Durrant, Philip Lee (2008) High frequency collocations and second language learning. PhD thesis, University of Nottingham.

High frequency collocations and second language learning

Philip Durrant, BA, MA

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

October 2008

Abstract

This thesis explores the implications of high frequency collocation for adult second language learners. It addresses three main questions. First, it asks to what extent high frequency of occurrence in a corpus indicates that collocations are independently represented in the minds of native speakers. A word association study indicates that high frequency of occurrence is a fairly reliable predictor of mental representation, though this methodology does not allow us to determine the precise strength of the relationship. A series of lexical decision studies also show a relationship between frequency and representation, but effects are limited to those collocations which are sufficiently salient to also register as associates. This suggests that psycholinguistic 'priming' models may not be the best way of understanding collocation. Second, the thesis examines the idea that adult second language learners usually fail to retain the collocations to which they are exposed. This is tested through a lab-based training study and a learner-corpus study. Results suggest that adult learners are capable of learning collocations from input, but that 1) the relatively low levels of input to which most learners are exposed mean that they nevertheless tend not to attain native-like profiles of collocation use, and 2) input which provides repeated exposure to collocations can dramatically improve learning. Third, the thesis asks whether a useful pedagogical listing of frequent 'academic collocations' can be compiled. Results suggest that an academic collocation list is viable, but that important caveats need to be made concerning the nature of the collocations included and the range of disciplines for which such a listing will be useful. Moreover, listings of two-word collocations should be seen only as a starting point for more comprehensive phraseological listings. Suggestions will be made for ways in which we might go beyond such two-word listings.

Acknowledgements

I owe several debts of gratitude to friends and colleagues for their support in the completion of this thesis. Firstly, thanks are due to my supervisor, Norbert Schmitt, whose rigorous academic coaching, personal support, and inexhaustible enthusiasm have been an invaluable spur and resource. Many of the studies presented here also benefited greatly from discussions with Kathy Conklin, without whose psycholinguistic expertise, this thesis would have been far weaker.

I have been fortunate to be part of a highly supportive group of research students at the University of Nottingham, whose scrutiny has been invaluable at all stages of preparing this thesis. Thanks in particular to Irina Dahlmann, Myq Larson, Phoebe Lin, Li Jie, and Anya Siyanova for their suggestions and criticisms.

I also owe thanks to many other researchers and teachers who have made this work possible, either through material support or through the ideas they contributed. Walter Van Heuven and Kathy Conklin kindly gave me access to their psycholinguistic facilities and provided technical support. Thanks also to Zhang Taoli for helping me get to grips with DMDX and to Anya Siyanova, who helped me to conduct part of the lexical decision research. Much of the research reported here relies on the contributions of my former students at Bilkent and Durham Universities, who kindly allowed me to use samples of their writing, and to Robin Turner who gave me access to his own corpus of learner writing. Thanks also to Jasper Holmes and Hilary Nesi for providing me with access to the BAWE corpus, to Sylviane Granger for her help with the ICLE corpus, an to Jakub Marecek for programming assistance. Many researchers and teachers have lent me their time to discuss the ideas and research that appear in this thesis. I owe thanks in particular to Zoltan Dörnyei, Nick Ellis, Adele Goldberg, Sandra Haywood, Michael Hoey, Steve Kirk, Kon Kuiper, Mike Scott, and Antoine Tremblay.

Thanks are also due to the Economic and Social Research Council, who funded this thesis through a 3 year postgraduate studentship.

Finally, I owe a special personal debt to the friends and family who have helped me throughout my postgraduate career. In addition to the friends listed above, thanks in particular to my parents, Peter and Pauline Durrant, and to the various pals in Nottingham, Durham, Ankara, and Istanbul, without whose support the last three years would have been a much harder and less rewarding experience.

Contents

Chapter 1
Introduction

1.1 Aims of the thesis

The last three decades have seen much interest within the second language teaching[1] community in the phenomenon of formulaic language. Learning formulaic language has been viewed as an essential element in achieving nativelike production (Pawley & Syder, 1983) and as key to the general language acquisition process (Peters, 1983). Formulas are gaining an increasingly prominent position in present-day teaching and reference materials and entire teaching approaches have been based around the learning of formulaic language (Lewis, 1993; Nattinger & DeCarrico, 1992). However, as Granger (1998, pp. 157-158) has pointed out, there is a danger in this enthusiasm of pedagogical practice outstripping linguistic knowledge. Recent years have produced much research in this area, but models remain rudimentary and tentative, and the stronger claims of the advocates of formula-based teaching are still unsubstantiated (see Section 3.2). If language teachers and learners are to engage effectively with formulaic language, many questions still need to be answered.

The present thesis aims to address some of these questions as they apply specifically to the phenomenon of high frequency two-word collocations. Such collocations have, as we shall see in Chapter 2, been central to the study of formulaic language. Indeed, some researchers believe that the principles at work in two-word collocations are archetypes of those involved in formulaic language as a whole, and even in language generally (e.g., Ellis, 2003; Hoey, 2005). Moreover, two-word collocations are relatively simple formulaic items, and a number of well-established methods exist for their quantification. Taken together, these considerations make them an ideal test-case for research into formulaic language.

The thesis will address three main questions. The first is the fundamental one of whether two-word collocations are, as has been hypothesised (e.g., Hoey, 2005;

[1] the phrases 'second language teaching' and 'foreign language teaching' will be used interchangeably throughout this thesis; the words 'acquisition' and 'learning' will also be treated as synonyms.

Sinclair, 1987), independently represented in the mental language systems of native speakers. Some researchers (e.g., Bley-Vroman, 2002; Herbst, 1996) have argued that collocation is an entirely textual phenomenon, which does not indicate anything of importance about how language is represented in the mind. On this view, collocations are not items that second language learners need to acquire; rather they arise spontaneously in text as an epiphenomenon of the meaningful use of language in context. If this position is correct, then a pedagogical focus on collocation would appear to be misguided. If, on the other hand, high frequency collocations correspond somehow to items which native speakers 'know', then they may represent good targets for learning. This thesis will aim to explore whether high frequency of occurrence does indeed indicate that word-pairs are independently represented in the mental language system of native speakers and tests one model of how such collocations might be represented.

The second question concerns how collocations are best learnt. Schmitt (forthcoming) has argued that the highly contextualized nature of collocations implies that they are best learnt implicitly, through extensive exposure to the target language. However, on the basis of her review of the acquisition literature, Wray has suggested that adult second language learners may not usually be able to acquire collocations in this way (Wray, 2002, pp. 206-209). She claims that learners' mature cognitive systems and the nature of their learning situations bias them towards a word-focused approach to learning which prevents them from retaining the collocations they meet. This model has strong implications both for how we understand the second language learning process and for how we approach the teaching of collocations, and therefore deserves a thorough evaluation. The second main aim of this thesis will be to provide such an evaluation.

The third question to be addressed is that of whether a useful learning inventory of target collocations can be specified for a particular group of learners. Listings of the individual words that second language learners are most likely to need have been used for many years (e.g., Thorndike & Lorge, 1944; West, 1953), and if the notion of vocabulary learning is to be extended to incorporate formulaic language, it would seem essential that our word lists also be extended to include such language (Coxhead, 2008). However, it is not yet clear how viable a pedagogical 'collocation list' would

be. The appeal of traditional word lists has lain in their ability to provide learners with a small number of very high frequency items which are thought to account for the majority of language likely to be encountered by the majority of learners. However, collocations are in general much rarer, much more diverse, and much more strongly tied to specific areas of discourse than are individual words. Given this, a collocation list is likely to provide far lower levels of text coverage with far higher numbers of items than traditional word lists, and it is unclear to what degree collocations will be sufficiently generic – i.e. found in a sufficiently wide range of discourse areas – to be useful to a reasonably wide spectrum of learners. The third main aim of this thesis will therefore be to test whether a pedagogically-useful listing of target collocations can be compiled for one particular area of English in which the word list approach has been particularly popular – that of English for academic purposes.

1.2 The shape of the thesis

The thesis has five central chapters. Chapters 2 and 3 provide general theoretical background: Chapter 2 describes the notion of high frequency collocation as it has been studied by corpus linguists, and the key properties that have been attributed to it. Collocation is taken to be one type of *formulaic language*, and the chapter indicates the wider significance of collocation by placing it within the context of this broader category. Chapter 3 looks at the relevance of formulaic language in general, and high frequency collocation in particular, for second language learning. It briefly describes the historical role of formulas in language teaching before evaluating the current-day arguments for focusing on such language and discussing some of the problems posed by formula learning. Chapters 4, 5, and 6 address in turn each of the three research aims described above. They include more detailed literature reviews pertinent to each issue and describe several original studies. Chapter 4 looks at the psychological reality of high frequency collocations. It examines how well various measures of collocational frequency predict psychological associations between words and tests one hypothesis as to how high frequency collocations might be represented in the mind. Chapter 5 tests the thesis that adult second language learners do not retain the collocations to which they are exposed. Chapter 6 explores possibilities for constructing a pedagogical listing of academic collocations as targets for learners of English for academic purposes.

Chapter 2
Collocations and Formulaic Language

2.1 Introduction

This chapter will introduce the linguistic phenomenon which forms the central focus of the thesis: high frequency collocation. It aims to clarify the concept of collocation, to outline some of its key features as they have been portrayed within different descriptive frameworks, and to indicate why collocations are thought to be important for linguistic theory. To achieve this, it will be necessary to discuss both the properties of collocation itself and the place of collocation within the broader context of 'formulaic language'. Section 2.2 will discuss collocations, while 2.3 will look at formulaic language in general. The implications of collocation and formulaic language for second language learners will be addressed in Chapter 3.

2.2. A frequency-based approach to collocation

The concept of collocation

In its non-technical sense, collocation is defined as the 'action of setting in a place or position, *esp.* of placing together with, or side by side with, something else' (Oxford English Dictionary, 2nd ed. 1989). Since (written) language involves a great deal of placing things side by side, it is not surprising to find the term appearing in several linguistic contexts through the centuries:

> **1750** HARRIS *Hermes* II.iv.Wks. (1841)197 The accusative..in modern languages..being subsequent to its verb, in the collocation of the words.
> **1751** JOHNSON *Rambler* No.88 p5 The difference of harmony arising..from the collocation of vowels and consonants.
> **1873** EARLE *Philol.Eng.Tongue.* (ed.2) §630 All languages use greater freedom of collocation in poetry than in prose.
>
> (Oxford English Dictionary, 2nd ed. 1989)

As a technical term in linguistics, however, collocation implies rather more than mere placing side by side. On the OED's formulation, it is the 'habitual juxtaposition or association, in the sentences of a language, of a particular word with other particular words' or 'a group of words so associated' (Oxford English Dictionary, 2nd ed. 1989). This technical sense of collocation differs from the lay sense in three ways. Firstly, collocation does not refer to the juxtaposition of just anything, but specifically of 'particular words'. For this reason, the 'collocations' discussed in the first two quotations above – which concern the juxtaposition of parts of speech and the juxtaposition of types of phonemes respectively – are not collocations in the technical sense. Secondly, the juxtaposition of words only counts as collocation if it is 'habitual'. One-off or rare word-pairings are not collocations. This criterion would appear to rule out the 'collocations' of the third quotation, since this appears to be claiming that unique or unusual collocations exist in poetry. Finally, collocation can refer not only to the act of juxtaposition itself (so that we can talk of, for example, *powerful argument* being a product of collocation), but also to the groups of words involved in such arrangements (such that the words *powerful* and *argument* can, together, be called 'a collocation').

The OED attributes this technical sense of collocation to J.R.Firth. Firth argued that the "habitual collocations" in which a word appears are part of that word's meaning, summarising his position in the now-famous dictum: "You shall know a word by the company it keeps' (1968, p. 179). To understand what Firth has in mind here, it is important to note that the 'meaning' of a word is not intended to be thought of as the 'concept' or 'idea' with which it is associated (1957, p. 196); rather Firth intends the broader sense of 'meaning' characterised by Wittgenstein's statement that "the meaning of words lies in their use" (Firth, 1968, p. 179). Thus, Firth claims that, since *dark* is characteristically used in conjunction with *night*, collocatability with *night* is one of the 'meanings' of *dark* (1957, p. 196). 'Meaning' here is simply a characterisation of the "other word-material" (Firth, 1968, p. 180) with which *dark* is often used. Since this sophisticated use of the term 'meaning' may be somewhat misleading, Firth's central insight is probably better summarised by the alternate (but synonymous) formulation that habitual collocation is a type of "mutual expectancy" between words (Firth, 1968, p. 181). Collocating words, that is, predict one another, in the sense that where we find one, we can expect to find the other.

It is this idea of mutual expectancy which lies behind the influential modern formulation of collocation as "the relationship a lexical item has with items that appear with greater than random probability in its (textual context)" (Hoey, 1991, p. 7). That is, words are 'collocates' of each other if, in a given sample of language, they are found together more often than their individual frequencies would predict (Jones & Sinclair, 1974, p. 19). Words which stand in such a relationship can be said to 'predict' one another because the presence of one makes the presence of the other more likely than it would otherwise be (Sinclair, 1966, pp. 417-418). This is the sense in which collocation will be used in the present thesis.

The significance of collocation

There are two main reasons why collocation in this sense has been considered linguistically interesting. The first is that a word's typical collocates are thought to give us important information about its semantics. Following Firth's lead, various types of link have been posited between collocation and meaning. The collocational setting in which we encounter a word enables us, it has been argued, to choose between the various possible senses of an ambiguous word. Thus, *commit* - which may mean 'perform'/'carry out', 'take on an obligation', or 'learn by heart'- is not ambiguous in context because each sense has its own distinctive collocates (e.g., *commit a crime, commit oneself, commit to memory*) (Bartsch, 2004, p. 72). While other types of context (e.g. the situational or broader meaning context) may make collocation somewhat redundant for human language users, it has the potential to provide important clues to computerised natural language processing systems in their resolution of ambiguity (Bartsch, 2004, p. 21).

Similarly, the typical collocates of a word provide a profile which can differentiate it semantically from other words with similar meanings. This possibility was pointed out by Halliday (1966), who noted that apparent synonyms, such as *strong* and *powerful*, can have characteristically different collocations (c.f. *strong/*powerful tea*;. **strong/powerful engine*). This idea has been developed by, amongst others, Partington (1998, p. Chapter 2), who shows how near synonyms like *sheer, pure, complete, utter* and *absolute* can be distinguished in terms of their typical collocates. In a similar vein, Hoey (2005, Chapter 5) shows how the different senses of

polysemous words are systematically distinguished by their characteristic co-occurrences, and how violation of these distinct preferences may lead to ambiguity or humour.

A further link between collocation and meaning has been proposed in the idea that the typical collocates of a word can reveal levels of connotation which might otherwise go unnoticed. This point has been developed by Sinclair under the heading of *semantic prosody*. "Many uses of words and phrases", Sinclair notes, "show a tendency to occur in a certain semantic environment" (1987, p. 322). As a corollary, such items may come to carry "an aura of meaning that is subliminal, in that we only become aware of it when we see a large number of typical instances together" (2004b, p. 18). One example is the word *happen*, which, according to Sinclair, characteristically appears together with "something nasty that has happened or is going to happen". Another is *set in* , which again collocates with "nasty things like bad weather" (2004b, p. 18). As Partington remarks, a "phrase like *good times set in* would be highly marked" (1998, p. 67).

While the term 'semantic prosody' is often used to describe cases, such as those cited above, in which a word's typical collocates lend it an aura of evaluative meaning, Sinclair also uses it more generally to refer to the phenomenon by which meaning is 'shared' between words and phrases. Much of the time, Sinclair argues, words do not "constitute independent selections". Rather, co-selection is the norm: "the choice of one word conditions the choice of the next, and of the next again". This sharing of meaning between items entails, Sinclair argues, that "[t]he meaning of words chosen together is different from their independent meanings", and so leads to a certain "delexicalization" of words. This is most obvious in the case of strikingly idiomatic collocations like *sitting duck* or *spill the beans*. It is also found in constructions using 'light' verbs, such as *have an argument* or *take a shower* (which could each be replaced by the single lexical verbs *argue* or *shower*). Less obviously, Sinclair cites such conventionalised phrases as *physical bodies, scientific experiment, full range,* and *general trend*, in which the adjective is heavily delexicalised, adding little in terms of substantive meaning. Such phrases, Sinclair contends, must also be examples of "co-selection" (2004b, pp. 19-20).

To account for the prevalence of such delexicalisation and for the divergent collocational habits of semantically similar words, Sinclair has proposed an 'idiom principle' of interpretation, according to which:

> a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments (1987, p. 319).

This is contrasted with an 'open choice principle', according to which text is composed item-by-item, the only restraint on the concatenation of words being grammaticality (1987, pp. 319-320). While text is sometimes interpreted on the open choice principle (for example, when unexpected lexical choices are encountered), the idiom principle is the normal default mode of interpretation (1987, p. 324).

The exact theoretical status of these two principles requires some discussion. Sinclair calls them "principles of interpretation" (1987, p. 319). This could have two very different meanings: it could mean principles of interpretation for the analyst (i.e. a recommendation for how linguists ought to see a text as being structured), or it could mean principles of interpretation for the text's addressee (i.e. a model of how readers/listeners actually decode language). The distinction is an important one: the first type of principle is a linguist's heuristic tool; the second is a psycholinguistic model of how language is normally processed in the mind. Sinclair does not explicitly distinguish the two possibilities. He clearly believes that linguists should view texts through the lens of the idiom principle (and his own grammars and dictionaries (e.g., 1990; 1995) provide good illustrations of this in action). At the same time, the wording of the principle ("a language user has available") and his diagnosis of its causes ("it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation" (1987, p. 320)) make it clear that he also believes the model to have psychological reality for language users. His dynamic portrayal of the principle at work confirms the point:

> For normal texts we can put forward the proposal that the first mode to be applied is the idiom principle since most of the text will be interpretable by this principle. Whenever there is good reason, the interpretive process switches

> to the open choice principle, and quickly back again. Lexical choices which are unexpected in the environment will probably occasion a switch…(1987, p. 324)

This is clearly intended as a description of the actual process of text interpretation by listeners. Moreover, the idiom principle is not merely – as Sinclair originally describes it - a model of interpretation, it is also a model of language production. We require the idiom principle, he argues, because "[w]e would not *produce* normal text simply by operating the open choice principle" (1987, p. 320, my italics). "At its simplest", he remarks, "the principle of idiom can be seen in the apparently simultaneous *choice* of two words" (1987, p. 321, my italics).

The idiom principle, then, is not merely a linguist's heuristic. It is a psycholinguistic model of language production and comprehension. This brings us to the second main reason why collocations have been considered linguistically interesting: they are thought by many to tell us something important about how language works in the mind. This is a radical step away from the theoretical framework originally outlined by Firth, who saw linguistics as a social science and discouraged linguists from drawing conclusions about the psychological workings of language (1968). The step is not an unproblematic one, as I will argue in Chapter 4. However, the idea that to explain collocation we need to posit some kind of psychological mechanism such as 'chunking' (Ellis, 2001) or 'priming' (Hoey, 2005) has been highly influential and has placed the study of frequent collocation at the centre of the growing body of research into a psycholinguistically-defined 'formulaic language' (Wray, 2002, p. 9), around which have been built models of idiomaticity and fluency (Pawley & Syder, 1983), first and second language acquisition (Ellis, 2003; Tomasello, 2003), language processing (Ellis, 2002a; Jiang & Nekrasova, 2007; Schmitt, Grandage, & Adolphs, 2004), and new approaches to second language education (Lewis, 1993; Nattinger & DeCarrico, 1992).

This 'psychologising' of collocation has been underlined by Hoey (2005) in a radically new definition of the phenomenon. He argues that traditional, frequency-based, definitions (such as his own earlier formulation, cited above) are not adequate since, while frequent co-occurrence is a sound criterion for *identifying* collocation, to

say that it *is* collocation "confuses method with goal". The mere fact of frequent co-occurrence, he points out "gives no clues as to why collocation should exist in the first place". For this, we must define collocation not in statistical, but in psychological terms. Collocation is, in his view, "a psychological association between words" which is merely "evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (2005, pp. 3-5). This psychological association is spelled out in terms of the psycholinguistic notion of 'priming' (to which we will return at some length in Chapter 4). On this model, "a 'priming' word can prompt a language user to recall a particular 'target' word"; e.g. a word like *heart* is recognised more rapidly if a subject has recently seen a related word, like *body*, than it is if they have seen an unrelated word such as *trick*; *body* is then said to prime *heart* (2005, p. 8). Hoey claims that "every word is mentally **primed** for collocational use" (2005, p. 8 original emphasis), i.e. that the selection of one word will make its regular collocates come to mind more readily. This priming is the result, he suggests, of the way in which a word is "acquired through encounters with it in speech and writing", by virtue of which it "becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context" (2005, p. 8).

2.3 Formulaic language

Introduction

If the linguistic patterns and psychological mechanisms proposed by Sinclair and Hoey operated only in collocation, they would constitute an interesting, but perhaps rather minor, aspect of language. However, many researchers believe that the principles at work in collocation spread far beyond this. Over recent decades, researchers working in a number of different fields (including computational linguists, lexicographers, discourse analysts, cognitive grammarians, psycholinguists, and language teachers) have emphasised the importance of what is coming to be called 'formulaic language'. Their work is leading to new ways of describing language, new models of language processing and acquisition, and new methods of language teaching. The present section will provide a brief overview of the major descriptive approaches to formulaic language. It will both discuss the main properties which have been attributed to formulaic language by various frameworks and outline some of the

ways in which these frameworks challenge the dominant Chomskyan model of linguistics. In the first section, we will look at how the 'neo-Firthian' corpus-linguists whose work was discussed above have extended their frequency-based analyses beyond collocation. We will then move on to consider three other perspectives on formulaic language: those of construction grammarians, phraseologists, and discourse analysts. Work on the role of formulas in language learning and teaching, and on the psycholinguistic processing of formulaic language, will be discussed in Chapters 3 and 4.

Frequency-based formulas beyond the word level

Introduction

The central insight of collocation is that some words appear together more frequently than we would expect on the basis of real-world coincidences or traditional linguistic rules (syntax, semantics, register, etc.), and we have seen that some linguists believe that to explain collocation we need to posit some form of psychologically-defined collocational knowledge. The principle of 'more frequent than expected co-occurrence' can be extended beyond word-to-word relations however. The account in Section 2.2 has already hinted at this: semantic prosody involves an abstraction beyond words to sets of related words, while the idiom principle proposes that normal language consists of 'semi-preconstructed phrases'. In fact, both Sinclair and Hoey hold that the types of relations found between words in collocation are replicated across the traditional linguistic levels of syntax, semantics, and even discourse. On this view, the principles underpinning collocation are elevated from the relatively minor role of accounting for a curiosity of lexical usage to a central principle of language in general. Thus, Hoey is able to open his monograph with the bold statement that he will "argue for a new theory of the lexicon, which amounts to a new theory of language" (Hoey, 2005, p. 1). This section will describe this extension of collocation-like mechanisms in the work of Sinclair and his colleagues and of Hoey.

Extended units of meaning and pattern grammar

As we have seen, Sinclair maintains that collocation-like restrictions are found not only between words, but also between words and sets of words with similar evaluative content. This is the thesis of semantic prosody, according to which, for example, the

verb *happens* shows a preference for subjects with a negative connotation, and the phrase *naked eye* tends to be used in structures emphasising difficulty (as in *barely/rarely visible to the naked eye*). Sinclair describes this as an abstraction at the pragmatic level – the word is associated not with a particular collocating word, but with an attitude which can be expressed in a variety of ways (Sinclair, 2004a, pp. 33-34). This is the highest level of collocational abstraction on Sinclair's scheme. One step down from this is *semantic preference* – an association between a word and a set of semantically-related words. This is seen, for example, in the preference of *naked eye* for words related to 'visibility' (*apparent, detect, see, visible,* etc) and in the preference of *brook* for words expressing 'intrusion' (*interference, criticism, contradiction,* etc) (2004a, pp. 32-33). Finally, words often show a tendency to co-occur with items of a particular grammatical type. Thus, *naked eye* tends to come to the immediate right of a preposition (usually *to* or *with*), while *brook* is often found to the right of a modal verb (often *will* or *would*). Sinclair calls this abstraction at the grammatical level *colligation* (2004a, p. 32). Drawing these various levels of association together, he proposes that, rather than taking the word as the basic unit of language, we should recognise extended 'units of meaning', consisting of several orthographic elements, which may be specified as particular words or in more abstract categories, such as grammatical types, semantic sets, or evaluative connotations (2004a, p. 34). Such units are, we can infer, the 'semi-preconstructed phrases' which were proposed on the idiom principle to be 'single choices' for language users.

