constructions from experiment seven

The tables in this appendix contain the entailment judgements for verb-particles constructions as reported in experiment seven. The columns contain the number of subjects that made each of the judgements contained in the header. The headers correspond to judgements as follows:

Y = Yes
N = No
U= Don't know

	$\models$ Verb			$\models$ Particle				$\models$ Verb			$\models$ Particle		
ITEM	Y	N	U	Y	N	U	ITEM	Y	N	U	Y	N	U
blow out	22	5	0	18	9	0	bring up	10	17	0	8	19	0
close off	25	2	0	11	16	0	come about	14	12	1	8	17	2
come back	25	2	0	18	9	0	draw out	10	16	1	15	10	2
drive out	17	10	0	22	4	1	drop off	19	8	0	14	13	0
drown out	13	14	0	11	16	0	fend off	17	10	0	17	8	2
fight back	26	1	0	10	17	0	figure out	19	8	0	1	26	0
fill up	26	1	0	3	24	0	fly off	22	5	0	13	12	2
follow up	18	9	0	2	25	0	go along	25	2	0	14	12	1
help out	27	0	0	1	25	1	knock down	13	14	0	22	5	0
live in	25	2	0	22	5	0	look up	23	4	0	21	5	1
open up	20	7	0	1	25	1	pack up	26	1	0	1	25	1
pay back	26	1	0	16	11	0	play out	14	11	2	1	24	2
print out	27	0	0	13	13	1	pull back	25	2	0	20	6	1
roll up	25	2	0	2	25	0	sell out	24	3	0	10	17	0
shake up	15	12	0	4	22	1	shrug off	13	13	1	16	9	2
shut down	21	6	0	7	19	1	sit up	22	5	0	19	6	2
smash up	26	1	0	2	24	1	split up	24	3	0	1	25	1
stir up	13	14	0	4	21	2	take back	22	5	0	19	7	1
take up	20	7	0	5	21	1	tie up	20	7	0	1	26	0
whip up	12	15	0	6	20	1	work up	16	11	0	7	18	2

Table G.1: Entailment judgements for subject group one

	$\models$ Verb			$\models$ Particle				$\models$ Verb			$\models$ Particle		
ITEM	Y	N	U	Y	N	U	ITEM	Y	N	U	Y	N	U
bail out	19	9	1	13	14	2	break off	18	9	2	15	10	4
bring down	17	10	2	16	9	4	bring in	26	3	0	22	7	0
burn down	25	3	1	15	12	2	burn up	20	8	1	3	23	3
buy back	25	4	0	14	13	2	clear out	21	8	0	20	8	1
curl up	24	4	1	7	21	1	cut down	16	13	0	14	15	0
drag out	23	5	1	19	8	2	draw back	10	19	0	20	8	1
draw up	15	14	0	8	20	1	drop out	8	20	1	16	12	1
dust off	14	13	2	13	13	3	eke out	14	8	7	5	15	9
fight off	18	11	0	13	15	1	find out	26	2	1	3	25	1
give up	13	15	1	4	23	2	go in	24	3	2	21	5	3
head off	4	22	3	6	19	4	lay down	11	15	3	12	14	3
let off	13	14	2	12	14	3	let out	18	9	2	24	1	4
look down	24	4	1	20	8	1	pull down	19	9	1	22	6	1
pull out	24	5	0	21	7	1	read out	26	2	1	8	19	2
send out	24	3	2	17	10	2	set out	12	15	2	10	17	2
slide down	22	7	0	19	10	0	spring up	10	16	3	14	11	4
stand up	27	1	1	20	6	3	stretch out	24	3	2	13	14	2
take in	13	13	3	17	8	4	take over	23	6	0	5	24	0
throw down	26	2	1	22	4	3	tick off	7	19	3	4	22	3
trickle down	24	3	2	24	2	3	wear down	8	18	3	13	13	3

Table G.2: Entailment judgements for subject group two

	$\models$ Verb			$\models$ Particle				$\models$ Verb			$\models$ Particle		
ITEM	Y	N	U	Y	N	U	ITEM	Y	N	U	Y	N	U
act out	27	2	0	5	24	0	back off	15	14	0	14	13	2
back up	17	10	2	4	23	2	beat up	26	3	0	2	27	0
break down	23	6	0	7	19	3	build up	26	3	0	9	19	1
close up	24	5	0	2	26	1	count out	27	2	0	8	20	1
do up	18	9	2	4	23	2	eat up	21	7	1	4	24	1
finish off	26	3	0	2	27	0	flush out	20	8	1	17	7	5
get in	5	24	0	27	1	1	go out	24	3	2	21	5	3
hold back	20	7	2	11	15	3	hold out	14	15	0	17	10	2
jump off	28	1	0	21	8	0	jump up	25	4	0	25	3	1
knock out	12	14	3	14	12	3	mix up	25	4	0	3	25	1
mop up	18	10	1	6	21	2	patch up	24	5	0	1	28	0
ring off	10	16	3	10	13	6	rise up	23	6	0	19	8	2
run down	17	11	1	11	16	2	shake out	26	2	1	15	13	1
shout out	29	0	0	16	13	0	shut off	15	13	1	22	5	2
sit down	28	1	0	21	8	0	slip out	20	9	0	20	5	4
speed up	24	5	0	7	22	0	stamp out	13	15	1	9	18	2
start off	28	1	0	1	28	0	stay on	26	3	0	9	20	0
strike up	6	22	1	3	24	2	turn on	7	20	2	18	8	3
turn over	12	16	1	13	15	1	walk down	28	1	0	10	16	3
wear off	16	12	1	14	13	2	wind up	12	16	1	6	22	1

Table G.3: Entailment judgements for subject group three

	$\models$ Verb			$\models$ Particle				$\models$ Verb			$\models$ Particle		
ITEM	Y	N	U	Y	N	U	ITEM	Y	N	U	Y	N	U
add up	24	0	0	6	17	1	bite off	17	7	0	14	7	3
brighten up	20	4	0	2	21	1	carry away	5	17	2	8	14	2
carry out	1	23	0	1	21	2	catch up	12	12	0	1	22	1
dig up	21	3	0	17	7	0	fall off	22	2	0	20	4	0
get back	21	3	0	15	9	0	get down	10	13	1	19	2	3
give off	17	7	0	14	9	1	go down	19	3	2	18	4	2
hand out	18	6	0	12	10	2	hang out	23	1	0	19	5	0
lay out	15	9	0	5	17	2	lie down	24	0	0	21	3	0
lift out	24	0	0	23	1	0	map out	17	7	0	2	22	0
mark out	20	4	0	5	19	0	move off	23	1	0	7	17	0
move out	24	0	0	19	5	0	pay off	23	0	1	3	17	4
play down	5	19	0	11	12	1	pull off	13	8	3	11	7	6
put off	7	17	0	8	14	2	roll back	9	15	0	17	3	4
run up	12	11	1	11	9	4	seek out	23	1	0	3	21	0
sell off	24	0	0	4	19	1	shake off	11	13	0	16	7	1
slow down	22	2	0	5	18	1	sort out	14	10	0	2	21	1
stay up	21	3	0	18	6	0	step off	23	1	0	22	2	0
stick out	12	12	0	21	3	0	throw back	16	8	0	20	4	0
throw in	13	10	1	13	9	2	throw out	13	11	0	20	3	1
trail off	5	19	0	13	11	0	wear on	4	19	1	6	17	1

Table G.4: Entailment judgements for subject group four

Bibliography

ANTILLA, RAIMO. 1989. *Historical and Comparative Linguistics*. John Benjamins.

BAAYEN, R.H., R. PIEPENBROCK, and H VAN RIJN. 1993. *The CELEX lexical database (CDROM)*. Linguistic Data Consortium, University of Pennsylvania.

BALDWIN, TIMOTHY. 2005. The deep lexical acquisition of english verb-particle constructions. *Computer Speech and Language* 19.398–414.

——, and ALINE VILLAVICENCIO. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*.

BANNARD, COLIN. 2005. Learning about the meaning of verb particle constructions from corpora. *Computer Speech and Language* 19.467–478.

——, TIMOTHY BALDWIN, and ALEX LASCARIDES. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

BARKEMA, HENK. 1994. Determining the syntactic flexibility of idioms. In *Creating and Using English Language Corpora*, 39–52. Rodopi.

BAUER, LAURIE. 1983. *English Word-formation*. Cambridge, UK: Cambridge University Press.

BÉJOINT, HENRI. 1989. "codedness" and lexicography. In *Lexicographers and their works (Exeter Linguistic Studies, Volume 14)*, ed. by Gregory James. University of Exeter.

BENTON, A.L., and R.J.JOYNT. 1960. Early descriptions of aphasia. *Archives of Neurology* 3.205–222.