Fundamental to Sinclair's picture is the idea that "there is a close relation between the different senses of a word and the structures in which it occurs" (1991, p. 53). Different senses of a word not only have their own characteristic collocations, prosodies, etc., but also their own characteristic syntactic realisations. Thus, for example, corpus analysis shows that the three major senses of the word *yield* – to give way; to produce; to lead to – are not distributed at random between the different syntactic forms of that word, but rather show a definite pattern: the first sense tends to be realized as an intransitive verb, the second as a noun, and the third as a transitive verb (1991, pp. 54-57). Words, syntactic forms, and sentential context do not, therefore, constitute independent choices. It is, Sinclair asserts, "folly to decouple lexis and syntax, or either of those from semantics". The concept of "a highly generalized formal syntax, with slots into which fall neat lists of words is suitable only

in rare uses and specialized texts". Most text "is made of the occurrence of common words in common patterns, or in slight variations of those patterns" (1991, p. 108). This constitutes a direct attack on the Chomskyan position that the language system comprises "two kinds of mental tissue": "a lexicon of words" and "a grammar of rules" and that the proper goal of linguistics is to describe highly abstract grammatical patterns which are blind to lexical content (Pinker, 1999, p. 14). This rejection of a strict lexis-grammar dichotomy is, we shall see, characteristic of many models of formulaic language.

Hunston and Francis (2000) have built on Sinclair's work to propose a description of language in terms of *patterns*. They define a pattern as "a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups and clauses that follow the word" (2000, p. 3). A pattern is identified "if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it" (2000, p. 37). The word *matter*, for example, is found often to occur in the expression '*a matter of* –ing' (as in *a matter of developing skills; a matter of learning a body of information; a matter of being able to reason coherently*) (2000, p. 2). The structure '*a ___ of* –ing' may therefore be described as a characteristic pattern of this word.

Like Sinclair's units of meaning, patterns are sense-structure wholes - units which co-specify meaning and form in a single linguistic choice – and no strict distinction is held to exist between lexis and syntax (2000, p. 30). Nevertheless, patterns can be analysed from either 'end'. Thus, where Sinclair focuses on the way in which the different words and different senses of a word can be distinguished by the different patterns in which they appear, Hunston and Francis concentrate on how particular patterns 'select' words with particular meanings. They illustrate the point with the pattern '*it* + link verb + adjective + clause' (as in *It is true/interesting/likely that* or *It is possible/useful/sensible to*). This form is, they point out, instantiated only with adjectives falling into particular 'meaning groups': modality, ability, importance, predictability, obviousness, value and appropriacy, rationality, truth. Moreover, most adjectives appearing in this pattern are associated with a particular type of clause – some (e.g. *true, interesting*) typically being followed by a *that*-clause, others (*useful, sensible*) by *to*-infinitive (2000, p. 29). In short, linguistic rules do not treat all lexical

items equally; or, as Hunston and Francis put it, "patterns occur with restricted lexis" (2000, p. 96).

Hunston and Francis claim that "the majority" of items instantiating a particular pattern will be assignable to broad meaning groups (2000, p. 95). There is, however, no one-to-one correspondence between pattern and meaning. Alongside the 'core' words which appear very frequently in a particular pattern, are a small number of infrequently-occurring 'non-core' items. These are described as existing in "an area of flux" (2000, p. 99) in which patterns may be creatively associated with non-standard lexis. Such creativity is in many cases based on a process of analogy: if a pattern is characteristically associated with a set of words belonging to a particular meaning group, other words of similar meaning may also come to be used in that pattern. Thus, while the verb *provide* is typically found in the pattern 'V n *with* n' (*provide him with money*), it is occasionally also used in the pattern 'V n *to* n**'** (*provide money to him*). This 'creative' use is, it is suggested, generated on analogy with the semantically similar verb *give*. Similarly, *attempt*, which is usually associated with a complementary '*to*-infinitive', is in a small minority of cases attested with a following 'V-*ing*', presumably by analogy with *try*.

In other cases, the only restrictions on extension may be that items fall under a very general meaning. Thus, while the pattern 'N *that*' is typically instantiated with a small number of core items (e.g. *concern, fear, expectation, disappointment*), there also exists a list of 'non-core' items which appear irregularly (e.g. *admiration, envy, joy*) and which appear to have in common only the property that they indicate "a feeling towards a situation" (2000, p. 100). Going beyond this, other patterns appear to supply a meaning of their own, rather than demanding one of their instantiating lexis. Thus while the pattern 'V *way* prep/adv' (as in *talked his way into the post; lie her way out of trouble*) is typically used with the verbs *talk, negotiate, bluff, charm, lie, argue* and *wheedle*, it is also attested with a number of very infrequent non-core items, such as *blather, communicate, persuade*. A precise common meaning is maintained across this diversity of lexis, - i.e. "someone uses clever, devious, or forceful language to achieve a goal, usually extricating themselves from a difficult situation, or getting into a desirable situation", a meaning which, the authors suggest, is supplied by the pattern

itself (2000, p. 100). This picture, as we shall see below, has much in common with the construction grammar model proposed by Goldberg (1995).

Lexical priming

Like Sinclair, Hoey (2005) also finds collocation-like relations at levels beyond the word. We have seen that Hoey sees collocation as a psychological association between words, whereby each word is primed to be used together with certain other words. This 'lexical priming' can be found, Hoey claims, at all levels of language. Like Sinclair, Hoey points to associations between words and other words ('collocation') and between words and groups of semantically-related words ('semantic association', equivalent to Sinclair's 'semantic preference'). While Sinclair's lexical units link into the realm of pragmatics through their semantic prosodies, Hoey calls the tendency of particular words to co-occur with words of a particular pragmatic function (e.g. the tendency of *sixty* to co-occur with 'vagueness markers' such as *about, around, over*) *pragmatic association*. All of these patterns are theorised to be products of lexical priming, working at different levels of abstraction. Hoey also notes that primings are held not only by individual words, but also – in a phenomenon he refers to as 'nesting' - by strings of words. Phrases, he argues, can have associations which are quite different from the associations of their component parts. Thus, *word* collocates with *say*, *say a word* collocates with *against*, and *say a word against* collocates with *won't* (Hoey, 2005, p. 11).

Also like Sinclair, Hoey claims that words are primed to occur with particular grammatical patterns; or, to put it in terms more commensurable with Sinclair's, that lexis and syntax are co-selected (2005, p. 40). He cites three different ways in which this co-selection can be seen, all of which he collects under the broad heading of 'colligation'. First, particular words (or nested groups of words) are primed to co-occur with (or avoid) particular grammatical functions (similar to Sinclair's colligations). Thus, for example, *in winter* is primed to occur with present tense verbs, while *that winter* occurs in Hoey's corpus exclusively with past tense verbs (2005, p. 39). Second, words (and phrases) are primed to occur in (or avoid) particular grammatical functions. Thus, the word c*onsequence* is primed to occur as part of an adjunct or complement, but to avoid occurrence as an object (2005, p. 46). Third,

words (and phrases) are primed to occur in (or avoid) particular sentence positions. Thus, *consequence* is primed to occur in Theme position (2005, pp. 49-52).

This last point brings us close to the realm of discourse, and here Hoey goes beyond Sinclair by claiming that primings exist for particular textual relations. Again, there are three separate claims. First, words (and phrases) are primed to occur in (or avoid) particular types of cohesive relations. He calls this tendency 'textual collocation', and it is demonstrated, for example, in the fact that 81% of occurrences of the word *army* in his corpus is part of a cohesive chain, whereas words such as *asinine, blink* and *particularly* are found to avoid such chains (2005, p. 119). Second, words (and phrases) are primed to occur in (or avoid) specific types of semantic relations, such as contrast, comparison, time-sequence, cause-effect, exemplification and problem-solution. Hoey reports, for example, that the word *sixty* is strongly primed to occur in contrast relations, as the problem component of problem-solution patterns, and – more weakly – in non-contrastive comparison relations (2005, p. 123). Third, words (and phrases) are primed to occur at (or avoid) the beginning or end of independently identifiable 'chunks' of text, such as a sentences, paragraphs, or whole texts. The word *sixty*, for example, is reported to prefer a sentence-initial position (with 200 of 307 occurrences the first word of a sentence) and, moreover, to occur as the first word in a text far more frequently than would be predicted by chance alone (Hoey, 2005, pp. 131-132).

All of these types of priming are cumulative products, Hoey claims, of our history of exposure to the language. Since each language user has a different linguistic history, primings are to a certain extent idiosyncratic to the individual, and – in a phenomenon he calls 'drift' – are liable to change over time (2005, p. 9). Idiosyncrasy has its limits, however. Hoey argues that our ability to communicate with each other points to the existence of certain "harmonising principles" which prevent our primings from diverging too widely. These principles include education, literary and religious traditions, the mass media, and reference works such as dictionaries and grammars (2005, pp. 181-182). It also needs to be noted that priming is not always simply a matter of the most frequent co-occurrences being primed the most strongly. Different types of input are hypothesised as having different degrees of impact on our primings, with particularly valued input (such as literary or religious texts, or the words of a

close friend) liable to be particularly salient, and so to have a disproportionately large effect (2005, p. 12). Different sources of input may, of course, point to different primings – leading to possible conflicts, which Hoey calls "cracks in the priming". A prominent example of this is where consciously learned rules conflict with naturally acquired primings: a sometimes uncomfortable experience which can leave speakers uncertain as to the best form to choose (2005, pp. 178-180). Related to this is the important point that primings tend to be genre- and domain-specific. Primings are acquired in specific situations and will take account of "who is speaking or writing, what is spoken or written about and what genre is being participated in" (2005, p. 13). Language users who intuitively understand this may overcome potential 'cracks' in their primings by reserving different primings for different contexts (2005, p. 179).

For Hoey, a major advantage of lexical priming as a model is that it seems able to explain 'naturalness'; i.e. why of two 'grammatically correct' stretches of language, one might seem idiomatic and the other not (2005, p. 6). While his main interest is with naturalness in this sense, however, Hoey acknowledges that a full account of language must also explain our capacity for linguistic creativity. To this end, he offers a model of how lexical priming might lead to a syntactic capacity capable of producing novel language. He observes that the words which we traditionally group together as nouns, verbs, or adjectives etc. typically share sets of primings. Thus, words like *consequence, aversion,* and *question* have common primings which are not shared by words like *taught* or *if*. The grammatical categories assigned to words, he argues, are simply "a convenient label" for some of these "most characteristic and genre-independent primings". Such categories are not, therefore, prior to lexis, but rather emerge from lexically-specific patterns of priming (2005, p. 154). Like other "nested combinations", they have in turn their own typical primings, and it is these primings which are captured by descriptions of syntax. In an echo of construction grammar models (see below), Hoey proposes that creative language can be explained in terms of these "more general primings", while idiomatic (formulaic) language can be explained by "the more specific primings" (2005, p. 166). As with other types of priming, grammatical categorisations are probabilistic rather than deterministic. Thus, while *winter* is typically used as a noun, it can also function in other ways (e.g. as a verb in *I'll winter in Brussels*) (2005, p. 155). Moreover, because primings are idiosyncratic to the individual speaker, and are subject to drift, there is no "single

grammar to a language" (2005, p. 47) and even individuals' grammars are "never complete" (2005, p. 162).

It should be noted at this point that, while Hoey's framework has enabled him to identify some interesting patterns in language (his analysis of the text-level patterning which certain items appear to follow being particularly original), his use of the psycholinguistic concept of 'priming' is a rather loose one which stands in need of further interrogation. Priming, for Hoey, appears to be a somewhat vague cover-all for a wide range of associations between many different types of linguistic entity. It is introduced with only a passing reference to the substantial psycholinguistic literature on the subject and at no stage are the descriptions of its apparent workings supported with psycholinguistic evidence. If his descriptions are to be taken as a literal model of how language works in the mind, then, much more work is required to back it up. Chapter 4 will aim to take some steps towards this by exploring in some detail whether collocation can indeed be characterised in terms of priming.

Frequency-based formulas and linguistic theory

Both Sinclair and Hoey contend that co-occurrence frequencies indicate principles of linguistic patterning across traditional levels of analysis: words co-occur with other words, with grammatical patterns, with semantic fields, with types of pragmatic force, and with patterns of discourse. I have already noted that this view undermines the strict dichotomy of lexis and syntax which has been maintained by researchers in the Chomsyan tradition. Another challenge to Chomskyan orthodoxy lies in the strongly empirical approach to language study taken by these researchers. The Chomskyan tradition is built on the principle that the primary goal of linguistics is not to account for language as it is manifested in linguistic *performance* – i.e. concrete instances of language use - but rather to describe speakers' *competence* - the abstract system of knowledge upon which this performance is based (Chomsky, 1965). Chomsky is interested, in his own terminology, not in *externalised (e-) language*, but rather *internalized (i-) language*. While i-language – the linguistic system as it is represented in the speaker's mind – is the proper subject of linguistics, e-language, "if it exists at all, is derivative, remote from mechanisms and of no particular empirical significance, perhaps none at all" (Chomsky, 1991 quoted in Cook and Newson, 1996, pp. 21-2). The work of Sinclair and Hoey, in contrast, is based entirely on the inspection of 'e-

language', as it is manifested in corpora. On Sinclair's view, the distinction between competence and performance, introduced to enable the linguist to abstract regularity from the chaos found in real language production, becomes obsolete once we start working with large-scale corpora, which enable us to identify the most typical patterns from amongst the 'noise' of performance errors without resorting to idealisations (Sinclair, 1991, p. 103).

Related to the rejection of these dichotomies is a further fundamental difference between the neo-Firthian and the Chomskyan schools: whereas the latter has aimed to describe rules which can generate all of the *possible* sentences of a language, the former has focused on describing what language is most *probable* in use. In Hoey's terms, they have aimed to account for 'naturalness' in language. We shall see that this focus on the natural and probable rather than the merely possible, and the attendant rejections of the lexis-syntax and competence-performance dichotomies are mirrored in a number of different approaches to formulaic language. We shall now turn to the major representatives of these.

Construction Grammar

Following Croft and Cruse (2004, p. 257), I will take the general heading of 'construction grammar' to cover a set of models incorporating Construction Grammar (Fillmore, 1979; Kay & Fillmore, 1999), Cognitive Grammar (Langacker, 1987; 1991), Cognitive Construction Grammar (Goldberg, 1995; 2006) and Radical Construction Grammar (Croft, 2001). Construction grammar is coming to provide one of the key descriptive frameworks for formulaic language, and is used extensively in research on the acquisition and processing of formulaic language (see, e.g., the chapters in Barlow & Kemmer, 2000; Robinson & Ellis, 2008). It will therefore be worth spending some time in discussing the details of this framework. The differences between the various construction grammar models will be dealt with only briefly here (but see Croft and Cruse, 2004:257-290 for a review); I will focus rather on the core shared tenets which together define a distinctive construction grammar model.

The main ideas behind construction grammar can be summarised as follows: The basic units of language are conventionalized pairings of form and meaning dubbed

'constructions'. Constructions vary in complexity, from single morphemes (e.g. the word *run,* or the 'plural' construction, which associates the suffix *–s* with the meaning 'plural') to longer utterances (e.g. *let the cat out of the bag*, associated with the meaning 'divulge hidden information'). They also vary in degree of lexical and structural specificity, from concrete lexical forms (such as those cited above) at one extreme to highly abstract patterns (such as the transitive construction: Subj-V-Obj) at the other, with various degrees of schematicity in between (e.g. partially-specified conventional templates such as *the __er the __er*, as in *the bigger the better*). Each construction is represented independently in the mind of the speaker, and knowledge of a language is knowledge of its constructions.

The impetus towards this model came in large part from dissatisfaction with traditional Chomskyan treatments of idioms. As was mentioned above, on the Chomskyan model, language is represented in terms of 'words' and 'rules' – fixed conventional symbolic units and highly general productive principles governing their combination. The latter account for all that is regular and productive in language, the former for all idiosyncracies (Pinker, 1999). Idioms pose a problem for this picture because they are both irregular and productive. Their irregularity requires that they be entered in the lexicon, while their productivity means that a fixed lexical entry cannot adequately characterise their behaviour. Theorists such as Katz (1973) and Fraser (1970) have suggested ways in which UG might accommodate idioms. However, there are reasons not to be satisfied with their proposed solutions.

Firstly, these models have relied on *ad hoc* stipulations, treating idioms as different in kind from the rest of the language system. While this is in keeping with the Chomskyan view that such features are 'peripheral' to language, construction grammarians have argued that it is illegitimate to marginalise idioms in this way. Far from being 'peripheral', figurative language is, they claim, "pervasive and fundamental". Indeed, it is asserted, "if figurative language were systematically eliminated from our database, little if any data would remain" (Langacker, 1987, p. 1). Being so central to the system, idioms ought to be accommodated by the mainstream of any theory of language, not shuffled off to the periphery and handled through special pleading. Construction grammars therefore take it as axiomatic that "the relatively general patterns of the language…and the more idiomatic patterns…stand

on an equal footing as data for which the grammar must provide an account" (Kay & Fillmore, 1999, p. 1) and believe that the theoretical machinery which underlies the former ought to be the same as that which underlies the latter, such that a satisfactory theory of idiomatic patterns ought also to provide an account of the language as a whole (Fillmore, Kay, & O'Connor, 1988, p. 535; Goldberg, 1995, p. 5).

Secondly, construction grammarians argue that the Chomskyan models fail to give an adequately rich description of the patterning which a more careful examination shows idioms to exhibit. Fillmore et al (1988) make this point through an extended analysis of the idiomatic form *let alone*. Their examination of the syntax, semantics and pragmatics of this 'formal idiom' reveal it to behave in a way which is productive and highly structured, but which could not have been predicted from a knowledge of the rest of the language. To use such an idiosyncratic phrase appropriately, they assert, "more is needed than a system of general grammatical rules and a lexicon of fixed words and phrases" (1988, p. 535). Rather, there must be something like "a special mini-grammar embedded within the general grammar, whose properties are not deducible from those of the larger grammar" (1988, p. 510). Following Fillmore et al's example - and Lakoff's (1987) comparable analysis of the '*There* construction' – construction grammarians have subsequently uncovered a range of similarly idiosyncratic constructions throughout the language system (see Croft & Cruse, 2004, pp. 240-247 for a review). Their analyses suggest that speakers must "possess an extraordinary range of specialized syntactic knowledge that goes beyond general rules of syntax and semantic interpretation on the one hand, and a list of substantive idioms on the other" (Croft & Cruse, 2004, p. 241).

In place of 'words and rules', construction grammarians have hypothesised that knowledge of language requires "a repertory of clusters of information including, simultaneously, morphosyntactic patterns, semantic interpretation principles…and, in many cases, specific pragmatic functions" (Fillmore et al., 1988, p. 535). These 'clusters of information' are *constructions,* and they are posited to be the basic units of language. Constructions are "conventional association[s] of linguistic form and content" (Kay & Fillmore, 1999, p. 2), which cut across the traditional division of language into separate components for phonology, syntax, semantic and pragmatics (Croft & Cruse, 2004, p. 247). They are defined as form-meaning pairings, some

aspect of whose "form or meaning is not strictly predictable from the properties of their component parts or from other constructions" (Goldberg, 1995, p. 4). That is to say, they are conventional in the sense of being "something a language user could fail to know while knowing everything else in the language" (Fillmore et al., 1988, p. 504).

Defined in these terms, constructions can be shown to exist across the continua of simplicity-complexity and concreteness-abstractness. Individual words are obviously conventional form-meaning pairings, as are complex phrases like *let the cat out of the bag* and schematic templates like *What's X doing Y?* (as in *what's this fly doing in my soup?*). Less obviously, more abstract forms with no fixed lexical components are also seen as conventionalized symbols. Goldberg shows how basic structures of English – amongst others, the *ditransitive structure* Subj V Obj Obj_2 (e.g. *Pat faxed Bill the letter*), the *caused motion structure* Subj V Obj Obl (e.g. *Pat sneezed the napkin off the table*) and the *resultative structure* Subj V Obj Xcomp (e.g. *She kissed him unconscious*) – are conventional form-meaning pairings, that "themselves carry meaning, independently of the words in the sentence" (1995, p. 1). She shows, for example, that the caused motion structure Subj V Obj Obl (as seen in *They laughed the poor guy out of the room; Frank sneezed the tissue off of the table; Mary urged Bill into the house*) has a core meaning ('X CAUSES Y TO MOVE Z') which cannot be derived either from the lexical items which instantiate it or from other structures in the language (1995, Chapter 7).

Fillmore et al (1988, p. 535) claim that even the most abstract descriptions of phrase structure grammar (such as VP $\rightarrow$ V NP) can be treated in this way and assigned (albeit highly general) conventional meanings. On this view, there is no ultimate distinction between lexis and grammar. "Lexicon, morphology, and syntax form a continuum of symbolic structures, which differ along various parameters but can be divided into separate components only arbitrarily" (Langacker, 1987, p. 3). Langacker suggests that a false dichotomy between lexis and grammar is generated when linguists "focus solely on representative examples from the two extremes of the continuum". Restricting their attention to forms like *giraffe* and *encyclopedia* on the one hand and forms like *–ing,* and *of* on the other, linguists find striking differences in terms of concreteness of sense, amount of semantic content, syntagmatic restrictions and openness of the class to new members. Greater attention to intermediate cases

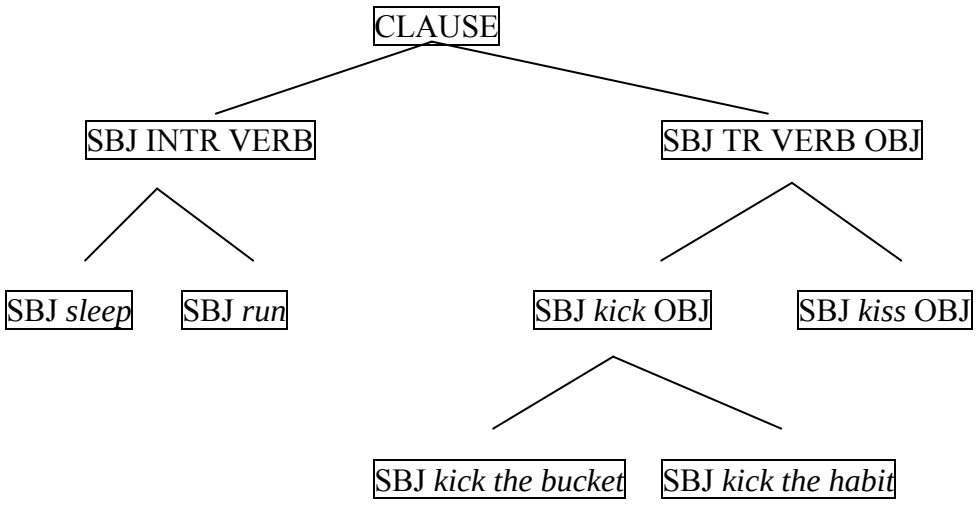
shows distinctions to be less clear-cut. Langacker points out, for example, that content words vary in their concreteness of meaning (c.f. *kick* > *talk* > *think* > *live* > *exist*) and level of semantic specificity (*giraffe* > *mammal* > *animal* > *organism* > *thing*), and that most grammatical morphemes are at least as meaningful as some lexical words (e.g. modals, quantifiers and prepositions do not seem to have less semantic content that the 'lexical' items *thing* or *have*). Moreover, there is often freedom of choice between grammatical morphemes, which may be alternated to express different construals of a situation (e.g. *try to complain* vs *try complaining*) while the choice between lexical items in a given sentential context is often more constrained than the traditional view would suggest (as in restricted collocation, such as *strong/*powerful coffee*) (1987, pp. 18-19).

Syntactic rules, then, take their place at one end of a continuum of schematicity that ranges from entirely fixed expressions, through expressions allowing minimal variations (as in *he's kicked/gonna kick the bucket*) and semi-fixed expressions with various types of open slot (*give* NP *the lowdown*; __*let alone* __) to completely schematic principles of phrase structure (as in the representation of the verb phrase as the construction [*V NP*]) (Croft & Cruse, 2004, p. 249). A parallel analysis can be given for semantics: utterances conforming to general compositional rules of interpretation differ in degree, not in kind, from conventional idioms. Following the analysis of Nunberg et al (1994), this view sees idioms not (as Katz (1973) and others have described them) as semantically non-compositional chunks to which we must assign holistic meanings, but rather as compositional structures whose sub-components have special meanings that apply only in the context of a particular expression. Thus, the idiom *spill the beans* can be analysed into two subcomponents: *spill*, which in this context means 'divulge'; and *the beans,* which here means 'information'. *Spill the beans* is a construction because the rules of interpretation linking its sub-components to these meanings are not found in the rest of the language, and so must be specified separately for this particular phrase. Various degrees of restricted collocation - such as *curry favour* (in which, uniquely to this context, *curry* means something like 'to attempt to win') and *table a motion* (in which the verb takes on a particular meaning only in conjunction with a small family of related nouns) – show that the same principle operates at different levels of generality. General

compositional rules are the extreme end of this continuum, the most abstract rules of interpretation (Croft & Cruse, 2004, pp. 249-263).

On the constructionist view, then, grammar is "a **structured inventory of conventional linguistic units**" (Langacker, 1987, p. 57 emphasis in original) at different levels of complexity and abstraction. A key issue theorists have sought to address is that of how this inventory is structured. All of the versions of construction grammar described here agree in representing relations between constructions in terms of taxonomic networks, in which each construction – i.e. each structure whose syntax or semantics is not predictable from properties of other constructions – is represented as a node. The taxonomic relationships between constructions capture the fact that most structures which require independent representation are at the same time instances of more abstract schemas. In this way, linguistic regularities at various degrees of generality can be uniformly represented alongside idiosyncracies. Croft and Cruse (2004, p. 264) give the example shown in Figure 1.

Figure 1: example of a construction network (Croft & Cruse, 2004, p. 264)



While *kick the bucket* and *kick the habit* require separate representations because of their idiosyncratic semantic interpretations, they are nevertheless both instances of the construction [SBJ *kick* OBJ], which itself requires independent representation to specify the argument structure of the verb *kick,* and is in turn an instance of the superordinate [SBJ TR VERB OBJ] construction, and so on. Not shown in the network is

the fact that locutions can have multiple parents. Since parent constructions are more abstract forms of their instantiating daughters, they must by definition leave some aspects of them unspecified. Aspects of a locution's structure on which one parent is neutral can be specified by another. Goldberg (2006, p. 21) illustrates this with the expression *A dozen roses, Nina sent her mother*, which has eleven parents:

a. Distransitive construction
b. Topicalization construction
c. VP construction
d. NP construction
e. Indefinite determiner construction
f. Plural construction
g. *dozen, rose, Nina, send, mother* constructions

Such combinations are constrained only by the condition that parents can be construed as not conflicting with one another.

Different versions of construction grammar disagree as to how redundant information should be stored in the taxonomy. Clearly, a full specification of the properties of a relatively concrete construction will repeat much that is already specified at higher levels. Thus, a full characterisation of *kick the bucket* will repeat information which is also listed for its parent [SBJ TR VERB OBJ]. An important difference between theories lies in how they handle this redundancy. On Kay and Fillmore's (1999, pp. 8-9) account, constructions in the lower nodes of the network *inherit* all of the features of their superordinates, and information is specified only at the highest possible level. Goldberg (1995, pp. 73-74) describes this transfer of all of a parent's features to its daughters as a *complete mode* of inheritance, and the concomitant non-duplicating mode of storage as *impoverished-entry*. She contrasts these models with a *normal mode* of inheritance and *full-entry* of redundant information. In the normal mode of inheritance, constructions lower in the taxonomy may block any information from their parents with which they conflict. The principle of full-entry means that each construction is individually fully specified, detailing within itself not only unique information but also all of the 'redundant' specifications which it inherits from its parents. Rather than an on-line process, then, inheritance is on this view 'a static

relation defined by shared information'. Normal-mode inheritance and full-entry are required, Goldberg claims, in order to specify 'partial generalisations', and to resolve conflicts arising when multiple parents contain contradictory information (1995, pp. 97-98). Such a model looks inefficient, but, as Croft and Cruse point out (2004, p. 278), the profligacy of full-entry models in terms of *storage* is balanced by savings in terms of *computing*. A complete inheritance model achieves storage parsimony at the price of greater on-line processing, as information must be accessed from elsewhere. A full-entry model, on the other hand, is parsimonious in terms of computing by making all information ready-to-access at all points.

This view is supported by Barsalou's (1992 quoted in Croft and Cruse, 2004, p. 278) assertion that "concepts and properties in human knowledge are organized with little concern for elegance and [storage] parsimony". As Croft and Cruse point out, however, an assumption that full-entry must be preferred in every instance is as misguidedly *a priori* as its opposite (2004, p. 278). For this reason, most construction grammarians now adopt a *usage-based model* of language processing (Kemmer & Barlow, 2000), according to which constructions are detailed in full only if they are used with sufficient frequency for the resultant savings in computing effort to outweigh the cost of storage (Goldberg, 2006, p. 64). We have seen that construction grammarians have defined constructions as form-meaning pairings, some aspect of whose "form or meaning is not strictly predictable from the properties of their component parts or from other constructions" (Goldberg, 1995, p. 4). However, considerations of cognitive efficiency have recently led many to extend this definition to include a frequency-based component. On this broader view, linguistic patterns are also "stored as constructions even if they are fully predictable as long as they occur with sufficient frequency" (Goldberg, 2006, p. 5), a position which has much in common with that of the neo-Firthian corpus-linguists, described above.

Before we move on from construction grammar, it should be noted that usage-based views of language extend in a more radical way the challenge to the Chomkyan distinction between competence and performance described above. Usage-based models hold that aspects of language use (Chomsky's *e-language*) such as frequency, far from being irrelevant to the linguistic system (Chomsky's *i-language*), are partially

responsible for determining the very structure of that system. According to a strong version of this view:

> "the structure of the linguistic system is not separate in any significant way from the (cumulative) acts of mental processing that occur in language use. The speaker's linguistic ability, in fact, is *constituted* by regularities in the processing of language" (Barlow & Kemmer, 2000, p. xi).

Phraseology

Under the heading of 'phraseology' I will group together the pedagogically-oriented work initiated by Palmer (1933) with that of the so-called 'Russian' or 'Soviet' lexicologists (as described by, for example, Weinreich (1963) and Cowie (1998b)) and contemporary writers such as Cowie (1981b; 1994) and Mel'cuk (1998). The work of Palmer and the Soviet lexicologists developed independently, but their descriptive schemes have much in common. They share a primary focus on the construction of reference materials for foreign language learners, and Cowie (1998b) notes that contemporary EFL lexicologists have drawn on both approaches. In recent years, applied linguists with an interest in analysing learner language (Howarth, 1998; Kaszubski, 2000; Nesselhauf, 2005) have also drawn on this dual tradition. Phraseological approaches to collocation are often contrasted with the frequency-based approach of Sinclair and Hoey (e.g., Herbst, 1996; Nesselhauf, 2005). However, I shall argue that there is considerable overlap between the two traditions, and that both can be accommodated within a usage-based construction grammar framework.

In his 'Second Interim Report on English Collocations', Palmer (1933) grouped together under the heading of 'collocation' such diverse 'comings-together-of-words' as *to strike while the iron's hot, the United States*, *thank you*, *all at once, next week, onlooker* and *to commit suicide*. What these items had in common, he claimed, was:

> that (for various, different and overlapping reasons) each one of them must or should be learnt, or is best or most conveniently learnt as an integral whole or independent entity, rather than by the process of piecing together their component parts (1933, p. 4)

The most characteristic types of collocation, on Palmer's analysis (1933, pp. 8-10), include:

- *heterosemes*: "in which at least one of the component words assumes a new and particular meaning by reason of being collocated with the other component", e.g. *to fall out* (= to quarrel), *the Civil Service*, *in order to*;
- *verb x object collocations*: regular pairings which the learner needs to know to avoid making a mistake, e.g. *ask a question* (not *make a question*); *do a favour* (not *perform a favour*); *give trouble* (not *do trouble*);
- *verb x preposition collocations*: again regular pairings which need to be learned, e.g. *agree with, help oneself to, rely upon*;
- *absence of article*: e.g. *to go to bed, to get hold of, to give way*;
- *coined collocations*: deliberately created by an individual or association, e.g. *Chamber of Commerce, Proper noun, one-way road*;
- *collocations without a break*: "comings-together of two words" which are written as a single word e.g. *downstairs, lighthouse, birthday*;
- *construction-patterns*: phrases which provide examples of grammatical rules which are specific to certain items of lexis and so could not be learnt except through the memorization of such specific instances. He gives the example of *to be difficult for somebody to do something*, which instantiates the generative pattern: *be* (*get, grow*, etc.) x ADJ (x for x INDIRECT OBJECT) x to x INFINITIVE (x OBJECT).

Palmer's main aim was to provide a classified listing of collocations which could serve as a basis for designing improved versions of the 'limited vocabularies' used for writing simplified readings for learners, and for creating better learner dictionaries or grammars. Soviet school lexicologists were also driven primarily by the need to create foreign language resources, particularly bilingual dictionaries (Weinreich, 1963, p. 61). Cowie, whose own lexicographical work was influenced by both Palmer and the Soviet school, notes that, while the former has the virtue of offering a careful syntactic classification of collocations, the latter manages to address an important variable absent from Palmer's classification: the 'degrees of variation' allowed by collocations,

and the connection between this and the idiomaticity of their elements (Cowie, 1998b, p. 213).

The Soviet tradition specified as its object of interest the 'phraseological unit', defined by Ginzburg et al (1979, cited in Cowie, 1989b, p. 214) as "non-motivated word-groups that cannot be freely made up in speech but are reproduced as ready-made units". The primary focus of work in this tradition has been on classifying units according to the two criteria of semantic opacity (the degree to which words are used with their 'dictionary' meanings) and 'fixedness' of combination (the degree to which elements of a phrase can be substituted) (Cowie, 1998a, pp. 4-5; Weinreich, 1963, p. 73).

Vinogradov is credited with laying the foundations of this work. He made a three-way distinction between 'phraseological-fusions', 'phraseological unities', and 'phraseological combinations' (or 'collocations') (Cowie, 1998a, pp. 4-5; Weinreich, 1963, p. 73) . Phraseological fusions are "'unmotivated' (or semantically opaque)" combinations which are "generally structurally fixed" (Cowie gives the English example of *spill the beans*) (1998a, p. 5). Phraseological unities are partially motivated phrases. They are semantic wholes, but there is a non-arbitrary, figurative, connection between the phrasal meaning and the usual meanings of the component words (e.g. *blow off steam*, where an originally transparent meaning has been extended by metaphor). Phraseological combinations are phrases comprising two open-class words, one of which maintains its literal sense, while the other is used figuratively (e.g. *meet the demand*, where *demand* has its usual meaning, but the sense of *meet* is highly context-dependent (Cowie, 1998a, p. 5)). Amosova further divided this last category into two sub-groups. The first is 'phrasemes', in which the figurative sense of the bound word is found only in conjunction with a single collocate; e.g. the three different combinations *small talk*, *small hours* and *small change*, where the three meanings of *small* ('trivial', 'early', 'of low value') are exclusive to these phrasal contexts. The second is 'phraseloids', in which the bound word can carry its meaning in conjunction with several other items (e.g. *pay one's respects/a compliment, court to someone*) (Cowie, 1998b, p. 215).

Modern phraseologists have essentially followed these categorisations, though with different terminology. Cowie (1994), for example, divides phrasal language into 'formulae' (corresponding to the sentence-like units described above), 'idioms' (which may be 'pure' – Vinogradov's fusions – or 'figurative' – Vinogradov's unities) and 'restricted collocations' (phraseological combinations). Word combinations which allow for more-or-less open substitution of elements are referred to as 'free' (Cowie, 1981a, p. 226). While this overall division remains essentially the same, modern writers have both extended and more finely categorised the class of collocations. Mel'cuk (1998, pp. 30-31) provides a four-way division of collocations. He splits the traditional class (in which the meaning of the bound word differs from its dictionary definition) into cases where the bound word is 1) 'empty' (as in *do a favour, take a step*), or 2) carries a signification it only has in combination with this partner word or a few other similar words (*black coffee; french window*). Going beyond the traditional definition, he also includes as collocations some pairs in which the meaning of the dependent lexeme is the same as its dictionary meaning – i.e. semantically transparent pairings. In this case either 3) this meaning cannot be expressed in conjunction with the paired word by any synonym (*strong (*powerful) coffee*) or 4) the meaning is specific to the meaning of the paired word, and so 'bound' by it (*acquiline nose; rancid butter*).

Howarth (1998, pp. 169-170), dealing exclusively with verb-noun pairs, identifies five 'levels' of increasingly restricted collocations. At Level 1, there is "[f]reedom of substitution of the noun" but "some restriction on the choice of verb" (e.g. *adopt/accept/agree to* a *proposal/suggestion/recommendation/convention/plan, etc.*). At Level 2, "[s]ome substitution of both elements" is permitted. That is, "a small range of nouns can be used" with "a small number of synonymous verbs" (e.g. *introduce/table/bring forward a bill/an amendment*). At Level 3, there is "[s]ome substitution of the verb" but "complete restriction on the choice of noun". That is, a "small number of synonymous verbs" can be used with a particular sense only in conjunction with a particular noun (e.g. *take/pay heed*). At Level 4, there is "[c]omplete restriction on the choice of verb" but "some substitution of the noun". That is, for "a small range of nouns", only one verb can be used for a particular meaning (e.g. *give the appearance/impression*). At Level 5, there is "[c]omplete restriction on the choice of both elements. The verb cannot, with its given sense, be

used with any other noun, and there are no synonymous verbs that can be used in its place (e.g. *curry favour*). Cutting across these levels, Howarth (following Aisenstadt and Cowie) also identifies three different types of semantic specialization: *figurative* (e.g. *bring up children, reach a conclusion*); *delexical* (e.g. *have a chance, make an investment*); and *technical* (e.g. *carry a motion, obtain a warrant*).

A key area of interest in the phraseological approach has been the interface between its two defining criteria of semantic opacity and fixedness. It is a basic tenet of Soviet lexicology that most words are polysemous, and that their relevant submeanings are determined "according to the grammatical and phraseological context in which they occur" (Weinreich, 1963, p. 67). Restricted collocations can be characterised on this view as word pairings in which one element carries a meaning which it only has in combination with its given partner (or with a small group of such partners). The 'semantic opacity' of the bound element, and the 'non-substitutability' of its partner therefore go hand-in-hand. The identification of collocations can also proceed in one of two ways: from the perspective of semantic opacity or from the perspective of substitutability. Traditional Soviet lexicology favoured the former approach, defining collocations as pairs which "contain one component used in its direct meaning while the other is used figuratively" (Arnold, 1986, quoted in Cowie, 1989b, p. 215). However, we have seen that Mel'cuk (1998) also uses the term *collocation* for pairings in which both elements have their usual meanings, but in which at least one element cannot be substituted for a synonym. Writers such as Cowie (1994) and Howarth (1998), meanwhile, make substitutability the primary criterion for collocation, seeing figurative meaning as a typical, but not necessary, property. Thus, Cowie defines collocations as pairs "characterized by arbitrary limitation of choice at one or more points" and, like Mel'cuk, recognises as collocations "combinations whose elements have neutral meanings" (e.g. *cut/*slash one's throat; slash/*cut one's wrists*) as well as those in which "one of the constituents is used in a figurative sense" (1994, p. 3169). Nesselhauf makes a case for abandoning the criterion of semantic opacity altogether and uses arbitrary restrictions on substitution to distinguish collocations from both free combinations on the one hand and idioms on the other (2005, pp. 25-34).

At this point, it seems that the analyses of the phraseologists have the potential to intersect with those of the neo-Firthians, with whom they have often been set in contrast (e.g., Herbst, 1996; Nesselhauf, 2005). 'Restrictions on substitution' must ultimately be evidenced empirically, and it may be that frequency measures which are capable of measuring the degree of mutual predictability between words (such as mutual information, see Section 4.4) will prove a reliable way of achieving this. This suggests that there may be more overlap between the two approaches than is often recognised. The exact nature of this overlap would be a question worthy of further research.

It is also worth noting that the central criteria of the two approaches - frequency in the neo-Firthian tradition, and semantic/substitutional anomaly in the phraseological tradition – correspond exactly to the two sides of the definition of 'construction' used by usage-based construction grammars (see above): i.e. linguistic items which are independently represented in the language system, either because they are not predictable on the basis of other knowledge or because they are sufficiently frequent for their independent storage in long term memory to be cognitively efficient (Goldberg, 2006, p. 5) (this definition itself, of course, echoes that of Palmer (above) of collocations as items which "for various and overlapping reasons" require independent learning (1933, p. 4)). The two frameworks are probably best seen, therefore, as overlapping and complementary, rather than alternative, approaches to the study of collocation.

Discourse analytic approaches

A number of researchers have focused on formulaic language from a broader discourse or ethnomethodological perspective. This work offers an important new perspective in that it engages more fully than other approaches with the relationship between formulaic language and differing contexts of production. Whereas other frameworks have concentrated largely on the benefits of formulaic language in reducing the processing load of fluent language use, discourse analysts have often twinned this motivation with social goals.

Central to discourse analytic approaches has been the recognition that, though grammar provides us with a theoretically infinite range of utterances, only a small

proportion of these are considered appropriate in context by native speakers. Like Hoey (2005), therefore, they hold that linguists need to account not only for what is possible, but also for what is natural, or probable in a certain situation. Coulmas (1981) spells out the point in his discussion of *conversational routines*. While the great creative power of an unfettered combinatorial grammar would predict "that almost every sentence has an occurrence probability of close to zero", in fact "a great deal of communicative activity consists of enacting routines making use of prefabricated linguistic units in a well-known and generally accepted manner". Much of our interaction is repetitive and, "as similar speech situations recur, speakers make use of similar and sometimes identical expressions". "Conversational routines", on this view, are "highly conventionalized prepatterned expressions whose occurrence is tied to more or less standardized communication situations" (1981, pp. 1-3). This definition covers a broad range of phenomena – including idioms, collocations, and even the large-scale conventionalised structures which shape the accepted formats of a routine conversation. "Wherever repetition leads to automatization", Coulmas writes, "we would call a performance a routine" (1981, p. 3). Many conversational routines can be characterised in Gricean terms as conventional implicatures - indirect speech acts whose interpretations do not have to be calculated during the course of conversation, but are known in advance thanks to their frequent use (1981, p. 7). While conversational routines are partly a form of social cement – "tools which individuals employ in order to relate to others in an accepted way" (1981, p. 2) - Coulmas also stresses their benefits for individual speakers in terms of psycholinguistic processing. Quoting Ladefoged's remark that the central nervous system "has rapid access to items in a very large memory, but comparatively little ability to process these items when they have been taken out of memory", he suggests that "routine formulae can be drawn from the memory without much effort, and, at the same time, they give us time for conversational planning" (Coulmas, 1981, pp. 9-10).

Kuiper (2004) also argues that routine performance supports fluency. He reviews Lord and Parry's studies of illiterate oral bards in the former Yugoslavia, who achieve the difficult task of composing poems in real time while maintaining fluency and adapting to audience reactions. Kuiper comes to the conclusion that these bards rely on formulaic performance - including both "formulaic phrases which are traditionally keyed to specific episodes" (2004, p. 37) and generic plot outlines. Because "it relies

on the resources of the tradition, formulaic performance is only possible in routine contexts. That is, in situations where there is an expectation that things will happen in much the same way as they have happened before" (2004, p. 39). Building on this idea, Kuiper goes on to report his own research into the use of formulaic language in two other high-pressure but routine situations - sports commentary and auctioneering. He again finds a correlation between formulaicity and performance pressures, and also notes the socio-cultural significance of the formulas used – the insider knowledge to utilise the scripts provided by a tradition plays a significant role in "the construction of the social self" (2004, p. 44). As Kuiper puts it: "We play parts, and a good deal of what it means to play a part is learning the lines" (2004, p. 44).

The social significance of formulaic language is underlined by broader sociolinguistic studies. Kuiper cites work by Ji on the use of routine formulas before, during and after China's Cultural Revolution (Kuiper, 2004, pp. 45-46). It was found that old formulas, bound up with old ways, were "either proscribed or altered to represent the new order", and that the formulaic inventory mirrored "each twist and turn of ideological and political direction" during the Revolution. Kuiper argues that "linguistic engineering through young people's desire for conformity in being like their peers came to be exploited for socio-political ends", and concludes that "formulaic speech is not only sensitive to socio-cultural change but can be manipulated by the powerful for socio-political ends" (2004, p. 46). This idea has also been explored by Stubbs (1996).

Another discourse-based approach to formulaic language is found in the work of Tannen (1989), who looks at what she calls 'prepatterning' in conversation. She claims that "all discourse…is more or less prepatterned". All text consists of prefabrications of various sorts, and, since all meaning is derived through previous associations, semantics itself is "a matter of prior text" (1989, pp. 42-43). Tannen presents a model on which prepatterning varies along three scales of fixity: *fixity of form, fixity of context* and *fixity of time*. Highly fixed in both context and form are *situational formulas*: expressions which are "always uttered in exactly the same way and are associated with – indeed expected in – certain situations", to the extent that their "omission would be noticed and disapproved". Such formulas are not common in English, but much use is made of them in, for example, Arabic, Turkish, and Greek (1989, pp. 38-39). Equally fixed in form but less so in context are proverbs and

sayings. Again, Tannen notes a suspicion of such fixed phrases by Americans, and points out that speakers of English often produce variations on canonical forms, utilising them as a resource for creativity (a point taken up at greater length by Carter (2004)). Prepatterning is also seen at higher levels of discourse – in terms both of recognisable patterns of discourse organisation and, more abstract still, of culturally-specific notions of "what seems self-evidently appropriately say, indeed, to think, feel, or opine" (1989, p. 44).

Fixity with respect to time refers to the "relative longevity" of prepatterning (1989, pp. 45-46). At one end of the scale is "ephemeral language which is picked up and repeated verbatim in a given conversation and then forgotten"; at the other are those phrases and texts which remain lodged in the cultural lexicon for centuries (Biblical and Shakespearean quotations being the obvious examples). In between these extremes, we find the *private languages* developed by individuals and groups and fashionable terms and phrases which regularly pass in and out of a culture.

Like the other discourse analysts discussed here, Tannen notes the importance of such language in easing the cognitive burdens of language production and comprehension and in asserting socio-cultural identities. Taking the latter aspect one step further, she asks why we should be driven towards repetition; why fixity is emotional and distinctive, rather than boring and bland. Quoting approvingly Freud's assertion that "[r]epetition, the re-experiencing of something identical, is clearly in itself a source of pleasure", she speculates that this drive could serve the purpose of underwriting learning (1989, p. 94). The idea that a love of repetition might provide us with an evolutionary survival advantage is discussed in more detail by Cook (2000).

Key themes in formulaic language

Wray has observed that the variety of methodological approaches and research agendas of linguists interested in formulaic language makes it difficult to identify any single "standard view of what formulaic language is" (2002, p. 261), and doubts that formulaic language constitutes a "single linguistic phenomenon" (2002, p. 44). However, a number of common themes have emerged from the literature reviewed here.

Firstly, an interest in formulaic language tends to go together with a primary focus on what is idiomatic, natural, or socially-acceptable in language, rather than simply on what is syntactically permissible. Secondly, all of the approaches described here are interested in linguistic restrictions which are not predicted by the general rules of the language. In particular, there is a focus on rules which are of limited scope in that they are tied to particular lexical items, or particular situations, or which appear to embody a syntax not predicted by what is found elsewhere in the language. Thirdly, no strict distinction is held to exist between lexis and syntax; these terms are held instead to form extreme points on a linguistic continuum. Finally, most of the approaches mentioned here are interested in the idea that a mental language system which frequently recycles memorised forms may be more efficient (and so is probably more psychologically plausible) than one which generates every utterance 'from scratch' on each occasion of use.

One reason why the formulaic language movement is a crucial one in contemporary language study is that all of these standpoints are in direct contrast to the axioms of the traditional Chomskyan approach to linguistics. The Chomskyan paradigm has held that linguists should be concerned with what is possible, rather than what is natural, with rules at the highest possible level of abstraction (ultimately in a language-general 'universal grammar'), rather than in limited-scope restrictions, has relied on a sharp distinction between syntax and lexis, and has emphasised the importance of descriptive economy, rather than cognitive plausibility, in evaluating models of grammar. In questioning these Chomskyan principles, the formulaic language movement aims to effect a major change in our conception of language and how it should be studied.

2.3 Summary and conclusions: collocations and formulaic language

This chapter has introduced the concept of high frequency collocation and discussed why researchers have been interested in it. Collocation has been seen by some as an archetype of the broader phenomenon of formulaic language. We have looked at some of the major approaches to describing formulaic language and have seen that these pose a radical challenge to traditional Chomskyan linguistic theory. Such a major shift

in our view of the nature of language raises important questions about how languages should be taught, and this will be the subject of the next chapter.

Chapter 3
Formulaic language and second language teaching

3.1 Introduction

Formulaic language has long played a role in second language teaching. Most obviously, beginner learners and holiday-makers have always been offered ready-made situational phrases for greeting people, asking directions, ordering a drink, etc. as a quick and easy route into communication (Wray, 2000, p. 463). The importance of formulas for inexperienced users of a language is reflected, as Ellis (2001, pp. 59-60) points out, in the American Council for the Teaching of Foreign Languages' Proficiency Guidelines for speaking. The speech of n*ovice-low* learners is described as consisting of "isolated words and perhaps a few high frequency phrases"; *novice-mid* speech "continues to consist of isolated words and learned phrases"; and *novice-high* speech is "[a]ble to satisfy partially the requirements of basic communicative exchanges by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements" (ACTFL, 1985).

Importantly, however, the ACTFL guidelines appear to see formulas as entirely the preserve of the beginner: learned phrases do not feature in descriptors above the novice levels, suggesting that ACTFL views formulaic language as a temporary expedient, to be dropped once fuller linguistic competence is in place. Indeed, the phrasing of even the novice descriptors seems to imply a pejorative view of memorised phrases: that novice-mid speech "continues to consist" of such language seems to imply that this is an undesirable state of affairs (akin to the use of "isolated words") which will be overcome as learners develop, while the picture of novice-high speakers' "relying heavily on learned utterances" suggests that learned phrases are a crutch, rather than a mark of true competence. Progress is marked by "expanding [phrases] through simple recombination of their elements".

The view that formulaic language is merely an temporary tool for novice learners, to be replaced in time by a truly creative linguistic competence has been explicitly endorsed by some second language acquisition theorists (Krashen & Scarcella, 1978). However, others have allowed formulas a more central role in language pedagogy. We saw in Chapter 2 that in the early decades of the twentieth-century, Palmer was extending the definition of vocabulary to incorporate those "successions of words" that needed to be learnt "as an integral or independent entity, rather than by the process of piecing together their component parts" (1933, p. 4) and attempting to create listings of such items to form the basis of reference books and limited vocabularies. We have also seen that phraseological work in the Soviet Union from the 1970s onwards was motivated largely by second language learning needs (Cowie, 1998b, p. 209).