BERRY-ROGGHE, G.L.M. 1974a. Automatic identification of phrasal verbs. In *Computers in the Humanities*, ed. by J.L. Mitchell, 16–26. Edinburgh University Press.

—— 1974b. The computation of collocations and their relevance in literary studies. In *The Computer and Literary Studies*, ed. by A.J. Aitken, R.W. Bailey, and N. Hamilton-Smith. Edinburgh University Press.

BIBER, DOUGLAS. 1988. *Variation across speech and writing*. Cambridge University Press.

BLOOMFIELD, LEONARD. 1933. *Language*. New York: Holt, Rinehart and Winston.

BOBROW, S., and S. BELL. 1973. On catching on to idiomatic expressions. *Memory and Cognition* 1.343–346.

BOD, RENS, 2000. The storage vs. computation of three-word sentences. Talk presented at AMLaP-2000.

——. 2001. Sentence memory: Storage vs. computation of frequent sentences. In *Proceedings of the 14th Annual CUNY Sentence Processing Conference*.

——, JENNIFER HAY, and STEFANIE JANNEDY (eds.) 2003. *Probabilistic Linguistics*. MIT Press.

BOGURAEV, B.K., and E.J. BRISCOE. 1989. Introduction. In *Computational Lexicography for Natural Language Processing*, ed. by B.K. Boguraev and E.J. Briscoe. London,UK: Longman.

BOLINGER, DWIGHT. 1976. Meaning and memory. *Forum Linguisticum* 1.

BOND, FRANCIS. 2005. *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*. CSLI Publications.

BOWER, GORDON H. 1969. Chunks as interference units in free recall. *Journal of Verbal Learning and Verbal Behavior* 8.610–613.

BOWKER, LYNNE, and JENNIFER PEARSON. 2002. *Working with Specialized Language – A practical guide to using corpora*. Routledge.

BRISCOE, TED, and JOHN CARROLL. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 1499–1504.

BURNARD, LOU. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

BYBEE, JOAN. 2003. Mechanisms of change in grammaticization: The role of frequency. In *The Handbook of Historical Linguistics*, ed. by B. D. Joseph and J. Janda, 602–623. Blackwell.

——, 2005. From usage to grammar: The mind's response to repetition. LSA Presidential Address, 2005.

——, and JOANNE SCHEIBMAN. 1999. The effect of usage on degrees of constituency: the reduction of don't in english. *Linguistics* 37.575–596.

CALLISON-BURCH, CHRIS, COLIN BANNARD, and JOSHUA SCHROEDER. 2005. Scaling phrase-based statistical machine translation to larger corpora and larger phrases. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

CARROLL, JOHN, TED BRISCOE, and ANTONIO SANFILIPPO. 1998. Parser evaluation: a survey and new proposal. In *Proceedings of the First International conference on Language Resources and Evaluation*, 447–454.

CHINCHOR, N., L HIRSCHMAN, and D LEWIS. 1993. Evaluating message understanding systems: an analysis of the third message understanding conference (muc-3). *Computational Linguistics* 19.

CHURCH, K., W. GALE, P. HANKS, and D. HINDLE. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, 115–164. Lawrence Erlbaum.

CHURCH, KEN, and WILLIAM GALE. 1994. Poisson mixtures. *Journal of Natural Language Engineering* 1.163–190.

——, and PATRICK HANKS. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 16.22–29.

CLARK, RUTH. 1970. Performing without competence. *Journal of Child Language* 1.1–10.

CODE, CHRIS. 1987. *Language, Aphasia, and the Right Hemisphere*. Chichester: Wiley.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20.37–46.

COHEN, P. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

COLLINS (ed.) 2000. *Collins Cobuild Dictionary of Idioms*. Harper Collins.

CORLEY, MARTIN, FRANK KELLER, and CHRISTOPH SCHEEPERS, 2000. *Conducting psychological experiments over the world wide web*. Unpublished manuscript, University of Edinburgh and Saarland University.

COVER, T., and J. THOMAS. 1991. *Elements of Information Theory*. Wiley and Sons.

CURRAN, JAMES, 2003. *From Distributional to Semantic Similarity*. University of Edinburgh dissertation.

DAILLE, B. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, ed. by P. Resnik and J. Klavans, 49–66. MIT Press.

DEERWESTER, SCOTT, SUSAN DUMAIS, GEORGE FURNAS, THOMAS LANDAUER, and RICHARD HARSHMAN. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41.391–407.