The most influential teaching approach of recent decades - communicative language teaching (CLT) – has been somewhat ambivalent over the role of formulaic language. CLT's starting premise that learners need to acquire what Hymes called *communicative competence* – i.e. mastery not only of what is grammatically possible, but also of what is *feasible* given our psycholinguistic limitations, of what is *appropriate* in a given social context, and of what is most likely to be attested (Hymes, 1972, pp. 281-286) – seems, as Widdowson (1989) has pointed out, well-matched with a focus on formulaicity. Moreover, some key texts within the CLT tradition have found a place for formulaic language. Van Ek and Alexander's influential *Threshold Level English*, a cornerstone of CLT, recognised that linguistic competence includes control not only of grammar and vocabulary, but also of some remembered utterances (1980, p. x), and their specification of the linguistic forms which learners are likely to need is replete with what many current-day linguists would call formulas. To take a typical example, the language function *expressing surprise* (1980, p. 47) lists the following linguistic realisations:

> This is a surprise!
> Fancy + V_{ing}…!
> How nice V_{to}…!
> What a surprise!
> It's surprising!

I'm surprised + *that*-clause

Enthusiasm for formulaicity is far from universal in CLT circles, however. It is noteworthy that Van Ek and Alexander's syllabus has been widely criticised for yielding only "situationally appropriate phrases" which are "no more interesting than the phrase books for tourists and businessmen that had been available since the Renaissance" (Yalden, 1987, p. 76). The chief focus of much work in CLT syllabus design has, rather, been on specifying the more abstract grammatical structures learners are thought to need to express key functions (e.g., Wilkins, 1976, pp. 66-68).

The last three decades have seen a marked increase in pedagogical interest in formulaic language, however. Motivated initially by Pawley & Syder's (1983) observations regarding the importance of memorized sequences in native speech, and later by an ever greater awareness of the types of patterning revealed by corpus research and by developing models of language acquisition, formulaic language in general – and high frequency collocation in particular – is being pushed towards the centre of the language teaching agenda. The recent literature has proposed three main rationales for teaching formulas. Formulaic language is held, first, to promote natural, nativelike language use; second, to increase fluency; and, third, to drive the acquisition of the language system. The next section will discuss each of these motivations in turn.

3.2 Rationales for teaching formulaic language

Formulaic language promotes natural language use

I have noted Hymes's (1972) argument that communicative competence entails knowing not just what it is possible to say, but also what is most likely actually to be said. Pawley and Syder (1983, p. 192) call this latter facet of competence "nativelike selection". Native speakers, they point out "do *not* exercise the creative power of syntactic rules to anything like their full extent". Indeed, "if they did so they would not be accepted as exhibiting nativelike control of the language":

> The fact is that only a small proportion of the total set of grammatical sentences are nativelike in form – in the sense of being readily acceptable to

> native informants as ordinary, natural forms of expression, in contrast to expressions that are grammatical but are judged to be 'unidiomatic', odd or 'foreignisms'. (Pawley & Syder, 1983, p. 193)

Choosing the most natural option from the wide range of grammatically possible sentences in any given situation requires, then, something more than a knowledge of syntax. Pawley and Syder argue that part of the extra knowledge required is a store of what they call *lexicalized sentence stems* (1983, p. 205). These are strings of language which are completely or partially pre-specified, recognized as standard expressions by the speech community, and used to denote standard concepts in that community. They include both idioms (e.g. *a stitch in time saves nine*), and more literal conventionalised expressions, such as:

> NP be-TENSE sorry to keep-TENSE you waiting
> (as in, "I'm sorry to have kept you waiting")
>
> Who (the EXPLET) do-pres NP_i think PRO_i be-PRES
> (as in, "Who the hell do you think you are?")
>
> P thinks nothing of V-ing
> (as in "Dave thinks nothing of walking 50 miles")

Pawley and Syder have disappointingly little to say about exactly how such formulas make language more nativelike, focusing mainly on their role in supporting fluency (see below). Indeed, they note that their role in supporting nativelike selection is likely to be "a limited one", since lexicalized sentences form "only a small subset" of nativelike sentences (1983, p. 214). However, they do emphasise that one marker of nativelike competence is to understand fully the restrictions on partially-specified lexicalized forms. A characteristic learner error, they maintain, is to assume that such expressions allow more variation than convention actually warrants (e.g. *You are pulling my legs*; *I intend to teach that rascal some good lessons he will never forget*) (1983, p. 215).

While Pawley and Syder believe that lexicalized sentence stems play a relatively small part in nativelike selection, Kjellmer (1990) – who looks at the broader concept of high frequency collocation – gives formulaic language a much more central role. He believes that collocation is ubiquitous in native language, arguing that "large parts" of native speakers' vocabulary is organised in terms of collocations, and claims that in producing discourse, speakers "very largely make use of chunks of prefabricated matter". The second language learner, in contrast, "having automated few collocations, continually has to create structures that he can only hope will be acceptable to native speakers". For this reason, in both speech and writing, "his output will often seem contrived and unacceptable to native ears" (1990, pp. 123-124). With this in mind, Kjellmer calls for "a new approach to the teaching and learning of foreign languages", in which emphasis is shifted "from individual words to the collocations in which they normally occur" (1990, p. 125). The creative competence emphasised by previous teaching approaches must, he implies, take a temporary back seat if learners are ever to achieve nativelike selection, since "[it] is only when the student has acquired a good command of a very considerable number of collocations that the creative element can be relied on to produce phrases that are acceptable and natural to the native speaker" (1990, p. 125).

The argument that learning formulaic language can contribute to nativelike selection is a persuasive one. However, a note of caution must be sounded. Kjellmer's point has often been taken to imply that the more learners use formulaic language, the more natural their production will be. Cortes (2004, p. 398), for example, notes that the "use of collocations and fixed expressions has been considered a marker of proficient language use", and approvingly quotes Haswell's (1991, p. 236) claim that "as writers mature they rely more and more on collocations". Similarly, in their studies of the development of collocational knowledge in non-native writers, both Nesselhauf (2005, pp. 234-236) and Kazsubski (2000, p. 33) assume that increased proficiency will correlate with increased use of conventional collocations. As we shall see in Chapter 5, however, this equation is probably oversimplistic. While we can expect learners' overall repertoire of formulaic language to increase over time, over-reliance on formulas may also – as the ACTFL guidelines discussed above describe – be a mark of non-nativeness. Similarly, as we shall see below, research into child language has suggested that language acquisition may involve a progression in which initially

memorized forms are gradually broken down as more creative linguistic competencies develop (Tomasello, 2003). The relationship between extent of formula use and nativelike selection is, therefore, likely to be rather more complicated than some researchers have assumed.

Formulaic language promotes fluency

We saw in Chapter 2 that many researchers have suggested that native use of formulaic language may be partly motivated by what Sinclair calls "a natural tendency to economy of effort" (1987, p. 320). Calling on memorised formulas is believed to be less cognitively demanding that constructing new utterances from scratch, and so it is thought that formulas may help speakers cope with the demands of real-time language production and comprehension while maintaining fluency (Coulmas, 1981; Kuiper, 2004; Pawley & Syder, 1983). If these researchers are right, then second language learners who do not have a stock of ready-made formulas to draw on are likely to have great difficulty in achieving fluent language use; either in production or in comprehension (Nattinger & DeCarrico, 1992, p. 159).

Theoretical support for the idea that formulaic language underwrites fluency is provided by psychological models of automaticity in language processing (see the reviews in De Keyser, 2001; Schmidt, 1992; Segalowitz, 2003; Segalowitz & Hulstijn, 2005). Though different models have disagreed both as to how automaticity should be defined and as to how it is achieved, most point to a role for formulaicity in achieving fast and efficient processing.

On Logan's *instance theory* (1988), the enactment of a skill involves a 'race' between an algorithm-based performance and a memory-based performance. An individual who is learning a skill will rely at first on the algorithm, but with each performance a memory will be stored of the action executed. As the stock of memories increases, retrieval of the information needed to perform the act will become gradually faster. Eventually, the individual will reach a point of mastery where memory retrieval is always faster than executing the skill by rule. At this stage, automatization is said to have taken place. Schmidt (1992, p. 371) notes that retrieval of past linguistic performances may include "phrases and both completely formulaic and partly open

clause structures", and that Logan's model therefore support the ideas that formulaic language is important in attaining fluency.

Other models suggest that *chunking* may be key to automatic skill performance (Anderson, 1983; Ellis, 2001; Newell, 1990). The notion of chunking was first introduced by Miller (1956) to explain why the span of human short-term memory remains at a more-or-less constant 7 items, regardless of the amount of information encoded by each item. The number of binary digits that can be handled is no greater than the number of decimal digits, of letters, or letters plus digits, or of monosyllabic words, in spite of the increasing information load carried by each of these item types. Miller argued that the capacity of short-term memory is not tied to the amount of information in a message, but to the number of chunks of information. By recoding more simple items (such as letters) into more complex chunks (such as words), we can massively increase the amount of information our memory can handle. Anderson (1983) and Newell (1990) have argued that chunking plays a key role in the automatisation of practised skills, while Ellis (2001) has suggested that the same principle might lie behind formulaic language. On Ellis's model, two or more words which frequently co-occur are recoded as a chunk and henceforth treated as a single entity. This process is recursive, with chunks themselves subsequently available for combination into still larger units, enabling language users to encode progressively greater amounts of information in short-term memory, so increasing the efficiency (and therefore the fluency) of communication (Ellis, 2001, pp. 38-40).

Attempts to demonstrate empirically the role of formulaic sequences in supporting fluency have been problematic because of the difficulty of distinguishing formulaic from non-formulaic language. However, a number of researchers have presented evidence that second language learners can use formulas in this way. Raupach (1984) describes the use of formulas in the speech of two German learners of French before and after a stay in France. He identifies a number of different types of formula at different levels of what construction grammarians would call complexity and specificity (see Section 2.2) and argues that increased control over a wider and more nativelike range of such formulas can act as a time-buying device, reducing the number of pauses and other hesitation phenomena. Towell et al (1996) also discuss the speech of two learners of French (this time with L1 English) before and after stays in

France, and argue that "memorized sentence builders" play a key role in extending the number of syllables learners can produce between pauses. Similarly, Wood (2006) distinguishes a number of ways in which Spanish, Chinese, and Japanese learners of English use formulas to extend linguistic runs between pauses. It is interesting to note that, in contrast to the ACTFL grading descriptors, which see learners as coming to rely less on formulas as their ability increases, both Raupach (1984, p. 134) and Wood (2006, p. 30) describe how development in fluency is marked by an increasingly sophisticated and nativelike use of formulaic constructions. In other words, more fluent language use is not marked by less (or more) use of formulas, but rather by a more nativelike command of a wider repertoire of formulas.

Formulaic language is the basis of acquisition

Introduction

The traditional focus on grammar as the prime object of language learning is largely motivated by the belief that it is knowledge of grammar which enables speakers to generate new sentences. As Wilkins puts it in *Notional Syllabuses*:

> grammar is the means through which linguistic creativity is ultimately achieved and an inadequate knowledge of the grammar would ultimately lead to a serious limitation on the capacity for communication (1976, p. 66).

From this perspective, studying 'phrasebooks' of formulaic language appears an unattractive long-term learning strategy. While memorised formulas can provide a quick route into fluent and nativelike speech, learners will ultimately need to tailor their utterances to new situations and to express ideas for which they do not have a formula. For this, control of a creative language system is required, and this means knowledge of the more abstract patterns of the language.

However, proponents of formula-based approaches to learning have argued that – perhaps paradoxically – it is precisely through the learning of specific formulas that mastery over the creative, abstract patterns of language is best achieved (Lewis, 1993, pp. 95-98; Nattinger & DeCarrico, 1992, pp. 114-116). This is probably the most contentious of the motivations for a formula-based approach to language teaching. It draws on models of first language acquisition which have argued that formulas play a

central role in the development of creative linguistic ability (Clark, 1974; Lieven & Tomasello, 2008; Peters, 1983; Pine & Lieven, 1993; Tomasello, 2003; Tomasello & Brooks, 1999). However, even researchers who broadly endorse these models with regard to child first language acquisition have raised question as to whether they apply equally to adult second language learners (Ellis, 2003; Wray, 2000). The present section will briefly describe the key points of the formula-based models of first language learning before turning to the question of whether adult L2 learning is likely to follow a similar route.

Formulas in first language learning

It has long been recognised that young children make use of memorised multi-word utterances (Brown & Hanlon, 1970). One reason for the prevalence of such language is probably to do with the perceptual nature of the learning task. Faced with an unbroken stream of speech, children are likely first to notice and remember the most salient units which are recurrently associated with obvious functions, and these will often be multi-word utterances (Peters, 1983, p. 5). Other reasons for formula use relate to processing limitations. Clark (1974, p. 2) has suggested that children may repeat chunks of language verbatim, without altering them to suit their contexts, in order to reduce the processing load of language production and reception. Tomasello and Brooks (1999, p. 166) speculate that formulas may be used because children can only attend to limited parts of the utterances they hear, or are only able to process one unit of language at a time.

While some researchers (Bates, Bretherton, & Snyder, 1988; Brown & Hanlon, 1970) consider early multi-word utterances to be a short-term communication tool, unrelated to the larger acquisition process, others have argued that formulas play a central role in the development of a mature language system. Increasingly sophisticated versions of this model have been developed over the last three decades, but the basic idea has remained the same. That is, that child language "becomes creative through the gradual analysis of the internal structure of sequences which begin as prepackaged routines" (Clark, 1974, p. 9). Researchers describe child language as developing gradually from an initial repertoire of concrete memorised utterances, through stages of gradually increasing complexity and abstraction, to adult-like mastery (Peters, 1983; Pine & Lieven, 1993; Tomasello, 2003; Tomasello & Brooks, 1999). The dominant

framework within which this formula-based learning process is being studied today is that of usage-based construction grammar (Barlow & Kemmer, 2000; Lieven & Tomasello, 2008), as described in Section 2.3.

Space limitations do not allow a full discussion of usage-based acquisition models here. However, a number of key points should be highlighted. Firstly, on usage-based models, both the route of acquisition and the structure of the mature language system are seen as strongly dependent on the nature of the input children receive. This stands in contrast to the Chomskyan view that language is largely innately-specified, and that experience is merely a mechanism for setting parameters within these pre-specified systems (Barlow & Kemmer, 2000, p. xi). The relationship of input to learning is complex because it involves the interaction of many different variables. However, key factors are the token and type frequencies of constructions (the former referring to the frequency with which a particular concrete item appears; the latter to the number of different items which are met in a variable slot), the consistency of mapping between form and function, and the complexity of constructions (defined in terms of factors such as the number of parts a construction has, whether its functional cues are local or distributed, and how it relates to already-known constructions) (Lieven & Tomasello, 2008, pp. 172-183). While brain structures are important in shaping the way language is learned (unlike Chomsky, usage-based models see the cognitive limitations of language users as playing a key role in structuring the language system), it is not thought necessary to posit innately-specified linguistic structures. Rather, structure emerges from a complex interaction between mind and environment (Elman et al., 1998; MacWhinney, 1999).

Secondly, the formula-based route of acquisition is held to be at least partly responsible for the formulaicity of adult speech. In the simplest case, set phrases which are adopted during childhood may continue to be stored in memory even after the language system is capable of analysing them fully (Peters, 1983, p. 71). A more subtle source of adult formulaicity is found in the way schematic constructions are abstracted only gradually from concrete instances, such that grammatical patterns are found initially used only with certain lexical items (Tomasello, 2003, pp. 114-122). It has been suggested that the adult constructions which emerge from this process never become entirely abstract; that is, that the adult's abstract syntactic patterns are always

"tied to, i.e. activated in concert with, specific instances of those patterns" (Barlow & Kemmer, 2000, p. ix). This may be the root of the preference of certain grammatical forms for certain lexis which is described by Pattern Grammar (see Section 2.3).

Thirdly, usage-based models see language acquisition as a product of the same psychological mechanisms which operate in other – i.e., non-linguistic - types of learning. This contrasts with the Chomskyan view of a discreet language acquisition device, working on different principles from the rest of the mental system and isolated from general cognitive limitations (Lieven & Tomasello, 2008, p. 168). Researchers in the usage-based tradition see acquisition as the product of such general mechanisms as association (Ellis, 2001, p. 42), categorization, analogy formation, and functional distribution analysis (Tomasello, 2003, pp. 122-124, 163-173)

Finally, many researchers have emphasised the extent of variation between children in the degree to which they focus on learning phrases or individual words. Peters (1977) identified separate *gestalt* and *analytic* learning strategies and recorded:

> a continuum of children, varying from those who are very Analytic right from the beginning, through those who use mixes of Analytic and Gestalt speech in varying proportions, to those who may start out with a completely Gestalt approach and have to convert slowly and painfully to an Analytic approach (1977, pp. 570-571).

This picture of two distinct strategies adopted differentially by different children, is corroborated by Lieven et al (1992), who find a significant negative correlation between nouns and frozen phrases in children's early vocabularies, and show that a preference for one or the other is constant through the early stages of learning, suggesting that they constitute two different strands of vocabulary development. Pine and Lieven (1993) argue that these two 'strands' correspond to different 'routes' into the language system. In a longitudinal study of seven children, they find a strong relationship between the proportion of frozen phrases in a child's first 100 words and the nature of their early productive patterns: while the early constructed utterances of 'non-phrasal' children (those with a low proportion of frozen phrases) tend to be built out of words already present as individual items in their single-word vocabularies,

'phrasal' children's early constructions can be traced back to originally unanalyzed 'frozen phrases'. According to the authors, the 'phrasal' strategy is the more common of the two. They find that relatively few patterns in early language are built up of two pre-existing vocabulary items, with 66% of constructions overall being traceable back to frozen phrases, and even the most non-phrasal of the children generating 40% of her constructions in this way (1993, p. 567).

While many writers accept the existence of different strategy preferences, there is disagreement on the long-term effects of such variation. Nelson, for example, claims that early differences of approach may lead to the establishment of "different rule systems" (1981, p. 172), and Peters (1977, p. 571) suggests that they may be linked to different approaches to second language learning later in life. Pine and Lieven, on the other hand, claim that strategies converge as learning progresses and find that "neither approach appeared to confer any relative long-term advantage on the children adopting it" (1993, p. 552).

Peters suggests four factors which may combine to influence a child's choice of strategy. First is their communicative needs. Noting that the child in her study tended to use analytic utterances "in referential contexts, such as naming pictures in a book, whereas Gestalt speech tended to be used in more social contexts, such as playing with his brother, or in commenting about objects rather than naming them", she suggests that referential communicative needs lead children to extract word-length labels for things, while the needs of social interaction "will drive them to extract from the speech stream the necessary language for conducting such pragmatic interactions", primarily "multi-word formulas or sentences" (1983, pp. 22-23). A second variable is the nature of the input stream (1983, pp. 23-28). Peters notes that differences have been found between the types of speech which adults of different cultures and social classes direct towards their children, and speculates that this must have an effect on the extraction task. The internal structure of families may also be significant in this respect. A first-born child experiences rather different kinds of input from their younger siblings: while the former tend to receive a great deal of exclusive carer attention, a situation fostering Referential input, the latter have, in the interactions between their older sibling and parent, a model of more Expressive, interactive language, such as directives and requests. It may be a combination of these factors

which biases children towards a certain strategy. Peters cites data demonstrating only a nonsignificant tendency for first-born children to be Referential and second-born children Expressive, but when parents' educational background was taken into account, the picture is less ambiguous: in Peters' data, *all* first-born children of parents with a least a college education were found to be Referential (1983, p. 24). The third variable discussed is that of the "culturally determined set of expectations regarding appropriate language use and acceptable language style" (1983, p. 28). Such expectations, Peters notes, may "form the unconscious basis of feedback from caretakers to children concerning whether their early vocalizations 'count' as language" (1983, p. 28). This sort of feedback may influence both the types of units which a child extracts and their perceptions of appropriate occasions of use. Finally, acknowledging that a child's environment cannot fully predict their course of development, Peters proposes that individual differences of personality and neurology may have a significant impact on extraction strategies, on propensity to imitate speech, and on the process of segmentation (1983, p. 28). Though Pine and Lieven discourage the idea that variation might be the result of any such inherent differences between children, their view appears not to have been widely accepted, and subsequent writers have continued to emphasise the role of individual differences. Wray, for example, writes that:

> A child's preference for a referential and analytic, or else for an expressive and holistic style might be determined by internal factors relating to personality, neural organization or early nonlinguistic experience (2002, p. 114)

Commenting on the finding that some children tend to use constant + variable structures while others use variable + variable constructions, Peters suggests that the former style may be preferred by children who are eager to talk, and who accordingly rely on formulas rather than waiting on a complete analysis, while the latter style may be characteristic of children who are productively cautious and who carry out a great deal of analysis before producing a lot of speech (1983, p. 70). She also speculates that analytic strategies may be associated with the dominant (i.e. for most right-handed people, the left) hemisphere and gestalt strategies with the minor (right) hemisphere (1977, pp. 571-572). This speculation has been developed by a number of

other theorists, as we shall see in Chapter 4 (e.g. Van Lancker-Sidtis, 2004; Wray, 1992).

Differences between child first and adult second language learners

Some proponents of formula-based approaches to second language learning have proposed that formulas might play a role in adult second language acquisition similar to that suggested for first language learning. On this view, it is argued that learners who study formulas in an appropriate way may be able to extrapolate a creative and nativelike language system through processes of abstraction paralleling those discussed above (Lewis, 1993, pp. 95-98; Nattinger & DeCarrico, 1992, pp. 114-116). However, other researchers have warned against accepting too readily the idea that adult second language learning will mirror child first language learning (Ellis, 2003; Wray, 2000).

There are a number of *prima facie* reasons for thinking that formulaic language may not play the same role in adult second language learning as has been hypothesised for first language learning. Firstly, the adult L2 learner is more cognitively mature than the first language learner. In L1 learning, language develops in tandem with the general cognitive mechanisms on which it depends. Ellis notes that this distinguishes the child from adult learners in two important ways. Firstly, where the child's world knowledge develops simultaneously with their linguistic knowledge, the adult builds their linguistic knowledge on pre-existing concepts. Secondly, the adult has from the start important analytical competences which are lacking in the infant. As Ellis puts it, "adult learners have sophisticated formal operational means of thinking and can treat language as an object of explicit learning, that is, of conscious problem-solving and deduction, to a much greater extent than can children" (2003, p. 72).

These cognitive issues interact with elements of the adult learner's social situation and needs. L2 learners with mature knowledge systems will want, from the beginning of the acquisition process, to express meanings which L1 learners are not able to conceptualise until later. Particularly important for formulaic language may be the *diversity* of forms which adults wish to use with particular constructions (a tendency no doubt compounded by pedagogical practices, which tend to encourage learners to practise new constructions with a broad range of lexis). We have already noted the

importance to child acquisition of item-based learning, whereby structures become fully generalised only gradually, and are initially tied to a small range of individual instantiations. Similarly, Goldberg claims that appropriate category formation is facilitated if input is 'skewed' – i.e. if it consists primarily of very few, very high frequency, items, rather than a representative sample of possibilities (2006, p. 84). The adult need to use constructions with a range of forms from the very beginning may well, therefore, subvert the natural course of grammar learning and category formation. More prosaically, the cognitive (and physical) maturity of older learners may lessen adult use of formulas by weakening the imperative to communicate. Where the infant may be driven to use formulas to cope with such urgent communicative needs as getting fed, the adult can circumvent communication, either by adjusting their needs to avoid linguistically difficult situations, or by meeting their needs through non-linguistic means (Wray, 2002, p. 175). Furthermore, situational factors may encourage the adult to employ conscious problem-solving abilities to which children do not have access. Cultural norms in general, and classroom practices in particular, often encourage a focus on the explicit analysis of input. It is also possible that literacy, which Wray claims "effects a major transition from working with larger complex units to smaller ones" (2002, p. 137), may lead adult learners to focus on individual words more than do children (2002, p. 194).

As well as a mature cognitive system, adult L2 learners also have a pre-existing first language in place at the start of their acquisition process. This means that, where the first language learner must build their knowledge of productive syntactic categories out of lexically specific patterns, second language learners have, as Ellis points out, "already acquired knowledge of these categories and their lexical membership for L1, and this knowledge may guide creative combinations in their L2 interlanguage to variously good and bad effects" (2003, p. 72). As with cognitive maturity, the existence of an L1 also affects the communicative needs of the learner. Just as the adult L2 learner has recourse to non-linguistic means of meeting their needs, they may also (depending on the situation in which they are operating) have the option of bringing in their first language when the communicative going gets tough (Wray: 2002:147).

A further relevant difference between first and second language acquisition concerns the nature of the input to which the two types of learner are exposed. Ellis suggests that, whereas an important feature of L1 exposure is the tendency for caregivers to "naturally scaffold development", the environment of the language classroom "can distort the patterns of exposure, of function, of medium, and of social interaction" (2003, p. 72). Furthermore, it may be that language directed to second language learners is stripped of formulaicity. Irujo (1986) has claimed that learner-directed language frequently omits idioms, while Wray and Grace have suggested that native-non-native encounters may foster "conscious strategies on the parts of both learner and native speaker to effect the regularisation of irregularities, the rationalisation of partial patterns, the re-expression of impenetrable conventionalised expressions" (2007, p. 557).

Research into the role of formulas in adult second language acquisition

While there are many *a priori* reasons for thinking that formulas may not operate in the same way in adult L2 and child L1 acquisition, there are not yet empirical data of sufficient richness to draw a strong conclusion either way (Ellis, 2003, p. 74). In their reviews of the literature, Wray (2002) and Yorio (1989) conclude that there is little evidence that adults who learn a second language naturalistically analyse formulaic language effectively. However, some researchers have suggested that classroom-learners may make use of such a strategy.

Bolander (1989), in a study of the spontaneous and elicited speech of 60 adult learners of Swedish, suggests that memorized sequences can play a role in the acquisition of word order rules. She finds that correct application of a number of such rules was associated with sentences using high frequency lexis or 'stereotyped' phrases. This suggests that the learning of these rules may have been tied to particular words or formulaic instantiations. Bolander also reports that overgeneralizations of word order rules (i.e., cases of inversion in inappropriate contexts) often involved particular verb-subject combinations which were frequently and correctly inverted, suggesting that they were produced as memorized formulas. While these findings are suggestive, however, they suffer for lack of quantitative data. In particular, no actual figures are given for the relative frequencies of occurrence and correct usage of different lexical items, and no real indication is given of the total range of items with which the various

constructions are used. Moreover, while her data appear to show that early use of rules is tied to particular items, this remains compatible with the view that such usage is a communicative expedient and that the mature development of these rules proceeds independently of formulas.

Myles et al (1998) claim to find evidence that formulas actually feed into the acquisition process of tutored learners. They identified three locutions in the early productions of 16 (L1 English) school learners of French (aged 11 or 12 at the start of the study) which appeared to be formulaic (*j'aime, j'adore, j'habite* – 'I like', 'I love', 'I live'), and traced the use of these chunks and their constituent parts over two years. While there was much variation between individuals, the authors assert that all but one of these learners employed the formulas as a communicative strategy early on, and that as their communicative needs moved beyond what the chunks could provide (in particular, as they started to make reference to the likings, lovings and living arrangements of third persons), these formulas were gradually 'broken down'. At first, the breakdown involved simply adding a third person referent to an unaltered chunk (as in *j'aime le sp- elle j'aime le sport* – 'I like sp- she I like sport'); full segmentation occurring only later (*il…j'ador- il adore la livre?* – 'he…I lov- does he love the book?'). In a second study using the same data set, Myles et al (1999) record a similar course of development for the question formula *comment t'appelles-tu?* ('what is your name?') as learners attempt to establish a way of asking the question for third-person referents. While there is again much variation between individuals, with some learners never proceeding beyond simply using the unanalysed formula for all referents, the authors nevertheless claim to identify a "common general route" of progression (1999, p. 67) similar to that seen for the declarative chunks:, i.e.:

1. Chunk inappropriately used, overextended (e.g. *comment t'appelles tu?*);
2. Chunk overextended, but with lexical NP tagged on to clarify reference (*comment t'appelles-tu le garçon?*);
3. Chunk starting to break down: e.g. subject omitted or replaced by a NP (*comment t'appelles (la fille)*)
4. Reflexive pronoun changes to *s'* – apparently through analogy with *il/elle s'appelle* (*comment s'appelle?; comment s'appelle…garç-un garcon?*)
5. Third-person pronoun used (*comment s'appelle-t-il*)

For successful learners, the authors conclude, rote-learned chunks are a "linguistic database" which they use "as a springboard for creative construction". Rather than dropping chunks once the productive system starts to come into force, learners "seem willing to keep working at them over sustained periods of time, presumably until they merge entirely with an evolving grammatical competence" (1999, p. 76).

Myles et al's data are intriguing; however, some important questions are left unanswered. The authors report that only one learner ever used the correct third-person question form (*comment s'appelle-t-il*), and that this individual used the form more-or-less consistently from the very beginning of the study. In other words, no learner ever actually achieved the target via the course of progression outlined above. It is not clear, then, whether it is actually possible to attain target-like performance via this route. More fundamentally, we need to ask whether the emergence of third-person forms in these data really demonstrates 'segmentation' in action. The authors record that at least some learners were exposed to the correct third-person form in their classroom input. One learner evidently picked this up, and was able to use it from the beginning. That others also at least partially learned the form is indicated by the fact that even the weakest of the learners whose progress is detailed invokes garbled forms of it (*comment s'appelle-tu?; comment t'appell-euh s- s- comment s'appelle-tu?* (1999, p. 64)) early on in the study. We might speculate here that the third-person form was not picked up and used as readily as the second-person because it did not occur in class as frequently, and so either did not become entrenched, or else had its recall seriously inhibited by its over-learned semantic and phonological neighbour. If the latter is the case, it may be that examples such as *comment t'appell-euh s- s- comment s'appelle-tu?* indicate not segmentation, but a struggle to overcome this inhibition. Similarly, with respect to the declarative formulas discussed above, examples like *il..j'ador – il adore la livre* ('he..I lov- does he love the book?') could well be explained in terms of over-learned sequences intruding on, rather than 'feeding into', word-based language production. While some attempts at asking the third-person question clearly involve manipulation of the second-person formula then, it should also be considered whether some such productions may in fact be examples of something akin to the 'blending' seen in errors of lexical retrieval (Aitchison, 1987), which they appear to resemble. This would seem to be the most natural interpretation

of the fact that the one learner who successfully produced the target form, having demonstrated its correct use in early rounds, later came up with such mixed forms as *comment t'appelle't-il*.

Wray (2004) also claims to find evidence for the spontaneous break-down of formulaic chunks. She presents a case study of a woman learning Welsh intensively for one week as part of a TV programme 'challenge'. The learner's task was to learn enough Welsh to present a brief cooking programme in the language. Learning took place almost entirely through the memorisation of formulaic sequences tailored to her needs for the programme. Wray remarks that, though the learner's greatest chance of success lay in reproducing her rote-learned formulas exactly, the formulas were in fact "subject to accidental editing" (2004, p. 265). In some cases, this involved the substitution of a word with a Welsh synonym; in others, a word was substituted by its English translation. Such interference was, Wray argues, "gratuitous", and suggests a process of "analytic activity interfering with which ought to have been a very straightforward process of faithful reproduction" (2004, p. 265). Further evidence of segmentation is found in the learner's coming to omit the obligatory unstressed grammatical particle *yn*, having initially produced the form accurately. Wray reasons that the learner must have noticed the form, having used it early on, but that a subsequent conscious or unconscious analysis of the relevant formula may have taken place. The *yn* particle would be likely to be among those parts of the formula with no semantic role assigned to it by such an analysis; as an unstressed clitic, it would then be susceptible to omission (2004, p. 266). Wray speculates that this "underlying propensity to engage in analysis" may mean that the repeated use of formulaic utterances "might ultimately bootstrap the learner into a kind of extrapolated knowledge that was both flexible and rather more nativelike than usual, being based, as the young child's is, exclusively on the delivery of real language in use." (2004, p. 267). However, since her analysis is based on a rather anecdotal report of the productions of single learner, this conclusion remains highly speculative.

In sum, while it is not possible to rule out the possibility that some adult second language learners might break into the creative language system at least partially through item-based learning and the spontaneous segmentation of memorised sequences, strong evidence for such a process has yet to be provided. Bolander's data

suggest that early correct production of rules may be based on formulas, but do not provide evidence that such formulas play any active role in the rule-learning process. Myles and her colleagues claim to observe segmentation in action, but it is not obvious that the language they report must be interpreted as segmentation, rather than some other process, such as the blending of chunks. It is also not clear that the formulas reported ever develop into a fully functioning creative grammar. Finally, while Wray's case study does appear to show some spontaneous analysis of formulas taking place, her findings stand in need of replication with a larger sample of learners and more systematic analysis.

Conclusions: formulas in language acquisition

An important element in the argument for a formula-based approach to language learning is the idea that formulas can feed into the development of a more creative language system. If this is not the case, then focusing too heavily on formulaic language may provide learners with a useful mental phrasebook of utterances for specific situations, but leave them unable to adapt their language to new situations or to express more novel ideas. While there is good evidence that a process of this sort operates for children learning their first language, there are also good reasons to believe that the case may be different for adult L2 learners. Until more empirical data are available, we must therefore remain cautious about the claims made for this route of acquisition. It is also important to bear in mind the large individual differences found between children in their use of primarily word-based or primarily phrase-based routes into language. While it is not clear if such biases are mainly due to long-term characteristics of the learners, to short-term preferences, or to features of the environment, it is possible that a one-size-fits-all approach will be inappropriate for second language learners.

Summary and conclusions: why teach formulaic language?

We have seen three main motivations for giving formulaic language a central role in the second language syllabus. Knowledge of formulas is important for achieving nativelike selection and nativelike fluency, and is held by some to be a central component in the acquisition of a creative language system. Though not all of the claims made for formulaic language are undisputed (in particular, the case for its involvement in acquisition is widely questioned), taken together these rationales make

a strong case for at least some focus on formulas in the second language learning syllabus.

Doubts can still be raised, however, regarding the extent to which this conclusion applies to the main items of interest in the present thesis – high frequency collocations. The formulas which have been discussed in relation to the three rationales above are probably best united under Wray's (2002, p. 9) definition of formulas as sequences which are "stored and retrieved whole from memory". We have, in other words, strong arguments for a pedagogical focus on sequences which are independently represented in the mind. However, as will be discussed below (see Section 4.1), it is not clear to what extent the class of high frequency collocations overlaps with the class of independently represented formulas. One of the aims of this thesis (pursued in Chapter 4) will be to evaluate the relevance of high frequency collocations to second language learning by examining this overlap in more detail.

3.3 How collocations are learned

Introduction

The discussion so far has focused on reasons *why* second language learners should pay attention to high frequency collocations. If we are to integrate collocation into the language syllabus in a principled way, however, we also need to understand *how* collocations are acquired and the types difficulties collocation learning might present to second language learners. The present section provides an introduction to these issues. It will describe Ellis's model of L1 collocation learning and discuss why some applied linguists have considered collocation learning to be especially problematic for adult L2 learners. The ideas raised here will form a theoretical backdrop to the empirical discussion of collocation acquisition in Chapter 5.

A model of L1 collocation learning

Ellis (2001) has claimed that, for first language learners, the acquisition of collocations involves an implicit process of 'chunk' formation (see Section 3.2) driven by a principle of associative learning which he calls the 'Law of Contiguity'. This is the rule that, "[o]bjects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be

thought of also" (James, 1890, quoted in Ellis, 2001, p. 42). Under the influence of this law, the frequent co-occurrence of two words in linguistic input (and hence in short-term memory) will lead to their becoming associated in long-term memory, and consolidated into chunks. In this way, long-term memory becomes 'tuned' to frequent collocations, such that when the same items are again encountered in subsequent input they are perceived by the learner as units, and so further reinforced as long-term representations.

The chunking of frequently co-occurring forms proceeds, Ellis maintains, through implicit processes; i.e. without the learner's conscious attention. It provides speakers with a vast amount of knowledge about the transition probabilities of sequences in their language; knowledge of which they may not be consciously aware, but which is evidenced by their performance in language use and in psycholinguistic experiments (2002b). However, Ellis is careful to point out that, while such processes constitute a large part of collocation learning, there is more to collocation than purely formal associations, and more to chunking than implicit learning: meaning also plays an important role. Sound sequences which are regularly associated with a useful communicative function will be more salient to the learner, and so more likely to be learnt than those which are not (2001, p. 41). Whereas the formation of formal associations merely requires the learner to match sound with sound (in more technical terms, to create links within a single cognitive 'modality'), the mapping of form to function requires that sound be matched with other aspects of experience (they must establish 'cross-modal' links). Forming such links requires, Ellis claims, a conscious focus of attention (2005). Consciousness both brings together the input of different cognitive modalities and provides the necessary conceptual structuring to determine to what aspect of a particular experience a sound is referring. Whereas implicit association formation is a slow cumulative process, following a power law of practice, conscious learning can be instantaneous. This is one reason why knowledge of collocation is not entirely determined by input frequencies (Ellis & Larsen-Freeman, 2006). Once an association is consciously made, however, the resultant chunk is itself subject to implicit tallying processes and so open to frequency effects (Ellis, 2005).

Adult L2 learners' difficulties with collocation learning

The small body of literature which has examined the acquisition of collocations by adult second language learners abounds with claims that collocations pose special difficulties for learners (e.g., Bahns & Eldaw, 1993; Farghal & Obeidat, 1995; Granger, 1998). A typical conclusion is that of Bahns and Eldaw, who find that "EFL learners' knowledge of general vocabulary far outstrips their knowledge of collocations" (1993, p. 108). On reviewing the literature on adult L2 formulaic language learning, Wray similarly finds that, despite picking up formulaic sequences with "apparent ease" in the early stages of learning, "by the time the learner has achieved a reasonable command of the L2 lexicon and grammar, the formulaic sequences appear to be lagging behind" (2002, p. 182).

The tacit assumption here that learners' levels of achievement in word knowledge, grammatical knowledge, and collocation knowledge can be meaningfully compared is perhaps a dubious one: it is hard to see how we could specify a level of collocational knowledge that would be properly 'on a par with' (rather than 'lagging behind') mastery of the third conditional. More important, however, is Wray's inference that adult L2 learners' knowledge of formulaic language is so weak that we must assume that they do not normally acquire collocations from the input they receive (2002, pp. 206-209). On the basis of this claim, Wray proposes a model according to which, when the adult learner is exposed to language input, they primarily notice and remember not (as L1 learners are thought to do) meaningful chunks of language, but rather individual words. To take Wray's own example, whereas a first language learner, on encountering a collocation such as *major catastrophe*, would note the string as a single sequence and remember it as the idiomatic way to refer to 'large disasters', the adult learner would instead "break it down into a word meaning 'big' and a word meaning 'disaster' and store the words separately, without any information about the fact they went together". If called on to talk about major catastrophes in the future, "they would have no memory of *major catastrophe* as the pair originally encountered, and any pairing of words with the right meaning would seem equally possible" (2002, p. 209).

This is not to say that adult learners do not adopt any formulaic language at all. Wray allows the possibility that, for advanced learners, there may be "some means of

building up the store of nativelike formulaic sequences post hoc, probably by residing and fully interacting for some time in the L2 environment" (2002, p. 210). However, whereas L1 collocations are "fully formulaic pairings which have *become loosened*" as learners start to meet their constituent parts in a wider range of contexts, the post-hoc approach of adult L2 learners means that their collocations are "separate items which *become paired*". Because of this, they do not usually establish the appropriate "strength of association" between words (Wray, 2002, p. 211 original emphasis). In the terms of Ellis's L1 model, because adult L2 learners consciously pair up collocates, rather than extracting them from an implicit tallying of their co-occurrence frequencies in input, they never effectively establish the appropriate transition probabilities between words.

Wray suggests that this fundamental difference between child L1 learners and adult L2 learners comes about through a convergence of social and cognitive factors. On the social side, adult learners (especially those in a classroom environment) rarely have the pressing need to communicate which drives L1 learners to memorise helpful communicative sequences. Indeed, since in many cases the surrounding social pressure may be largely from the L1, rather than the L2 community, it may actively discourage the adoption of native-like formulas. These effects will be further compounded by traditional classroom teaching methods, which often focus on grammatical form and on the introduction of a wide range of new words. On the cognitive side, the mature mental faculties of adult learners – and, in particular, the fact that they are likely to be literate, and so aware of the word as a basic unit of language - will mean that they are likely to feel uncomfortable not knowing how sequences break down into their component words (2002, pp. 205-206). While it is possible to overcome these influences (so that learners do occasionally achieve full nativelike competence), this will be rare, given the "great many obstacles which their social and intellectual experience and their learning situations will set up to prevent it" (2002, p. 213).

If this model is right, it is of considerable theoretical and practical interest. Ellis claims that the associative learning which is demonstrated in the chunking of collocations is no mere linguistic side-show, but rather a central process in the formula-based process of language acquisition. If this mechanism is not normally

employed by adults, this may go much of the way to explaining the significant differences between first and second language learners across the linguistic spectrum. On the practical side, it will influence how collocation teaching should be approached. Schmitt has suggested that the heavily contextualized nature of collocations means that learners should be encouraged to acquire them implicitly, through "massive exposure to the L2" (Schmitt, forthcoming). However, if Wray's model is right, such implicit learning may be blocked. It may be necessary instead to help learners to build up their collocational associations through a process of proceduralization involving the automatic planning and assembly of utterances (Wray, 2002, pp. 201, 211). Given these implications, evaluating Wray's model is a matter of central importance for collocation teaching. Chapter 5 will present such an evaluation.

3.4 Summary and conclusions: formulas in second language learning

This chapter has provided a brief and selective overview of the place of formulaic language in adult second language learning. We have seen that, while formulaic language has always been of some interest to teachers and learners, the last three decades have witnessed a far more concerted focus on formulaicity, chiefly motivated by the ideas that formulas are essential to attaining nativelike fluency and selection and that they may play a key role in the acquisition process. We have also seen that the way in which adult L2 learners approach collocation learning is thought by some to be fundamentally different from the approach of child L1 learners, and that this may have far-reaching implications for language learning in general.

The remainder of this thesis will examine in more detail some of the issues that remain outstanding from the preceding discussion. The first relates to the psychological reality of high frequency collocations. I have noted that an essential step in the argument for a pedagogical focus on high frequency collocations is the assumption that such collocations are independently represented in the minds of speakers. Chapter 4 will aim to assess this assumption. It will review the existing evidence and present two sets of studies investigating the relationship between frequency of collocation in a corpus and mental representations. The second concerns the way in which collocations are learnt. If we do wish to teach collocations, the

methods that we use need to be informed by an assessment of the effectiveness of different kinds of input. Accordingly, Chapter 5 will evaluate and develop Wray's contention that adult L2 learners do not tend to acquire the collocations they meet. It will review the literature on formulaic language learning and describe two studies which aim to test Wray's model directly. The third issue concerns a practical consequence of the other two: if high frequency collocations are indeed good targets for learning, and if learners cannot be relied on to acquire them naturally, then it is incumbent on teachers to build collocations into their syllabi. This raises the difficult issue of how collocations can be selected for teaching. Chapter 6 will examine this problem and attempt to construct an inventory of target collocations for one particular set of learners – students of English for academic purposes.

Chapter 4
Are high frequency collocations 'psychologically real?'

4.1 Introduction

We saw in Chapter 3 that formulaic sequences are thought by many applied linguists to be important targets for second language learning. Formulas are believed to be important contributors to the idiomaticity and fluency of nativelike language and – more contentiously – to be key to the language acquisition process. It is important to note that these ideas rest on a conception of formulas as linguistic sequences which are somehow individually represented in the mental systems of competent language users. If this were not the case – if, in other words, formulas were not things which native speakers 'know' – then it would be difficult to argue that learners need to learn them. However, while it seems implausible that some types of formula (idioms or clichés, for example) are not individually represented in this way, some researchers have doubted whether the same applies to the items which are the central concern of the present thesis – high frequency collocations.

At the root of these concerns is the problem that the psycholinguistic models of collocation put forward by corpus linguists (Sinclair and Hoey being the prime examples) have been based entirely on descriptions of text, with little or no reference to psycholinguistic research. Such linguists are guilty, in other words, of what Lamb calls *introjection* - the direct ascription of features of the external world (in this case, features of text as viewed in large-scale corpora) into the mind, without further consideration of the facts of processing or biology (2000, p. 96). Introjection is fallacious because it is not clear, without further support, whether patterns found in language are a product of the mental linguistic system or of something else. This point has been put forcefully by Herbst (1996), who observes that the frequent collocations found in corpora may simply reflect real-world coincidences. *Dark night*, he comments, is a significant collocation "because nights tend to be dark and not bright". On this view, "the fact that certain words tend to co-occur must be attributed to certain

facts of the world – together with the way this world is conceptualised in language" (1996, p. 384). Bley-Vroman makes similar comments with regard to the collocation *profound ignorance*, which is, he claims, a product of "human cognition and the use of language to express meaning, rather than of calculating word transition probabilities based on the analysis of a corpus". On this view, recurrent patterns are merely a product of the use of language to express intentions in context, and do not have strong direct explanatory force (Bley-Vroman, 2002, p. 210). Newmeyer memorably sums up this position with the comment that frequency-based analysis "is no more defensible as an approach to language and the mind than would be a theory of vision that tries to tell us what we are likely to look at" (2003, p. 697).

Sinclair acknowledges this problem in his original explication of the idiom principle, where he concedes that some collocational regularity can be explained by "the nature of the world around us", and by "[t]hings which appear physically together", "concepts in the same philosophical area", and "organising features such as contrasts or series". Moreover, he notes, register exerts a constraining influence which limits the options open to a speaker. However, even given all of these influences, Sinclair maintains, "there is still far too much opportunity for choice". The idiom principle is therefore held to be necessary to account for the full extent of "unrandomness" in the system (1987, p. 320). This may be right, but it requires further argument, which Sinclair does not provide. As we shall see below (Section 4.3), the use of statistics to determine how much "unrandomness" a corpus exhibits is a very approximate science indeed. It is certainly not possible in our present state of knowledge to assert on statistical grounds what proportion of any patterning found might be explained by each of the factors Sinclair mentions, or to determine how much is 'left over' once these have been taken into account. One line of defence open to Sinclair and Hoey would be to point out that models of automaticity (see Section 3.2) suggest that, even if collocations originally attain their high frequency of occurrence by means of other types of regularity, their regular occurrence would subsequently lead to a process of chunking. However, this argument, again, stands in need of direct psycholinguistic confirmation.

A further problem with the putative link from corpus to mind is that the large corpora from which data on collocations are usually drawn cannot be said to represent the

linguistic experience of any individual speaker. We saw in Section 2.2 that the Firthian approach to linguistics was developed to study language as a social, rather than a psychological, phenomenon. Accordingly, while corpus-based studies can give us a fair idea of, say, what words are collocations in the English language (and so are ideal for such applications as dictionary writing), we are not able to say what words are collocations for individual speakers. A corpus of the production of a particular individual might give us some approximation to what we are after here, but the relative infrequency of even common collocations (in comparison to individual words) means that their study requires very large corpora, of the sort which few if any individual language users are likely to produce within a sufficiently short span of time for the sample to count as a synchronic snapshot of their language use.

Hoey is well aware of this problem. Indeed, he takes the strong position that "the personal 'corpus' that provides a language user with their lexical primings is by definition irretrievable, unstudiable and unique". However, he reasons that though corpora "cannot tell us what primings are present for any language user", they can at least "indicate the kinds of data a language user might encounter in the course of being primed". He sees corpora, then, "as a kind of laboratory in which we can test for the validity of claims made about priming" (Hoey, 2005, p. 14). It is possible to dispute Hoey's assumption that the patterning found in large-scale corpora will be similar in type to the patterning found in a personal corpus: the latter will include a much narrower range of text types, speakers, and topics, and will incorporate the internal monologues in which speakers engage, the nature of which is extremely hard to determine, and it is conceivable that such factors will lead to the emergence of quite different types of patterning. However, even if we allow Hoey's claim that corpora enable us to test the validity of claims about how language works in general, this would not be sufficient for our present purposes. What teachers want to know is whether their learners need to acquire the specific high frequency collocations found in corpora. Clearly, the set of collocations in any given corpus and the sets of collocations known to each individual speaker will be somewhat different. The key question is how much remains constant. If there is a strong overlap between the knowledge of large numbers of speakers and the corpus, then what is found in the latter is likely to be worth learning. If the knowledge of different individuals and of the corpus are extremely diverse, however, they probably are not.

In sum, high frequencies of co-occurrence in a corpus have been hypothesized to indicate that collocations are represented in the mental systems of competent language users, and so that word pairs identified in this way may be good targets for second language learning. However, the non-randomness introduced by factors other than collocation, and the gap between what is in any given corpus and what any given individual experiences, mean that the link from what is in a corpus to what individuals know- and so to what learners ought to learn - is likely to be at best an indirect one. It is the aim of the present chapter to explore this link in more detail. It is hoped that by combining methodologies from corpus linguistics and psycholinguistics to investigate the relationship between frequencies of occurrence in corpora and the representation of collocations in the mind, we will gain a clearer idea of what frequency information might mean for language learning. Section 4.2 will review existing evidence on the psychological status of formulaic language. We will see that, though there is a good case for the idea that some forms of formulaic language – especially idioms – are processed differently from other types of language, the link between corpus frequency data and processing remains unclear. Section 4.3 will provide a framework for studying the effects of frequency on the processing of collocations by describing the major methods which have been used to quantify collocation frequency. Subsequent sections will then evaluate empirically what these methods can tell us about the likely representation of collocations in the mind. Section 4.4 will consider how well each method is able to predict psychological 'word associations', while Section 4.5 will examine Hoey's claim that high frequency collocations 'prime' each other.

4.2 Evidence on the processing of formulaic language

Introduction

The idea that formulaic language is processed by the mind differently from non-formulaic language is not a new one. Wray (2002, p. 7) traces it back as far as John Hughlings Jackson, who, in the mid-nineteenth century, noticed that aphasics who were not able to construct novel utterances were nevertheless able fluently to recite rhymes, prayers and routine greetings. The importance of formulaic language for non-impaired speakers, meanwhile, was recognised by Saussure (1916/1965) and Jespersen (1924/1976), who both suggested that the mind might ease the burden of

language production by taking the 'short-cut' of bundling together common clusters of language into unanalysed formulas. Psycholinguistically-oriented researchers today continue to explore the apparently special status of formulaic language in aphasics and its supposed utility as a processing 'short-cut' for unimpaired speakers. The present section will briefly review each of these strands of research.