DIAS, GAEL. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

DUNNING, TED. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.61–74.

EVERAERT, M., E-J. VAN DER LINDEN, A. SCHENK, and R. SCHREUDER (eds.) 1995. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates.

EVERT, STEFAN. 2004. A simple lnre model for random character sequences. In *Proceedings of the 7émes Journées Internationales d'Analyse Statistique de Données Textuelles*.

——, 2005. Ucs/perl documentation. http://www.collocations.de/UCS/UCS-Perl-html/pod/ucsam.html.

——, and BRIGITTE KRENN. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188–195.

FERREIRA, FERNANDA, and CHARLES CLIFTON. 1986. The independence of syntactic processing. *Journal of Memory and Language* 25.348–368.

FILLMORE, C., P. KAY, and M. O'CONNOR. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64.501–538.

FIRTH, J.R. 1957. Modes of meaning. In *Papers in Linguistics 1934-1951*. Oxford University Press.

FORSTER, KENNETH I., and SUSAN M. CHAMBERS. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12.627–635.

FORSTER, K.I., and J.C. FORSTER. 2003. Dmdx: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments and Computers* 35.116–124.

FRANCIS, W, and H KUCERA. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin.

FREGE, G. 1961. *Die Grundlagen der Arithmetik. Eine logisch-mathematische Untersuchung über den Begriff der Zahl.* Georg Olms.

FREGE, GOTTLOB. 1977. *Compound Thoughts, Logical Investigations (Translated by P.T. Geach and R.H. Stoothoff)*. Blackwell.

GIBBS, RAYMOND. 1994. *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge University Press.

GIBBS, R.W., N.P. NAYAK, and COOPER CUTTING. 1989. How to kick the bucket and not decompose: Analysability and idiom processing. *Journal of Memory and Language* 28.576–593.

GOLDBERG, ADELE. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago, USA: University of Chicago Press.

GREFENSTETTE, GREGORY. 1994. *Explorations in Automatic Thesaurus Extractions*. Kluwer Academic.

HA, L.Q., J. MING E.I. SICILIA-GARCIA, and F.J. SMITH. 2002. Extension of zipf's law to words and phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*.

HAIMAN, JOHN. 1994. Ritualization and the development of language. In *Perspectives on Grammaticalization*, 3–28. John Benjamins.

HARRIS, ZELLIG. 1964. Distributional structure. In *The Structure of Language: Readings in the Philosophy of Language*, ed. by Jerry Fodor and Jerrold Katz, 33–48. Prentice-Hall.

HART, BETTY, and TODD R RISLEY. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Brookes.

HAWKINS, JOHN A. 2000. The relative order of preposition phrases in English: Going beyond manner – place – time. *Language Variation and Change* 11.231–266.

HINTON, PERRY R. 1995. *Statistics Explained: A Guide for Social Science Students*. Routledge.

HOWES, DAVID. 1957. On the relation between the intelligibility and frequency of occurrence of english words. *Journal of the Acoustical Society of America* 29.296–305.

——, and RICHARD L. SOLOMON. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology* 41.491–410.

JACKENDOFF, RAY. 1995. The boundaries of the lexicon. In (Everaert *et al.* 1995), chapter 7.

JACKSON, JOHN HEWLINGS. 1879a. On affections of speech from disease of the brain. In (Taylor 1879).

——. 1879b. On the nature of the duality of the brain. In (Taylor 1879).

JANSSEN, THEO M.V. 1997. Compositionality. In *Handbook of Logic and Language*. Elsevier.

JORGENSEN, JULIA. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research* 19.167–190.

JURAFSKY, DANIEL. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In (Bod *et al.* 2003), 37–96.

JUSTESON, JOHN, and SLAVA KATZ. 1995. Technical terminology: some linguistic properties and an algorithm for detection in text. *Natural Language Engineering* 1.9–27.

KATZ, SLAVA M. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering* 2.15–59.

KELLER, FRANK. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 317–324, Barcelona.

KILGARRIFF, ADAM. 2002. English lexical sample task description. In *Proceedings of the Senseval-2 workshop*.

KRIPPENDORFF, KLAUS. 1980. *Content Analysis: An Introduction to its Methodology*. Sage.

KRUG, MANFRED. 1998. String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics* 26.286–320.

KUIPER, KOENRAAD, HEATHER MCCANN, and HEIDI QUINN, 2003. A syntactically annotated idiom database (said), v,1.

LANCKER-SIDTIS, DIANA VAN, and GAIL RALLON. 2004. Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language and Communication* 24.

LANDIS, J.R., and G.G. KOCK. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33.159–174.

LAPATA, MARIA, and ALEX LASCARIDES. 2003a. A probabilistic account of logical metonymy. *Computational Linguistics* 29.263–317.

——, SCOTT MCDONALD, and FRANK KELLER. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 30–36, Bergen.

LAPATA, MIRELLA, and ALEX LASCARIDES. 2003b. Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 235–242, Budapest.

LEECH, G., P. RAYSON, and A. WILSON. 2001. *Word Frequencies in Written and Spoken English.* Longman.

LIEVEN, E, H. BEHRENS, J. SPEARES, and M. TOMASELLO. 2003. Early syntactic creativity: a usage-based approach. *Journal of Child Language* 30.333–370.

LIN, DEKANG. 1994. Principar—an efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94).*

——. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98).*

——. 1998b. Extracting collocations from text corpora. In *Proceedings of the First Workshop on Computational Terminology.*

——. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the ACL*, 317–24, College Park, USA.

LONG, THOMAS H., and DELLA SUMMERS. 1979. *Longman Dictionary of English Idioms.* Longman Dictionaries.

LORCH, ROBERT F., and JEROME L. MYERS. 1990. Regression analysis of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory and Cognition* 16.149–157.

LUM, CARMEL C., and ANDREW W. ELLIS. 1994. Is "nonpropositional" speech preserved in aphasia? *Brain and Language* 46.368–391.

LUND, KEVIN, CURT BURGESS, and RUTH ANN ATCHLEY. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, 660–665, Pittsburgh, USA.

MACDONALD, M.C. 1993. The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language* 32.692–715.

MANDELBROT, BENOIT B. 1953. An information theory of the statistical structure of language. In *Proceedings of the Symposium on Applications of Communications Theory*.

MANNING, C., and H. SCHUTZE. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, USA: MIT Press.

MCDONALD, S.A., and R.C SHILLCOCK. 2003. Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research* 43.1735–1751.

MCDONALD, SCOTT, 2000. *Environmental Determinants of Lexical Processing Effort*. University of Edinburgh dissertation.

MCENERY, TONY, and ANDREW WILSON. 1996. *Corpus Linguistics*. Edinburgh University Press.

MELAMED, I. DAN. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*.

MEL'ČUK, I. 1995. Phrasemes in language and phraseology in linguistics. In (Everaert *et al.* 1995), chapter 8.

MILLER, GEORGE A., RICHARD BECKWITH, CHRISTIANE FELLBAUM, DEREK GROSS, and KATHERINE J. MILLER. 1990. Introduction to WordNet: an online lexical database. *International Journal of Lexicography* 3.235–44.

MINNEN, GUIDO, JOHN CARROLL, and DARREN PEARCE. 2000. Robust, applied morphological generation. In *Proceedings of the First International Natural Language Generation Conference*, 201–208, Mitzpe Ramon, Israel.

MITCHELL, T.F. 1971. Linguistic 'goings on': Collocations and other lexical matters arising on the syntactic record. *Archivum Linguisticum* 2.35–69.

MONTAGUE, RICHARD. 1970. Universal grammar. *Theoria* 36.373–398.

——. 1973. The proper treatment of quantification in ordinary english. In *Approaches to Natural Language*, ed. by K. Hintikka, J. Moravcsik, and P. Suppes, 221–242. Reidel.

MOON, ROSAMUND. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford, UK: Oxford University Press.

MOORE, ROGER K. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech 2003*, 2582–2584.

NELDER, J.A., and R. MEAD. 1965. A simplex method for function minimization. *Computer Journal* 7.308–313.

NG, HWEE TOU, CHUNG YONG LIM, and SHOU KING FOO. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL-SIGLEX Wordshop on standardising lexical resources*.

NUNBERG, GEOFFREY, IVAN A. SAG, and TOM WASOW. 1994. Idioms. *Language* 70.491–538.

O'DOWD, ELIZABETH M. 1998. *Prepositions and Particles in English.* Oxford University Press.

PALMER, F.R. 1965. *A Linguistic Study of the English Verb.* Longmans.

PATEL, MALTI, JOHN BULLINARIA, and JOSEPH LEVY. 1997. Extracting semantic representations from large text corpora. In *Proceedings of the Fourth Neural Computation and Psychology Workshop*.