Evidence from aphasia research

Van Lancker Sidtis (2004, pp. 19-27) describes how, following Jackson's lead, aphasia researchers have repeatedly observed that a class of language variously described as 'non-propositional', 'automatic', or 'holistic' tends to be preserved in a wide range of aphasias. Types of language that are typically preserved include speech formulas, pause-fillers, expletives, sentence stems, serial speech, and proper nouns. Aphasia is typically the result of damage to the left-hemisphere of the brain, which is normally considered to play the dominant role in language processing. The persistence of non-propositional forms in aphasics has therefore led many researchers to suggest that they may be localised to the right hemisphere (Van Lancker-Sidtis, 2004, pp. 22-27; Wray, 2002, pp. 236-243). This interpretation fits well with the supposed specialised information-processing abilities of the two hemispheres: the left being associated with "sequential and computational operations", and the right with "holistic and configuration recognition" (Van Lancker-Sidtis, 2004, p. 31). It is also supported by a number of detailed empirical findings, as described by Van Lancker-Sidtis (2004, pp. 22-27). Greater opening of the left-side of the mouth has been found in aphasics producing 'automatic' speech, and greater opening of the right-side in the production of 'propositional' utterances, suggesting that the former is localised to the right-hemisphere and the latter to the right (Graves & Landis, 1985). Patients who have had their left hemisphere removed are reported to retain non-propositional language when the ability to form novel utterances has been lost (Van Lancker & Cummings, 1999), while a patient whose right basal-ganglia was removed lost the ability to recite previously familiar verses (Speedie, Wertman, T'air, & Hellman, 1993). Left-brain-injured patients whose right-brain was either temporarily made inactive by injection or was damaged by a new stroke showed a loss of previously preserved language (Cummings, Benson, Walsh, & Levine, 1979; Kinsbourne, 1971). Functional imaging has suggested that naming tasks are associated with increased activation of the left-hemisphere and counting with the right (Van Lancker, McIntosh, & Grafton, 2003),

while cerebral blood-flow studies have also associated automatic speech with right hemisphere activation (Ingvar, 1983; Larsen, Skinhoj, & Lassen, 1978; Ryding, Bradvik, & Ingvar, 1987). Finally, a double-dissociation has been found in language comprehension, with left-brain-injured patients performing poorly in understanding literal expressions, but better with idiomatic and formulaic language, whereas right-brain-injured patients showed the reverse pattern (Kempler, Van Lancker, Marchman, & Bates, 1999; Van Lancker & Kempler, 1987).

While these findings appear to make a good case for a link between the right hemisphere and non-propositional language in aphasics, Wray points out that it would be somewhat premature to claim that formulaic language in general is localised to the right brain. The category of non-propositional language both includes much that would not normally be called formulaic (e.g., expletives, pause-fillers, proper nouns) and excludes certain important classes of formulaic language. In particular, high frequency chunks that carry informational content are not included under this heading (Wray, 2002, p. 239). Moreover, it is not clear whether any association between the right-brain and non-propositional language is a normal feature of the language system, or a reaction to injury (Wray, 2002, p. 240).

The processing of formulaic language

The processing of idioms

The most extensive body of research on the processing of formulaic language by non-impaired speakers has focused on the comprehension of idioms . Early work in this area claimed that, since the meanings of idioms cannot be derived from their component parts, they must be stored in the mental lexicon as individual items, akin to 'big words'. Whereas the comprehension of literal strings of language is taken to involve decoding the component words and combining their meanings according to the general rules of the language, idioms are, researchers claimed, simply looked up as extended lexical items. Support for this picture came from the finding that idiomatic phrases (e.g. *break the ice*) are recognised as meaningful strings more rapidly than literal controls (e.g. *break the cup*) (Swinney & Cutler, 1979), which was taken to suggest the involvement of a rapid, holistic recognition mechanism. Models differed as to how and when literal and figurative readings became active. According to one version, the two readings are adopted selectively, according to preceding context

(Bobrow & Bell, 1973), according to others they operate either simultaneously (Swinney & Cutler, 1979), or one after the other, with a literal reading invoked only if the figurative fails (Gibbs, 1980).

Later research has largely rejected the idea that idioms are processed in an entirely holistic, word-like, manner (e.g., Cacciari & Tabossi, 1988; Gibbs, Nayak, & Cutting, 1989; Titone & Connine, 1999). Perhaps most influentially, Gibbs and Nayak (1989) note that approaches which see idioms as separate lexical entries fail to account for the productivity of certain idioms. That is, they cannot explain why some idioms can be syntactically altered while retaining their figurative meanings (e.g. 'John laid down the law' can become 'The law was laid down by John') but others cannot ('John kicked the bucket' cannot become 'The bucket was kicked by John' without losing its idiomatic meaning). To account for such cases, Gibbs and Nayak propose the *idiom decomposition hypothesis*, according to which idioms may be classified as either decomposable or non-decomposable. If an idiom is decomposable, its constituent parts can be associated with the components of its literal referent. Thus, when *pop the question* is glossed as *propose marriage*, the noun *question* clearly refers to the proposal and the verb *pop* to the act of making it. In the same way, the *law* of *lay down the law* refers to rules, *lay down* to the act of invoking them. Non-decomposable idioms, on the other hand, do not display such correspondences - no part of *kick the bucket* can easily be analyzed as referring to any part of dying, nor can *chew the fat* be broken down into components corresponding to parts of leisurely conversation. The idiom decomposition hypothesis claims that because the individual parts of decomposable idioms have recognisable meanings, those idioms will be syntactically productive - when *the question is popped*, the transformed components maintain their individual figurative meanings – while non-decomposable idioms cannot be altered without losing their figurative sense.

This division of idioms into decomposable and non-decomposable has been found to correspond to differences in processing. Examining reading times for the two types of phrase through a self-paced reading task, Gibbs, Nayak and Cutting (1989) found that decomposable idioms are read significantly faster than both similar literal phrases (e.g. *pop the question* was read faster than *ask the question*) and non-decomposable idioms. In contrast, non-decomposable idioms were found to be read significantly more slowly

than similar literal phrases (*kick the bucket* took longer to read than *fill the bucket*). This lead the authors to suggest a model on which readers always try to analyse phrases into component parts (an analysis which may, but need not, involve the activation of literal meanings). Decomposable idioms are still hypothesised to have directly stipulated meanings, but because their parts contribute systematically to these meanings, component analysis aids their recognition. Non-decomposable idioms, on the other hand, can only be processed holistically, the process of analysis doing nothing to aid their retrieval. It is this which makes their reading more difficult. While the reading of decomposable idioms is similar to that of literal strings, idioms are read faster because they are more familiar to readers.

Peterson et al (2001) have further refined this picture of holistic vs. componential comprehension by separating the action of semantic from syntactic processing. Using a naming task, they found that, regardless of whether the preceding context biases a literal (e.g., *the soccer player slipped when he tried to kick the*) or a figurative (e.g., *the man was very old and feeble and it was believed that he would soon kick the*) reading process, syntactically-congruent completions (e.g. *town*) were read faster than syntactically-incongruent completions (e.g. *grow*). This 'syntactic priming' effect held true regardless of the degree of decomposability of a phrase and suggests, they argue, that syntactic processing continues throughout the reading of an idiom. In a parallel study designed to detect 'conceptual priming', the same researchers found that literal phrases primed semantically congruent completions (e.g. after reading the stem *the soccer player slipped when he tried to kick the*, the concrete (and so kickable) noun *shelf* was named faster than the abstract (and so not kickable) *truth*), while matched figurative phrases (e.g. *the man was very old and feeble and it was believed that he would soon kick the*) did not. Peterson et al conclude that, though syntactic processing continues after a phrase has been recognised as idiomatic, processing of the literal semantics is halted.

The processing of corpus-derived formulas

Though the processing of idiomatic phrases has attracted the majority of research attention, it should be clear from Chapter 2 that such phrases make up only a small percentage of formulaic language. Moreover, it seems possible that the semantic opacity and high salience of such items may give them a different psychological status

from other types of sequence. Indeed, the apparent influence of semantic type (i.e. decomposable vs. non-decomposable) on processing appears to indicate that the most relevant factor for these items is not their frequency but their meaning. It is only in recent years that the sorts of high frequency but semantically regular formulas identified through corpus analysis which are our primary concern have started to receive serious research attention.

An early study in this area is that of McKoon and Ratcliff (1992, described in more detail in Section 4.5), who looked for evidence of priming between high frequency collocates. Though they found a small priming effect, the authors acknowledged possible problems with their source items, both because of the potential unreliability of the small corpus used and because effects of frequency were not distinguished from those of psychological association (see Section 4.4), and so decline to draw strong conclusions. They do tentatively suggest, however, that co-occurrence statistics may have some applicability as predictors of priming.

Schmitt et al (2004) used a dictation task to determine whether recurrent word clusters identified by corpus analysis are stored in the mind as holistic formulas. The clusters used were sequences of between two and six words, some taken from published listings and others found by corpus analysis to be recurrent contexts of certain key words. A set of 25 clusters were selected which varied in length, frequency, transparency of meaning and according to how intuitively 'holistic' they appeared. These clusters were embedded in a story, which was recorded and played to native and non-native speakers of English in 20-25 word segments. After hearing each burst, native speakers were asked to perform a simple mental arithmetic task and then to repeat what they had heard. Non-natives were only asked to repeat the segment. The thinking behind this task was that participants' working memories would be overloaded and they would need to reproduce the segment using their own linguistic resources. Word-for-word reproduction of the target clusters, it was hypothesised, would indicate that the clusters were likely to be holistically stored. For native speakers, a great deal of variation was found between different clusters, with some (e.g. *go away, I don't know what to do*) being faithfully reproduced by most participants, and others (e.g. *in the same way as, aim of this study*) usually being either avoided or reproduced only partially. Non-natives, unsurprisingly, performed rather

less well overall than natives, and again there was much variation between phrases. For both groups, accuracy of reproduction was not found to correlate with either the frequency or the length of clusters; the researchers suggest that semantically more transparent items (e.g. *go away*) may have been better recalled, while sentence stems (e.g. *in the same way as*) were less well remembered, but strong correlations were not found. They conclude that frequency data from corpora do not appear to be particularly strong predictors of holistic storage. However, the somewhat eclectic mix of clusters used in the study, and the use of a methodology (the dictation task) whose ability to tap holistic processing had not been independently validated, renders this conclusion rather speculative.

Schmitt and Underwood (2004), also failed to find evidence for the holistic storage of corpus-derived clusters. As in Schmitt et al (2004), some clusters were taken from published listings and others were identified by corpus analysis as recurrent contexts of certain key words. The final listing of 21 items included lexical phrases, transparent metaphors, saying/proverbs, and idioms. Sequences were between four and eight words long and were deemed to be relatively predictable from their initial words; the authors also report that items which appeared with low frequency in the British National Corpus or CANCODE (a corpus of spoken English) were excluded, though they do not specify what level of frequency was required for inclusion. The phrases were embedded in short contexts, which were presented to participants word-by-word on a computer screen, with participants pressing a button to bring up each new word. The thinking behind this method was that the time taken for participants to press the button would indicate how long it had taken them to recognize and process the word. If formulaic sequences are stored holistically, the authors reasoned, recognition times for the latter parts of these phrases should be hastened once the phrase has been recognised. However, neither native nor non-native participants demonstrated any advantage in reading the final words of formulaic sequences over the same words in non-formulaic contexts. As with the previous study, the mix of items used, and the failure to provide specific frequency data, makes these results rather difficult to interpret. Moreover, as the authors themselves suggest, the word-by-word presentation paradigm may have disrupted normal holistic processing strategies.

Other recent studies do appear to provide support for a link between the frequency of clusters in corpora and holistic storage. Using the same materials as Schmitt and Underwood (2004), Underwood et al (2004) used an eye-tracking paradigm to study native and non-native speakers' reading of formulas embedded in short contexts. They found that both natives and non-natives fixated on target words less often when they appeared as the final words of a formulaic sequence than when they appeared in other contexts. They also found that natives (but not non-natives) fixated on targets for shorter durations when they were found within formulas. Underwood et al conclude that these results are consistent with the idea that formulas are stored holistically by native speakers and suggest that the somewhat ambiguous results seen for non-natives (fewer, but not shorter) fixations, may indicate only partial knowledge of the formulas). The emergence of reliable effects here, using the same materials as those in Schmitt and Underwood (2004), suggests that the failure to find an advantage for formulaic sequences in that study may have been due to the acknowledged methodological problems.

Jiang and Nekrasova (2007) also claim to find evidence for holistic storage of formulaic sequences in both native and non-native speakers. They found grammaticality judgements for 26 formulaic sequences taken from previous corpus-based studies (again, no actual frequency data are provided) to be both faster and more accurate than judgements for matched control strings (in which formulas were changed by one word to create a more novel string). This suggests, they conclude, that the formulaic items are recognised holistically, obviating the need the full syntactic analysis which must presumably take place for the novel strings.

In a series of studies, Tremblay et al (in preparation) used self-paced reading and memory tasks to determine whether high frequency lexical bundles are holistically stored in the mind. The lexical bundles were either four-word strings found in the spoken part of the British National Corpus (BNC) with a mean frequency of at least 10 occurrences per million words, or five-word strings appearing in the same corpus at least five times per million words. Control phrases were created by substituting one word in each string with a replacement which was individually more frequent and (on average) shorter than the original and such that the new phrase was less frequent in the BNC than the original. In a word-by-word self-paced reading task, the replaced word

in the original bundle was found to be read significantly faster than its replacement in the control phrase; in a segment-by-segment self-paced reading task, entire lexical bundles were read significantly faster than control phrases, and in a sentence-by-sentence task, sentences containing the original bundles were read significantly faster than those containing their replacements.

Working on the idea that holistically-stored bundles should take up less space in working memory than novel strings of the same length, which need to be represented word-by-word, Tremblay et al also test the memory load of lexical bundles and their controls. They presented subjects with the lexical bundles or control phrases described above, along with a string of individual words and then asked them to recall the phrase and the words. When the input was presented visually, they found significantly better recall for both lexical bundles and their following words than for control phrases and their following words, suggesting that lexical bundles may indeed place less strain on working memory. When input was presented in the auditory modality, better recall was again found for the lexical bundles than for the control phrases, but not for their following words. The authors speculate that this may have been because natural intonation features of the lexical bundles had been deliberately stripped out of the recordings by using synthesised speech.

Summary and conclusions: formulas in the mind

We have seen that there is much evidence from aphasia research that 'non-propositional' language has a different psychological status from other types of language. This suggests that formulaic language may be somehow psychologically special, and some have suggested that it may indicate that formulaic language is stored in the right-hemisphere of the brain, rather than the left-hemisphere, which is usually dominant in language processing (Van Lancker-Sidtis, 2004; Wray, 1992). However, the implications of this work for high frequency collocations are not clear-cut. Non-propositional language is not usually taken to include information-bearing high frequency phrases (such as collocations), and it is also not clear whether the special status of such language reflects normal psychological organisation or is a reaction to language impairment. There is also much evidence that idioms are processed differently from literal language. Again, however, idioms are only one sub-type of formulaic language, and it seems likely that their special status is a result of their

semantics, rather than of their frequency. In recent years, a few studies have examined the processing of corpus-derived formulas, but results have been mixed. Though the balance of evidence seems to be in the direction that high frequency formulas are processed more efficiently than novel language, further research is clearly needed. A particular failing of previous studies is that few have provided thorough information on the actual frequencies of the phrases used; Tremblay et al (in preparation) and McKoon and Ratcliff (1992) being the only ones to provide well-defined frequency-based criteria for inclusion. From the perspective of the current thesis, it is also important to note that most existing research concerns the processing of relatively long word strings (usually four words or more). It seems likely that such sequences will be more salient – and so more likely to attain a special psychological status - than the two-word collocations which are the main focus of this thesis. We must conclude then, that the question of the psychological reality of high frequency formulas in general, and of high frequency collocations in particular, remains an open one.

4.3 Frequency-based methods of identifying collocations

Introduction

One factor which makes two-word collocations an appealing resource for the study of psycholinguistic frequency effects is the fact that a number of statistical methods have been developed for describing their frequencies. This section will discuss the most widely used of these. Sections 4.4 and 4.5 will then use these methods as a basis for testing the relationship between frequency and processing.

Raw frequency

The simplest frequency-based method of establishing whether a particular word combination is a collocation is simply to count the number of times that combination occurs. Thus, finding that *strong tea* occurs in the BNC 28 times, while *powerful tea* appears only 3 times, we may conclude that the former is the more conventional collocation. The problem with this approach is that many of the strongest collocations in any corpus would simply be those made up of the most frequent words; amongst the strongest collocations in any English corpus, for example, would be *a-the, of-and* and *to-was*. Such combinations appear to be frequent, not because the words stand in any particularly interesting relationship to each other, but simply because they are so

common that their regular co-incidence comes about by chance. Moreover, the simple frequency-based approach to collocation is not only in danger of finding collocational relations where there are none, it may also miss many genuine collocations, since strongly associated word pairs composed of words which are individually rare (*battering ram*, *zero-sum game, abject poverty*) would not register at all. Corpus linguists have used two main types of method to improve on raw frequency counts: asymptotic hypothesis tests and mutual information. The two approaches are conceptually different and typically produce rather different types of results. We shall deal with each in turn.

Hypothesis testing

The main hypothesis testing methods of identifying collocations are the *z-score, t-score, chi-squared* and *log-likelihood* tests. These test the null-hypothesis that words appear together no more frequently than we would expect by chance alone. They can therefore by seen as formalisations of Hoey's definition of collocations as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, p. 7).

All of the hypothesis testing methods start by calculating how many times we would expect to find a word pair together in a corpus of a certain size by chance alone, given the frequencies of its component words. To calculate this, we first determine how probable it is that any word pair, chosen at random from the corpus, will be the combination we are studying. This is usually calculated with the formula:

$$P(w1w2) = P(\text{w1}) * P(\text{w2})$$

This states that the probability that any randomly selected pair of words will be the combination *w1* w2 is equal to the probability of *w*1 occurring on its own multiplied by the probability of *w2* occurring on its own. For example, the word *strong* appears in the British National Corpus (BNC) 15,768 times, the word *tea* appears 8,030 times. Since the BNC has a total of 100,467,090 words, we can calculate the probabilities of occurrence of each as follows:

$$P(strong) = \frac{15{,}768}{100{,}467{,}090} = .00016$$

$$P(tea) = \frac{8{,}030}{100{,}467{,}090} = .00008$$

This tells us that if we select any word at random from the BNC, the probability that it will be the word *strong* is .00016 and the probability that it will be *tea* is .00008. We can then conclude that the probability that any two words, picked at random from the BNC will be the pair *strong tea* is:

$$P(strong\ tea) = .00016 * .00008 = 1.25e\text{-}08$$

Although this probability is very low, we would still expect *strong* and *tea* to occur together at some point, simply because there is such a large number of words in the corpus. In fact, *strong tea* can be predicted to occur:

$$1.25e\text{-}08 * 100{,}467{,}090 = 1.26 \text{ times}$$

Since we know that strong tea actually occurs 28 times, we can conclude that the pair collocates more frequently than chance. The aim of the hypothesis testing methods is to determine the statistical significance of this apparently greater than chance frequency (Manning & Schütze, 1999, pp. 162-163). A number of different statistics are commonly used.

The *z-score* is calculated with the formula:

$$z\text{-}score = \frac{O - E}{\sqrt{E}}$$

where O is the observed frequency of occurrence of the collocation, and E is the expected frequency of occurrence on the null hypothesis that there is no relationship between the words.

For *strong tea*, then, we get:

$$\textit{z-score} = \frac{28 - 1.26}{\sqrt{1.26}} = 23.82$$

A problem with the z-score is that, because it takes expected occurrence as its denominator, a misleadingly high score can be returned if the words involved are infrequent in the corpus (Evert, 2004). A measure which seeks to avoid this problem is the *t-score*, which takes observed occurrence as its denominator. The *t-score* is calculated as follows:

$$\textit{t-score} = \frac{O - E}{\sqrt{O}}$$

Thus, for the pair *strong tea*:

$$\textit{t-score} = \frac{28 - 1.26}{\sqrt{28}} = 5.05$$

Both z-score and t-score have been criticised on the grounds that they assume an approximately normal distribution of results. It has been argued that this tends not to be the case for rare events like collocations (Dunning, 1993). The hypothesis testing methods conceive of a corpus as a series of bigrams, each of which may have a value of 1 (the bigram is the word pair being examined) or 0 (the bigram is not the word pair being examined). Two-outcome tests of this sort (analogous to a series of coin-tosses) generate a binomial distribution. Where the mean number of positive outcomes is relatively high (as in the case of getting heads from a coin toss), the binomial distribution approximates the normal distribution. However, where the mean number of positive outcomes is relatively low (as in the case of collocation), the binomial distribution is heavily skewed, so violating the assumption of normality (Dunning, 1993, pp. 64-65).

In response to this problem, some researchers have recommended the use of non-parametric tests, which do not rely on the assumption of normality. One such test is

Pearson's chi-square. This relies on the following 2x2 contingency tables showing the observed and expected occurrences in the corpus of each word and of the collocate:

Observed

	w2 = X	w2 ≠ X	
w1 = Y	O_{11}	O_{12}	$= R_1$
w1 ≠ Y	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

($R_1 = O_{11} + O_{12}$; $R_2 = O_{21} + O_{22}$; $C_1 = O_{11} + O_{21}$; $C_2 = O_{12} + O_{22}$; $N = R_1 + R_2 + C_1 + C_2$)

Expected

	w2 = X	w2 ≠ X
w1 = Y	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
w1 ≠ Y	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Thus, for the word pair *strong tea,* we get the tables:

Observed

	w2 = tea	w2 ≠ tea
w1 = strong	28	15740
w1 ≠ strong	8002	97,596,164

Expected

	w2 = tea	w2 ≠ tea
w1 = strong	1.3	157,66.7
w1 ≠ strong	8,028.7	97,596,137.3

On the basis of these tables, Chi-square is calculated as follows:

$$x^2 = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}}$$

Thus, for *strong tea:*

$$x^2 = \frac{97{,}619{,}934(28 - 1.3)^2}{1.3 * 97{,}596{,}137.3} = 548.5$$

A problem with chi-square is that it is known to be inaccurate when small numbers are involved. Dunning therefore recommends instead using the likelihood ratio, which is more robust at lower frequencies (1993). Like chi-square, log-likelihood makes use of the contingency tables described above. It is calculated as follows (this version of the equation comes from Evert (2004)):

$$\text{log-likelihood} = 2\sum_{ij} O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

Thus, for *strong tea*:

log-likelihood =

$$2 * \left[\left(28 * \ln\left(\frac{28}{1.3}\right)\right) + \left(15740 * \ln\left(\frac{15740}{115766.7}\right)\right) + \left(8002 * \ln\left(\frac{8002}{8028.7}\right)\right) + \left(97596164 * \ln\left(\frac{97596164}{97596137.3}\right)\right)\right]$$

= 118.6

I have noted the rationale behind all of these statistics is that of testing the null-hypothesis that a word pair appears together no more frequently than we would expect by chance alone. Taking this conception literally, we can consult tables of critical values to see how confident we can be in rejecting the null-hypothesis. A t-score of greater than 2.576, for example, would enable us to reject the null-hypothesis with 99.5% confidence (Manning & Schütze, 1999, p. 164). However it is important to note exactly what is meant by a word pair's being more frequent that we would expect 'by chance'. The calculation of expected occurrence is based on a model in which words

are drawn as if from a hat, entirely at random. However, as Manning and Schütze note, language is far more regular than a "random word generator" (1999, p. 166). Grammar, semantics, and real-world occurrences all constrain the construction of real language (this is another way of stating Sinclair's point, noted above, that many factors might account for high frequencies of co-occurrence). It is therefore very common for word pairs to co-occur 'more frequently than random', regardless of specifically collocational relations. Given this, levels of 'statistical significance' are not usually thought to constitute useful cut-off points in identifying collocations. Rather, the statistical tests reported here are used to *rank* word pairs according to their relative likelihood of being a collocation (Manning & Schütze, 1999, p. 166; Stubbs, 1995, p. 33).

Mutual Information

Church and Hanks (1990) propose *mutual information* (MI) as a means of estimating the degree of association between words. This compares the observed number of occurrences of a word pair with its expected number of occurrences, as follows:

$$MI = \log_2 \frac{O}{E}$$

Thus, for *strong tea*:

$$MI = \log_2 \frac{28}{1.3} = 4.43$$

Mutual information can be conceptualised as a "measure of how much one word tells us about the other" (Manning & Schütze, 1999, p. 178). In other words, when we encounter one part of a word pair which has a high mutual information score, we can predict that the other part of the pair is likely to be nearby. This is importantly different from the hypothesis testing methods described above. Clear (1993, pp. 279-282) neatly sums up the point, noting that, whereas "MI is a measure of *the strength of association between two words*", hypothesis-testing methods are measures of "*the confidence with which we can claim there is some association*" (original emphases). This has important implications for the types of word pairs retrieved by the two

methods. Clear gives the word pair *taste-arbiters* as a typical example of a combination attaining a high MI score. Though the pairing is not particularly frequent, it accounts for a high proportion of the occurrences of its component words; in fact, Clear reports that one quarter of all appearances of *arbiters* are within two words of an appearance of *taste*. The two are strongly associated, then, in that where we find *arbiters*, we have a good chance of finding *taste*. However, its relatively low frequency of occurrence reduces the statistical reliability of this pattern – i.e. we cannot be very confident that the relationship will be generalisable to other samples of language. A typical example of a pair with a high score on hypothesis-testing methods, on the other hand, is *taste-for*. While the association between these words is much weaker than that between *taste* and *arbiters*, the pair occurs with much higher frequency. The connection, though weaker, is therefore more reliable.

Like *z-score*, MI takes 'expected occurrence' as its denominator. Like the z-score, therefore, it can give very high scores for collocations which include low frequency words, even if the total number of occurrences of the collocation is very low. To guard against accepting word pairs as strong collocations on the basis of minimal evidence, therefore, MI is often used in conjunction with a minimum frequency threshold (e.g., Church & Hanks, 1990, p. 24). A less widely-used method to correct for this problem has been to adjust the MI formula to give greater weight to the 'observed occurrences' part of the equation (Evert, 2004). Suggested corrections include *local MI*, *MI-squared,* and *MI-cubed*, which are calculated as follows:

$$local\ MI = O \times \log_2 \frac{O}{E}$$

$$MI^2 = \log_2 \frac{O^2}{E}$$

$$MI^3 = \log_2 \frac{O^3}{E}$$

Directional measures of collocation

All of the measures of collocation discussed so far are non-directional, in the sense that it makes no difference which part of the word pair is taken as node and which as collocate. However, this may be misleading. As Stubbs points out, though the pair *kith* and *kin* have the same score on all of the measures regardless of which word is taken as the node, the relationship between the two words is clearly not symmetrical: *kith* predicts *kin* with around 100% certainty, whereas *kin* can be found in other contexts (1995, p. 35). The non-directionality of these measures may be particularly problematic for our task of predicting the psychological correlates of frequency data, since it seems highly likely that any associative links running from *kith* to *kin* will be stronger that those running in the opposite direction. It would therefore be useful to have a statistic which reflects this.

A simple way of achieving a directional score would be to calculate the conditional probability of one word, given another. This could be done by simply dividing the frequency of the word pair by the frequency of the node. Since the conditional probabilities are usually likely to be rather small, this figure can be multiplied by 100 for ease of reading:

$$P\,(w_2|w_1) = 100 \text{ x } \frac{w_1 w_2}{w_1}$$

Thus, to return to our earlier example, the conditional probability of the collocate *tea*, given the node *strong*, is:

$$100 \text{ x } \frac{28}{15{,}768} = 0.178$$

while the conditional probability of the collocate *strong*, given the node *tea*, is:

$$100 \text{ x } \frac{28}{8{,}030} = 0.349$$

indicating that this collocation is rather more important for *tea* than it is for *strong*. This approach has not been widely used in corpus linguistics, though Handl (2008) has recently suggested a similar method, and psychologists have speculated that the formula described here may be related to word association norms (Anderson, 1990, p. 64).

Variables

Using any of the above methods will involve the analyst in two important decisions which we have not yet been addressed: how close together two words need to be to count as 'co-occurring' (the question of 'span'); and whether we should pool the counts for each inflectional/derivational form of a word - so that, for example, *argue strongly, argued strongly* and *strong argument* would count as three occurrences of a single collocation - or whether separate counts should be made for each form (the question of 'lemmatisation').

With regard to span, Jones and Sinclair report that the vast majority of a word's collocational influence is found within a span of four words to its left and right (1974, pp. 21-22). Though much longer-distance dependencies have been claimed to exist (Clear, 1993, p. 276), this '+/- 4 word' guideline has been widely accepted (Hoey, 2005, pp. 4-5). A less satisfactorily resolved issue related to span selection is that of whether association measures should be adjusted to take account of the span used. We have seen that standard association measures are based on comparing the number of times we would expect to find two words together if they were selected at random with the number of times we actually find them together. Clearly, however, the number of times we would expect to find two words directly adjacent to each other is rather lower than the number of times we would expect to find those words somewhere within a span of +/- 4 words of each other. Specifically, if the probability that *word2* is the word directly after *word1* is given by the formula:

$$P(word1) \times P(word2)$$

then the probability that *word2* is one of the eight words falling within a +/- 4 word span of *word1* is:

8 x *P*(*word1*) x *P*(*word2*)

To maintain the original logic of the association measures therefore, we would need to make this adjustment when calculating the 'expected frequency' part of the equations. While some publicly-available software for calculating association measures allows this adjustment to be made (e.g. *T/Z and Mutual Information Calculator* (Klarskov Mortensen, 2003)), others (e.g. *WordSmith Tools* (Scott, 1996)) do not make any adjustment for span. It could be argued that the latter choice violates the logic of the original formulas, leading to artificially-inflated scores when wider spans are used and to non-comparability between studies using different spans. On the other hand, the author of *WordSmith Tools* argues against including any adjustment for span on the grounds that word pairs which frequently co-occur directly next to each other (e.g. *rely-on*) should not, for that reason alone, be considered stronger than pairs which frequently appear at a certain distance from each other (e.g. *kith-kin*). "[I]f one CASTS ASPERSIONS on something", he asks, "is that more linked that when ASPERSIONS got CAST on it?" (Scott, personal communication). The issue of adjusting association measures for span remains, then, a moot one. Since much of the corpus-based work in this thesis depends on *WordSmith Tools*, I will follow Scott in not making any such adjustment.

On the question of lemmatisation, Halliday (1966, p. 151) has argued that collocation should be seen as existing between 'words' at a rather high level of abstraction. On this view, *strong, strongly, strength* and *strengthened*, for example, should all be regarded as "the same item"; and *a strong argument, he argued strongly, the strength of his argument* and *his argument was strengthened* are all "instances of the same syntagmatic relation". Halliday's argument is that restating the syntagmatic relationship for each form of the words involved would add complexity without a gain in descriptive power because, as far as the collocational pattern is concerned, differences between word forms are irrelevant. Since Halliday published these remarks, however, the assumption that differences between word forms are irrelevant to collocation has been widely questioned. Amongst other, Sinclair (1991, p. 8), Clear (1993, p. 277), Stubbs (1996, p. 38), and Hoey (2005, p. 5) have all argued that lemmatisation may disguise differences in the collocational preferences of different forms of a word. Clear, for example, notes that collocations such as *vested interest,*

crying shame, and *bodes ill* are all restricted to particular inflected forms, a point that would be lost in a lemma-based analysis. Moreover, Clear points out, lack of lemmatisation rarely if ever disguises a collocation, since "one of the inflected forms will appear as a significant collocate, and the potential for the other forms in the paradigm to collocate will be apparent to the human analyst" (1993, p. 277). In the studies that follow, no lemmatisation is used in tallying collocations unless specifically noted.

Evaluating frequency measures

Evaluation of the validity of the various frequency-based measures of collocation has generally been limited to the intuitive assessment of a few top-ranked items (Evert & Krenn, 2001, pp. 1-2). This is probably due to the difficulty of specifying and operationalising any independent criterion of accurate identification. As Clear notes, the most obvious course would be to compare frequency results with the results of an independent manual analysis. However, any such endeavour would be highly problematic: not only would a manual analysis be prohibitively time-consuming, but part of the point of frequency analysis is that it is thought to be capable of uncovering patterns which are not immediately evident to the human analyst (Clear, 1993, p. 282). Indeed, it has been a constant refrain of corpus-based collocation study that "intuition is typically a poor guide to collocation" (McEnery, Xiao, & Tono, 2006, p. 83).

A few studies have, however, attempted to compare language users' intuitions with frequency data. The earliest such study of which I am aware is that of Hoffman and Lehmann (2000), who elicited native and non-native speakers' intuitions regarding 55 word pairs which were found to be strongly associated in the BNC (as measured by log-likelihood). Each pair consisted of a 'low frequency' node (occurring between 50 and 100 times in the corpus) and a collocate found within +/-3 words. Most were adjective-noun (24/55) or noun-noun (19/55) pairs, but the listing also included other parts of speech. Hoffmann and Lehmann prepared a questionnaire in which each node was presented without its pair and asked 16 native and 16 nonnative-speaker informants to supply the collocates. It was found that, on average, native speakers supplied the 'correct' collate in 70% of cases, a figure which the authors judge to be surprisingly high, given the widespread scepticism about the accuracy of intuitions.

Non-native speakers, unsurprisingly, did less well, achieving an average accuracy of only 34%. The native speaker 'success rate' of 70%, however, appears to provide some support for the validity of log-likelihood. In a similar vein, Siyanova and Schmitt (2008) found a significant correlation between the frequency of collocations in the BNC and scores out of 6 given by both native and non-native speakers for the 'typicality' of the collocation (for natives r_s = .578, for nonnatives r_s = .440). Moreover, native (though not non-native) speakers were able reliably to distinguish 'medium frequency' (21-100 occurrence in BNC) from 'high frequency' (>100 occurrences in BNC) collocations.

While these studies offer some encouragement that frequency-based measures are able to detect items which have psychological reality for speakers, they do not attempt to assess the relative merits of the different statistics described above. One paper which does make such an attempt is that of Evert and Krenn (2001). They automatically retrieved around 4,500 adjacent adjective-noun pairs which occurred at least twice in an 800,000 word corpus of German law texts, and around 15,000 pronoun-noun-verb triples which occurred at least three times each in an eight million word portion of the Frankfurter Rundshau Corpus. Two native speakers were asked to identify those adjective-noun pairs which they perceived as 'typical combinations' (including idioms, legal terms, and proper names) and those pronoun-noun-verb triples in which there was a grammatical relation between the verb and the PP, and the triple could be interpreted as support verb construction and/or a metaphorical or idiomatic reading was available. Collocations were taken to be positively identified if they were picked by either informant. Association measures (raw frequency, log-likelihood, t-score, chi-squared, and mutual information) were also calculated for all items on the lists and separate ranked lists produced for each measure. Finally, 'precision' and 'recall' graphs were generated for each list. Precision graphs showed the percentage of items at each level of the lists which were manually-identified collocations; recall graphs showed the cumulative percentage of manually-identified collocation which had been found at each level of the lists. They found that for adjective-noun pairs, t-score and log-likelihood provided the best predictions, while for the pronoun-noun-verb triples, t-score and raw frequency were the best. In both cases, chi-squared and mutual information were the worst predictors.

While Evert and Krenn's paper is useful in showing (in the form of its precision and recall graphs) the sort of shape which a thorough examination of association measures might follow, the generalisability of its findings must be questioned given the small number of informants used (i.e. two, with identification by only one necessary to mark an item as a collocation). Moreover, their specification of items which are to count as collocations (idioms and metaphors, technical terms, proper names, support verb constructions) is rather narrower than the set of potentially psychologically-real word pairs in which the current thesis is interested. The study described in the following section will attempt to go beyond Evert and Krenn's analysis by considering how accurately the various frequency-based methods can predict psychological associations between words, making use of published norms of word association collected from large numbers of participants. It will also attempt to define some approximate rules of thumb as to what levels of each measure are likely to indicate psychologically real collocations.

4.4 Frequency measures and word association.

Introduction

The psychological associates of a word are those other words which first come to a person's mind when they see or hear it. There has been an interest in establishing 'norms' of association since the beginning of the nineteenth century, when they were used as a measure of sanity. Observing that a "derangement" in the "association of ideas" was one of "the most striking and commonly observed manifestations of insanity", Kent and Rosanoff (1910) attempted to establish the common types of association and the variation within normal populations by reading a list of 100 stimulus words to over 1,000 subjects and asking them to respond to each with the first word that occurred to them other than the stimulus word itself. Since the 1960s, word association has come to be used to be used in language studies, where it has been thought to provide evidence about first and second language acquisition and the structure of the mental lexicon (Fitzpatrick, 2007, pp. 320-321).

Word associations are of interest to us because they have been widely linked to collocation. Observing that many associated words appear to be collocates of each other, some psychologists have proposed that words may come to be associated

precisely because they are encountered together on a regular basis (Charles & Miller, 1989). This link between collocation and association has been tested empirically by Spence and Owens (1990), who showed that a group of 47 associated noun-noun pairs co-occurred more frequently, in spans of text ranging from 50 to 1,000 characters (about 10 to 250 words), in the one million-word Brown corpus than did matched non-associates. Moreover, strength of association (as measured by the percentage of respondents providing a particular response) was correlated with frequency of occurrence up to spans of 2,000 characters. This suggests, Spence and Owens conclude, that the co-occurrence of words in language is a major contributor to their being linked in word association norms.

If this is right, then at least part of what is being evidenced by word association tests is the proposed psychological representation of high frequency collocations which this chapter has set out to investigate. It should, therefore, be possible to use such norms to gauge the ability of the various frequency-based measures described above to detect collocations which are likely to be psychologically-real for speakers. The test is imperfect because, though we can conclude with some confidence that collocations which appear in word association norms are linked in the minds of at least some of the population sampled, non-appearance does not necessarily mean that words are *not* so linked. Indeed, it seems likely that only a small proportion of the total number of psychologically-real collocations will be tapped by association tests (especially since such tests typically elicit only one response per participant). Similarly, not all associates are necessarily collocations, since other relationships (e.g. between paradigmatically-related pairs) are also commonly found in association norms. We should not, therefore, expect either all mentally-represented collocations to appear in the association norms or all associations to be mentally-represented collocations. Nevertheless, it seems fair to assume that measures which are good predictors of those psychological collocations which are attested as associates will be good predictors of psychological collocation in general. The research reported in this section explores this possibility by comparing frequency data for collocations in the British National Corpus (BNC) with associations reported in a set of association norms. In particular, it asks how well a set of ranked lists of collocations produced by various frequency-based methods predict the reported associations.

Method

The norms used in this study are taken from the Edinburgh Word Association Thesaurus (EAT)[2]. This database was compiled by Kiss et al (1973) between 1968 and 1973. Researchers presented a range of stimulus words to informants and elicited for each the first word to come to mind. Each stimulus word was presented to 100 different people, most of whom were undergraduates at British universities. Clearly, the associations recorded in EAT are the associates of a particular group of people (British undergraduates) at a particular moment in time (1968-71). It is highly likely that if the same procedure were followed with a different group of participants, or at a different moment in time, many of the associations would be different. To take some simple examples, it is unlikely that if the same experiment were repeated today, the second most common associate of *politics* would be *Wilson*, or that *mobile* would fail to elicit the response *phone*; similarly if the data had been collected from bankers, rather than undergraduates, the most frequent associate of *student* would surely not have been *me*. In examining the relationship between high frequency collocations in the BNC and associates in EAT, therefore, we are examining whether the collocations are likely to have been psychologically real for one particular group of people at one particular moment in time. Far from being a weakness of the current approach, however, this is precisely the point. Many psychological collocations will vary from group to group (even from person to person) and from time to time. Equally, the collocations found in corpora will vary according to when data were collected and what sorts of texts were included. This, as I argued above, is one of the major reasons why drawing any inference from corpus to mind is problematic. The question we need to ask is, given such variability, how much stays constant, such that we can use a corpus like the BNC to make reasonably confident predictions about the mental associations of any particular group of native English speakers?

The base data for the study was a listing of several thousand modifier-noun combinations which had been retrieved from a variety of texts as part of a separate study (see Section 5.3). Frequencies of occurrence in the BNC and a range of association measures were calculated for these combinations (i.e., t-score, chi-squared, log-likelihood, mutual information (MI), z-score, and conditional probability). Since

[2] EAT is accessible online at www.eat.rl.ac.uk/.

association measures often work poorly with low frequency items, word pairs occurring fewer than 5 times in the BNC were excluded from further analysis. Also excluded were combinations whose modifier part was not listed as a stimulus in the EAT. For the remaining 3,168 combinations, the EAT was consulted to see whether the noun part of the combination was listed as an associate of the modifier part. Where the noun was an associate, the strength of association was also noted. On the basis of these data, two questions were addressed: 1) which frequency measure is the best predictor of psychological associations? and 2) what value of each measure is likely to indicate a psychological association?

Results and discussion

Question 1: Which frequency measure is the best predictor of psychological associations?

Spearman point-biserial correlations (Field, 2005, pp. 131-134) were used to compare the relative effectiveness of each frequency-based measure as a predictor of whether words are likely to be psychological associates. The correlation works by taking each word pair on our list as an individual case and assigning to each a value of either 1 (if it is an associate) or 0 (if it is not). Correlation scores (shown in Table 1) are the Pearson correlations between these values and the scores assigned to each pair by the various frequency-based measures.

Table 1: point-biserial correlations between frequency measures and association

Association measure	correlation*
raw frequency	.258
t-score	.268
chi-squared	.300
log-likelihood	.291
MI	.250
z-score	.299
conditional probability	.364

*all correlations are significant at the $p < .0001$ level

These data indicate that all association measures are reliable predictors of psychological association. Although the correlations are rather weak by usual standards of interpretation, it should be born in mind that the imperfect match between association norms and mentally-represented collocation will mean that there is a considerable amount of 'noise' on the association side of the comparison. Given this,

the maintenance of statistically significant correlations at around the $r = .3$ level is an encouraging sign. The best results are obtained for conditional probability. Of the more traditional methods, chi-squared is the best, followed closely by z-score. In fact, these two measures provide almost identical rankings for the combinations studied here. Only minor differences exist between the two ranked sets, and across all 3,168 pairs there is an overall Spearman correlation between the two scores of $r = 1$. Chi-squared and z-score are followed in their ability to predict association by log-likelihood, then t-score, then raw frequency, then MI.

While these figures give a good indication of the relative strengths of each measure, a more easily interpretable picture can be gained by considering what percentage of combinations at a given level of each score are associated. This can be calculated by first ranking combinations according to their scores on each measure and then dividing these ranked lists into percentiles; i.e. creating 10 different bands, corresponding to the top 10%, next 10%, etc. We can then evaluate what percentage of word pairs at each level are associates. For good predictors, we would expect to see high percentages in the upper percentiles and low or zero associates in the lower bands.

The results of this analysis are presented in Table 2. We can see that all measures provide reasonably good returns, with at least 43.5% of word pairs in the top band being associates. The best results are given, as we would expect, by conditional probability, with approximately 55% of pairs in the top band being associated. Since the associates reported in the norms are likely to be only a sub-set of the collocational associates found in the population, I would argue that this is an encouragingly good rate of return. We can speculate that the true percentage of collocations from this band which are psychologically real for the current population is likely to be rather higher.

Table 2: percentage of word pairs in ranked lists which are associated

	raw frequency	t-score	chi-square	log-likelihood	MI	z-score	conditional probability
top 10%	46.06	46.06	47.95	47.63	43.53	47.95	54.89
10%-20%	30.60	30.60	36.91	33.75	32.18	36.91	40.38
20%-30%	22.71	24.61	27.13	25.55	25.24	27.13	25.24
30%-40%	22.08	22.08	23.34	22.40	22.40	23.34	23.66
40%-50%	22.71	21.45	14.83	19.24	17.35	14.83	14.20
50%-60%	14.83	14.20	9.46	11.99	15.77	9.46	12.93
60%-70%	12.93	13.25	11.36	12.93	14.51	11.36	10.41
70%-80%	8.83	9.15	11.04	10.41	10.09	11.04	6.94
80%-90%	10.41	10.73	9.15	8.20	10.41	8.83	5.36
90%-100%	7.57	6.62	7.57	6.62	7.26	7.89	4.73

The analysis so far has considered only whether combinations are listed as associates or not. However, some associations seem to be stronger than others. Responses which are supplied by a number of participants are clearly more likely to be widespread in the population than those elicited from one person only (which may reflect individual idiosyncracies). It may be instructive, then, to consider the various measures' ability to predict word pairs which are 'robust' associates in this sense. Table 3 again shows Spearman correlations between the various measures and association, but this time only counts as associates those pairs which were provided by at least 5% of respondents.

Table 3: point-biserial correlations between frequency measures and robust association

Association measure	correlation*
raw frequency	.156
t-score	.164
chi-square	.229
log-likelihood	.197
MI	.223
z-score	.229
conditional probability	.244

*all correlations are significant at the $p < .0001$ level

Overall, we can see that the correlations are reduced somewhat. Again, conditional probability is the best predictor of association. Of the traditional measures, z-score and chi-square are again the best. MI has now moved ahead of log-likelihood, t-score, and raw frequency, which remain in the same order as before.

As before, we can also look at what percentage of pairs at each percentile of the ranked lists are robust associates (Table 4). The results here emphasise the improved performance of mutual information as a predictor if we look only at robust associates. On the new analysis, MI equals conditional probability in having the highest percentage of associates in its top 10% of pairs (20.82%). Raw-frequency and t-score perform worst, each returning 11.99% associates.

Table 4: percentage of word pairs in ranked lists which are robustly associated

	raw frequency	t-score	chi-square	log-likelihood	MI	z-score	conditional probability
top 10%	11.99	11.99	19.24	14.51	20.82	19.24	20.82
10%-20%	8.83	9.15	10.09	10.41	8.20	10.09	11.36
20%-30%	6.94	6.62	6.94	6.62	6.31	6.94	4.42
30%-40%	3.79	4.73	3.15	4.42	2.21	3.15	3.79
40%-50%	4.73	4.42	1.26	3.47	2.21	1.26	0.63
50%-60%	2.84	2.21	1.26	2.52	1.89	1.26	1.26
60%-70%	2.52	2.52	1.26	0.63	1.26	1.26	1.26
70%-80%	1.58	1.58	0.00	1.26	0.63	0.00	0.32
80%-90%	0.63	0.63	0.95	0.32	0.63	0.95	0.63
90%-100%	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Taking the two sets of results together, conditional probability is clearly the best indicator of whether a word pair will be associated. Of the traditional corpus measures, z-score and chi-square - which are virtually identical - perform the best. Log-likelihood outperforms t-score, which is in turn better than raw frequency. Mutual information is less good at predicting associates than these three methods, but is better at predicting the more 'robust' associates.

Question 2. What value of each measure is likely to indicate a psychological association?

Though it is generally agreed that association measures are best used to rank collocations, rather than to distinguish absolutely between collocation and non-collocation, the literature on association measures has also included some guideline 'cut-off' points at which word pairs are likely to be collocationally interesting. Church and Hanks (1990), for example, recommend paying attention to pairs with MI scores or three of more, and Stubbs (1995) suggests filtering out word pairs with t-scores of lower than 2 and MI score of lower than 3. It will be worth asking, then, whether any such 'cut-off' points are identifiable in the data analysed here. The 10% bands used

above may be rather too broad for these purposes (a glance across tables 2 and 4 shows rather sharp drops in the percentage of associates between levels). To provide a clearer view of the data, therefore, the ranked lists for each association measure were divided more finely into 30 segments (of 105 or 106 word pairs each). The percentage of word pairs which are associates (what I will refer to as the 'hit-rate') was then calculated for each segment.