PAWLEY, ANDREW. 1986. Lexicalization. In *Languages and Linguistics: The interdependence of Theory, Data, and Application (Georgetown University Round Table on Languages and Linguistics, 1985)*, ed. by D. Tannen and J.E. Alatis. Georgetown University Press.

PEARCE, DARREN. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, CMU.

PECINA, PAVEL. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the Student Research Workshop of the 43rd Annual Meeting of the Association for Computational Linguistics*.

PEREIRA, FERNANDO, NAFTALI TISHBY, and LILLIAN LEE. 1993. Distributional clustering of English words. In *31st Annual Meeting of the ACL*, 183–190.

PETERS, ANN. 1977. Language learning strategies: Does the whole equal the sum of its parts? *Language* 53.

PINKER, STEPHEN. 1995. *The Language Instinct*. Penguin Books.

POLLATSEK, ALEXANDER, and ARNOLD D. WELL. 2001. On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful anaysis. *Journal of Experimental Psychology: Learning, Memory and Cognition* 21.785–794.

PULLUM, GEOFFREY K. 2004. Snowclones: lexicographical dating to the second. *Language Log, January 16, 2004* . http://itre.cis.upenn.edu/ myl/languagelog/archives/000350.html.

PUSTEJOVSKY, JAMES. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

RIEHEMANN, SUZANNE, 2001. *A Constructional Approach to Idioms and Word Formation*. Stanford University dissertation.

RUBENSTEIN, HERBERT, LONNIE GARFIELD, and JANE A. MILLIKAN. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* 9.487–494.

SACHS, J.S. 1967. Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics* 2.437–444.

SCHONE, PATRICK, and DAN JURAFSKY. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, 100–108.

SCHÜTZE, HINRICH. 1998. Automatic word sense discrimination. *Computational Linguistics* 24.97–123.

SINCLAIR, JOHN M. 1987. Collocation: A progress report. In *Language Topics: Essays in Honour of Michael Halliday*, ed. by R. Steele and T. Threadgold, volume 2. John Benjamins.

—— 1991. *Corpus, Collocation, Concordance*. Oxford University Press.

——, and OTHERS. 1987. *Collins COBUILD English Language Dictionary*. Collins.

SOANES, CATHERINE, and ANGUS STEVENSON. 2005. *Oxford Dictionary of English*. Oxford University Press.

SPARCKJONES, KAREN. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28.11–21.

STEVENSON, MARK. 2003. *Word Sense Disambiguation: The Case for Combining Knowledge Sources*. CSLI Publications.

STOLCKE, ANDREAS. 2002. Srilm - an extensible language modelling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

SWINNEY, DAVID A., and ANNE CUTLER. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior* 18.523–534.

TAYLOR, J. (ed.) 1879. *Selected Writings of John Hewlings Jackson*. Staples Press.

THEAKSTON, ANNA. 2004. The role of entrenchment in children's and adult's performance on grammaticality judgment tasks. *Cognitive Development* 19.15–24.

TITONE, DEBRA A., and CYNTHIA M CONNINE. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics* 31.1655–1674.

TOMASELLO, MICHAEL. 1992. *First Verbs: A Case Study in Early Grammatical Development*. Cambridge University Press.

——. 2001. *The Cultural Origins of Human Cognition*. Harvard University Press.

——. 2003. *Constructing a Language*. Harvard University Press.

TOUTANOVA, K., and C. MANNING. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

TRUESWELL, JOHN C. 1995. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language* 35.566–585.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval*. Butterworths.

VILLAVICENCIO, ALINE. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? *Computer Speech and Language* 19.415–432.

——, and ANN COPESTAKE, 2002. On the nature of idioms. *LinGO Working Paper No. 2002-04*.

WEINREICH, URIEL. 1969. Problems in the analysis of idioms. In *Substance and Structure of Language*. University of California Press.

WERMTER, JOACHIM, and UDO HAHN. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics*, volume 2, 980– 986.

WIDDOWS, DOMINIC. 2004. *Geometry and Meaning*. CSLI Publications.

WITTGENSTEIN, LUDWIG. 1953. *Philosophical Investigations*. Oxford University Press: Blackwell.

WRAY, ALISON. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

——, and M. PERKINS. 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20.1–28.

YATES, S. 1996. Oral and written linguistic aspects of computer conferencing. In *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, 29–46. John Benjamins.

ZIPF, GEORGE KINGSLEY. 1935. *The Psycho-Biology of Language*. Houghton Miffin.