Figure 1: hit-rates at 30 bands of a ranked raw-frequency list

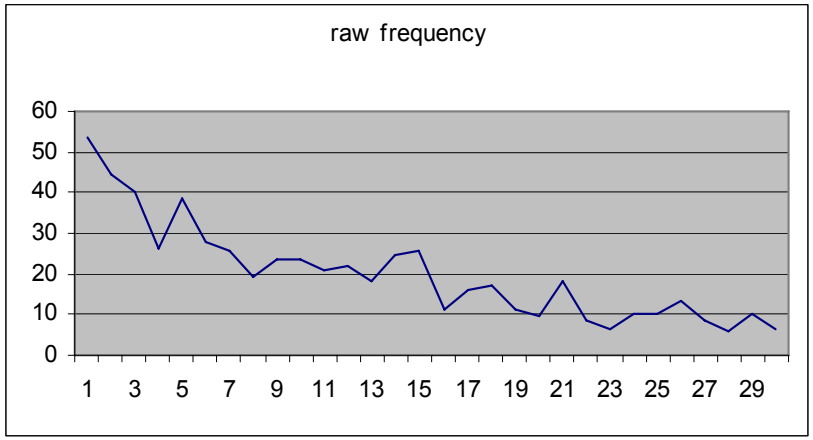


Figure 2: hit-rates at 30 bands of a ranked t-score list

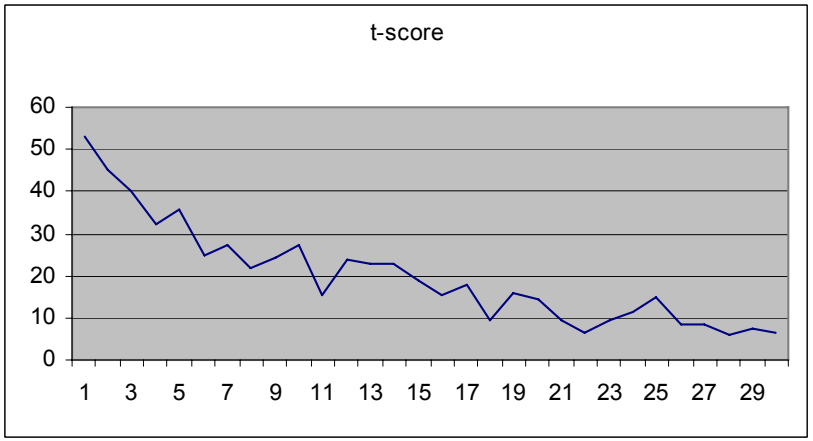


Figure 3: hit-rates at 30 bands of a ranked chi-squared list

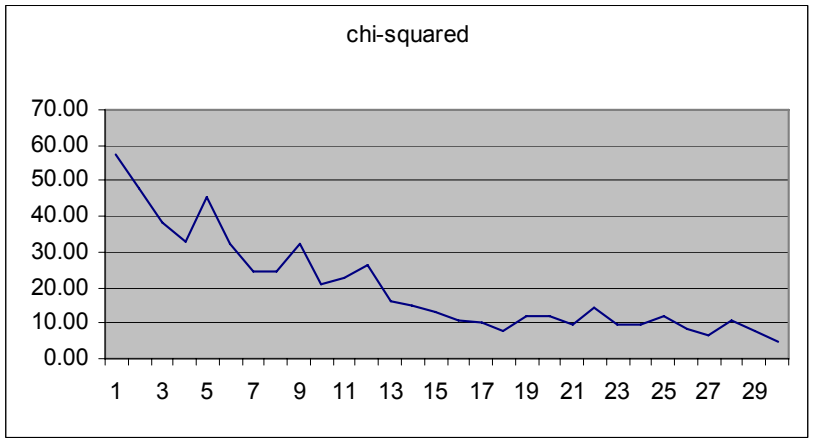


Figure 4: hit-rates at 30 bands of a ranked log-likelihood list

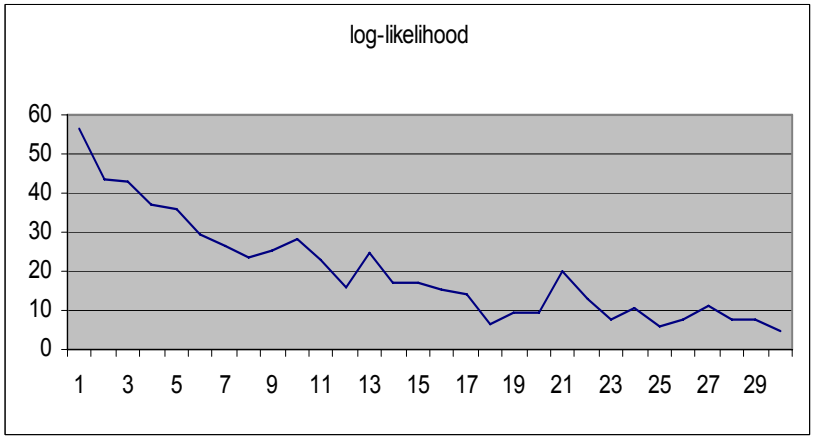


Figure 5: hit-rates at 30 bands of a ranked MI list

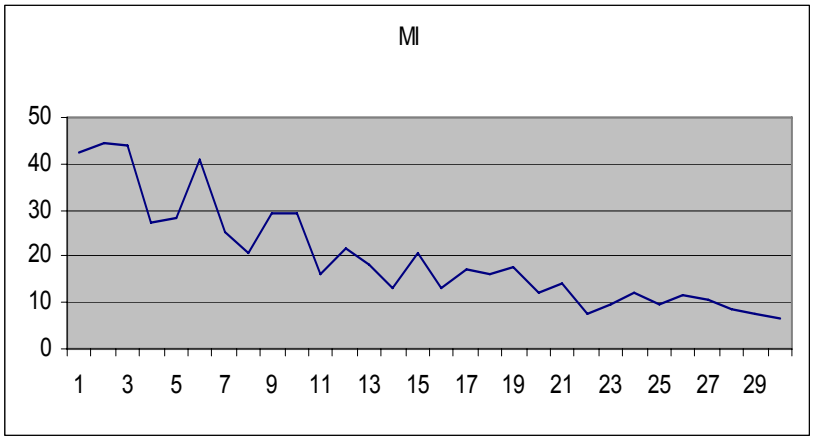


Figure 6: hit-rates at 30 bands of a ranked z-score list

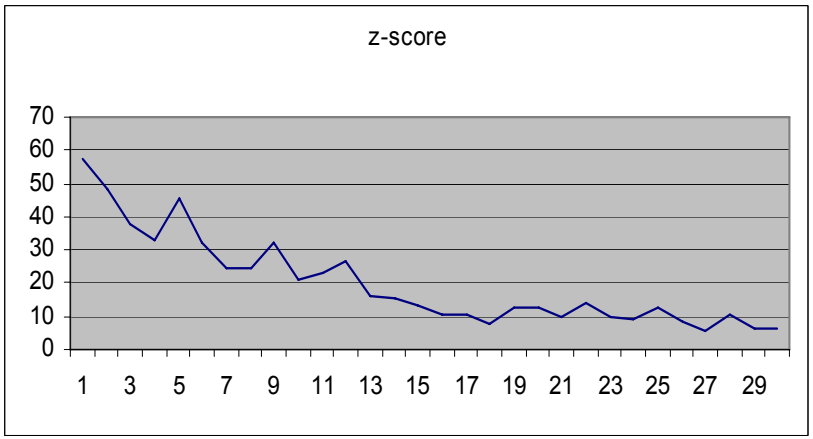
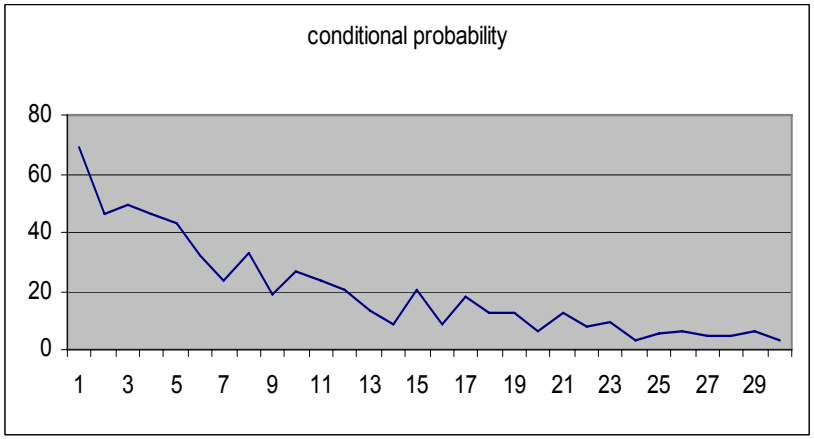


Figure 7: hit-rates at 30 bands of a ranked conditional probability list



Figures 1 to 7 show the results of this analysis. While no sharp cut-off point is evident on any of the graphs, it does seem to be possible to identify some helpful rules of thumb. Looking at the right-hand end of the graphs, the declining hit-rate tends to flatten out over the final segments, where it comes to fluctuate, apparently at random, between around 5 and 13% (the one exception being conditional probability, where rates occasionally drop to below 3%). Table 5 lists the segment number at which this flattening out occurs for each measure and the corresponding score on that measure. Below these levels, the association measures appear not to have much predictive effect. Correlations between scores and association (also shown in Table 5) are either non-existent or very weak, suggesting that the hit-rates seen here are more-or-less random. We can therefore propose that the values of the various measures listed in Table 5 are the minimum points at which they might tell us something interesting about psychological association; scores below these levels appear to tell us little if anything.

Table 5: level at which each measure becomes informative about associations

	segment	**score**	**Spearman correlation**
raw-frequency**	22	16	$r = .017$
t-score	21	3.9	$r = .030$
chi-squared	15	1520	$r = .050$*
log-likelihood	22	60	$r = .050$
MI	22	3.7	$r = .057$*
z-score	15	38	$r = .042$*
conditional probability	18	0.21	$r = .016$

*significant at $p < .05$
** occurrences in 100m words

At the other end of the scale, hit-rates tend to rise consistently with scores. The best rule here is clearly ‘the higher the better’. Table 6 shows the values above which each measure consistently achieves a hit-rate of 30% or more. Word pairs achieving scores above these levels are likely to be well worthy of attention.

Table 6: level above which each measure consistently achieves a hit-rate of 30% or better

	segment	score
raw-frequency*	3	270
t-score	7	10.3
chi-squared	7	10,297
log-likelihood	7	700
MI	3	8.3
z-score	9	101
conditional probability	6	1.16

* occurrences in 100m words

Results of a similar analysis for those associates which were reported by at least 5% of respondents are shown in figures 8 to 14. Because of the decreased probability of hitting associates at random, the flattening-out effect at the right-hand end of these graphs is rather more obvious, with most measures levelling off to hit-rates of around 0-3%. Table 7 shows the points at which this threshold is reached, again the correlation between frequency and association is lost below these levels. These would appear to be the points at which the measures can begin to give us useful information about which word pairs are very likely to be widespread associates in a population.

Table 7: level at which each measure becomes informative about robust associations

	segment	score	Spearman correlation*
raw-frequency**	22	16	$r = .009$
t-score	20	4.2	$r = .019$
chi-squared	12	3,112	$r = .026$
log-likelihood	17	142	$r = .016$
MI	16	5.0	$r = .034$
z-score	12	56	$r = .026$
conditional probability	11	0.56	$r = .036$

* no correlations are significant at $p < .05$
** occurrences in 100m words

As before, at the top end of the scale, the best rule is ‘the higher the better’. On this analysis, a hit rate of 30% is only achieved by conditional probability, and this only in its top segment (corresponding to values over 5.2). Chi-squared, MI, and z-score also achieve respectable rates of 27-30% in their highest segments (corresponding to chi-squared > 2,000; MI > 9.6; z-score > 445), while log-likelihood manages 22% (at

values > 4,890). Neither t-score nor raw-frequency manages to achieve more than an 18% hit-rate.

Figure 8: hit-rates at 30 bands of a ranked chi-squared list

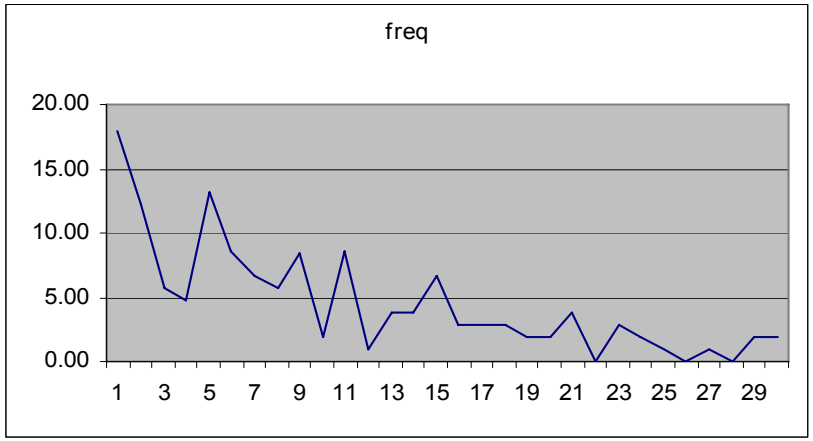


Figure 9: hit-rates at 30 bands of a ranked log-likelihood list

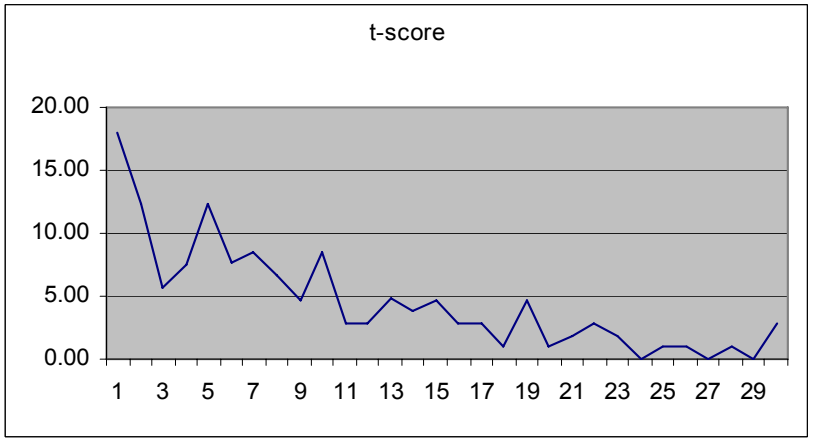


Figure 10: hit-rates at 30 bands of a ranked MI list

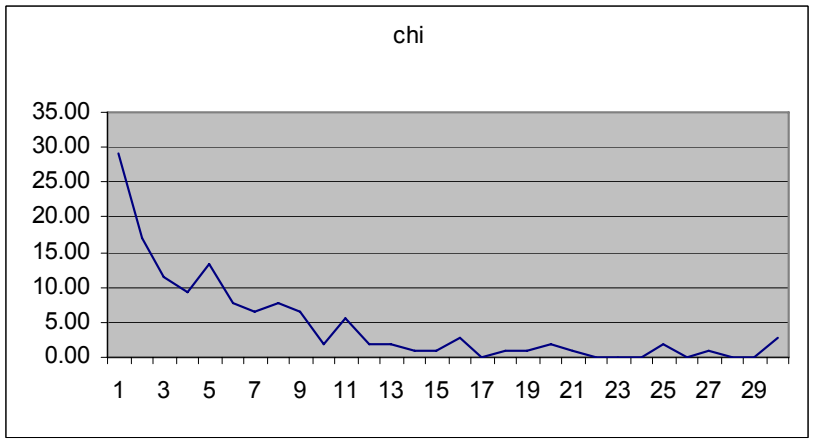


Figure 11: hit-rates at 30 bands of a ranked z-score list

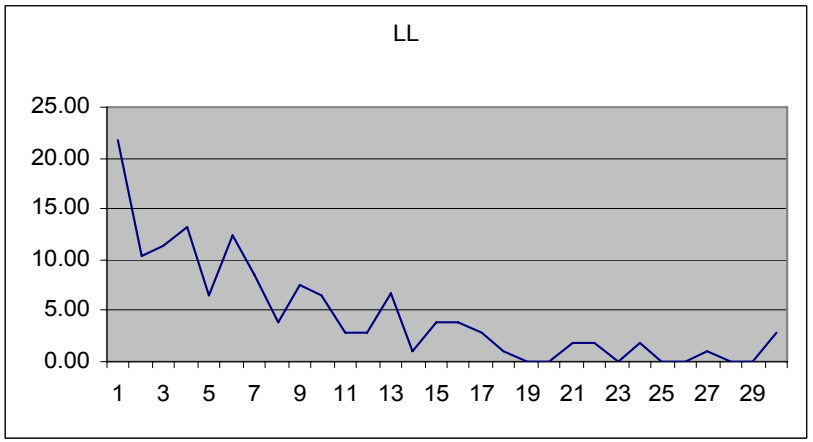


Figure 12: hit-rates at 30 bands of a ranked MI list

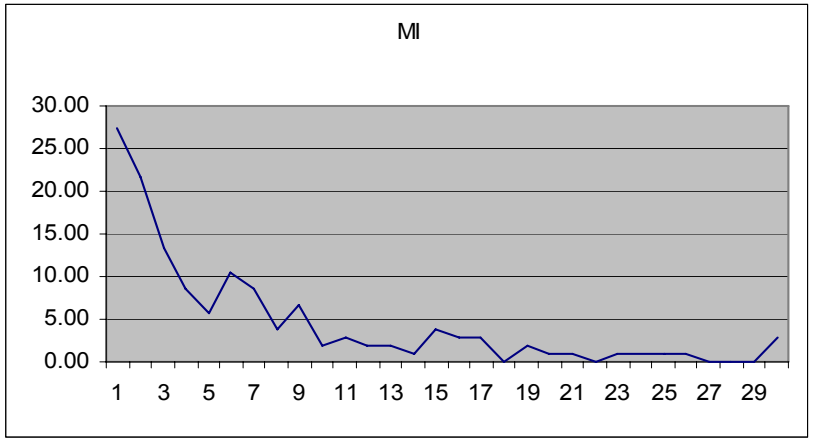


Figure 13: hit-rates at 30 bands of a ranked z-score list

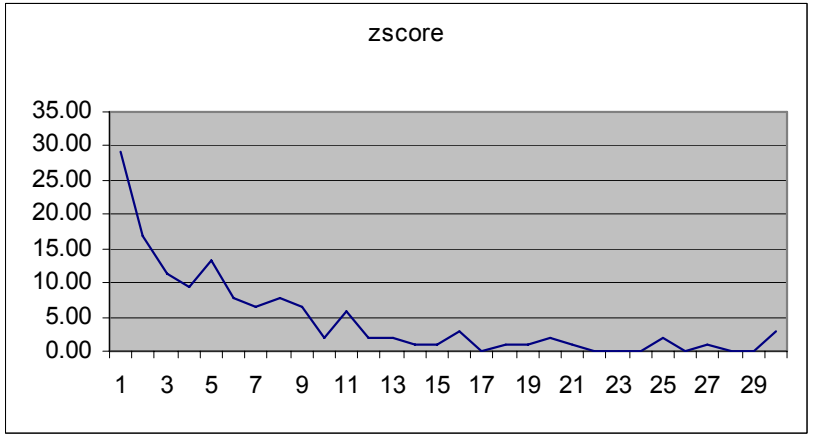
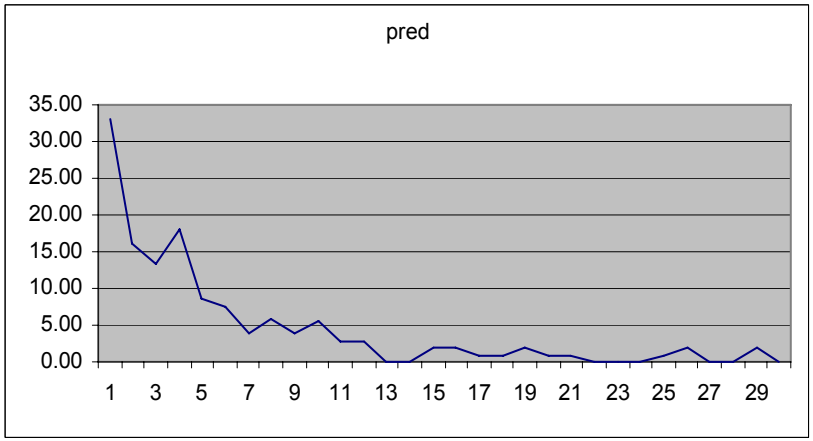


Figure 14: hit-rates at 30 bands of a ranked conditional probability list



Summary and conclusions: co-occurrence frequency and word association

This study has aimed to discover how reliably various frequency-based methods of corpus analysis identify psychologically-real collocations. I have noted that the approach used here cannot yield definitive measures of accuracy, since it is likely that many collocations which were psychologically real for the population studied were not reported in the word association norms. The hit-rates given above therefore almost certainly underestimate these measures' true accuracy. Moreover, we have been able to gauge only what information theorists call the 'precision' of the measures (i.e. the percentage of collocations suggested by each measure which are psychologically associated), without being able to say anything about their 'recall' (i.e. the percentage of psychologically associated collocations which are found by each measure). Another limitation is that we have worked with only one grammatical form: adjective-noun pairs. Different results may be found for different part of speech pairings. However, I would argue that these data are likely to provide a reasonable indication of the relative accuracy of each measure and of the levels at which they are likely to provide useful information.

We have seen that all BNC-based frequency measures are reliable predictors of psychological associations in the EAT population. Encouragingly, this suggests that collocations are sufficiently stable for generalisations from corpus counts to psychological associations to be made with a fair degree of confidence. The best predictor of psychological association appears to be conditional probability. This is followed by chi-squared and z-score, which return almost identical results. Log-likelihood is a better predictor than t-score, which in turn outperforms raw frequency. MI is a less accurate predictor of associates overall than the other measures, but is a good predictor of those associates which are robust enough to be attested by many respondents.

4.5 Do high frequency collocates 'prime' each other?

Introduction

Our study of word association norms has suggested that the frequency-based analysis of a corpus enables reasonably good predictions to be made about which word pairs

are likely to be 'psychologically real' collocations for native speakers, suggesting that corpus-derived collocations may be good targets for language learning. The studies presented in the present section aim to investigate further the relationship between frequency data and the mental representation of collocations by examining specifically Hoey's thesis of 'lexical priming'. As we saw in Section 2.1, Hoey believes that the frequent co-occurrence of words in a corpus can indicate a "psychological association between words", and that this association can be spelled out in terms of priming: when a given node word occurs, its high frequency collocates come to mind more readily than they would in other contexts (Hoey, 2005, pp. 5-8). Since priming is an extensively researched paradigm, for which well-established methodologies have been developed, Hoey's thesis is an eminently testable model of how frequency might relate to mental representations.

It will be worth investigating Hoey's thesis for a number of reasons. First, it will be of intrinsic theoretical interest to establish whether native speakers' knowledge of collocations can be validly described in terms of priming, or whether we need to look to other models to understand this phenomenon. Second, if knowledge of collocations can indeed be characterised in terms of priming, the relatively simple methodologies which are used to study priming might enable us to gain a more accurate and comprehensive picture of the relationship between corpus-frequency and the mental representation of collocations than was possible with the relatively 'noisy' method of word association. Finally, priming methodologies might provide us with a means of assessing learners' collocational knowledge and so with a useful tool for the study of language learning. This section will describe the notion of priming as it has been discussed in the psycholinguistic literature before going on to investigate existing research relating to collocational priming in particular. We shall see that, to date, no strong evidence exists to support Hoey's model. Four new studies testing the relationship between co-occurrence frequency and psychological priming will then be described.

The priming paradigm

Priming is the psychological phenomenon, first documented by Meyer and Schvaneveldt (1971), whereby the recognition of a word is facilitated by its preceding context. This is seen, for example, in the way a hearer or reader recognises a given

word faster if they have previously seen or heard a semantically related word. Thus, the word *girl* is recognised more quickly when it comes soon after the word *boy* than it does when it follows a semantically unrelated word. In such cases, the context is said to *prime* the target word.

Priming has been claimed to exist between words with similar orthographies and phonologies, between words which are related in meaning, and between syntactically congruous words (e.g. determiner - noun) (Balota, 1994, pp. 334-341). Priming is usually investigated by some variation of the two-stage task described by Neely (1991, p. 265): first, the informant is presented with a single word (the *prime*), to which they are not required to make any overt response. Second, they are presented with a letter string which may or may not be a real word (the *target*), and are required to respond either by making a word/nonword decision (the 'lexical decision task', or 'LDT'), or by saying the target aloud (the 'naming task'). When the target is a word, it is either related or unrelated to the prime. Priming is taken to exist where reaction times or percentage errors are significantly lower for those targets that are related to their primes than for those which are not.

Semantic priming in particular has received a great deal of attention in psycholinguistics, and several decades of research have identified a wealth of robust effects. Key variables which have been shown to influence priming include: the type of relationship between prime and target; the nature of the response required from subjects (LDT produces a different pattern of results from the naming task); length of stimulus onset asynchrony (the time between the first appearance of the prime and the first appearance of the target – hereafter referred to as SOA); whether subjects are consciously aware of the presence of the prime; whether the target is obscured in any way; the frequency in language of the target; whether subjects expect items to be related (either because they have been told to expect or not expect certain relations or because the ratio of related to non-related items is high enough for relations to be sufficiently prominent for such expectations to be learned on-task); whether an unrelated word is interposed between prime and target; and the depth of subjects' processing of the prime (Neely, 1991).

Evidence for collocational priming?

Much work on priming has been concerned with separating the effects of 'purely' semantic relations (the relations found between, for example, categories and their exemplars, co-hyponyms, or words with featural similarities between their concepts) from the effects of 'associative' relations (i.e. the normative psychological associations discussed in Section 4.4). The primary aim of this work has been to distinguish between models of the mental lexicon in which word form and word meaning are integrated, such that purely semantic information can influence word recognition, and models in which the two types of information are strictly separated, such that semantic information cannot feed back to recognition of a word form. The former model predicts that pure semantic priming is possible, whereas the latter predicts that it is not (Lucas, 2000, pp. 618-619). For us, the issue of semantic vs. associative priming is of special interest because it has been claimed that associative priming may be a product of the frequent co-occurrence of associated items in text (Charles & Miller, 1989). That is to say, associative priming may in fact be collocational priming.

If this were right, then the existing evidence for association priming would constitute direct experimental support for Hoey's collocational priming thesis. However, as I discussed above, though collocation and association are undoubtedly linked, they are far from identical. Re-analysis of the prime-target pairs used in studies of associative priming confirms the distinction. In Shelton and Martin's (1992) influential study, for example, pairs in the 'associated' condition have a whole range of MI-scores, from the very low (*min* = 0.90) to the moderately high (*max* = 8.85), with a median score which is above the traditional cut-off point of 3, but only modestly so (*Mdn* = 3.98). The same remarks apply to other measures of collocational significance (t-score*: min* = 1.81, *max* = 24.34, *Mdn* = 4.87; log-likelihood: *min* = -770.3, *max* = 1832.55, *Mdn* = 200.61).

A particular problem with using data on association priming as evidence for collocation priming is that those collocations which appear prominently on association norms are likely to be especially salient in a way that other collocations are not. Even if an associative priming study were based entirely on collocating items therefore, the

results may be generalisable only to the most salient of collocations. If this is the case, priming may not be a suitable paradigm for understanding collocation in general.

Studies of associative priming cannot, therefore, provide adequate support for the thesis of collocational priming. There are only a handful of studies which have aimed specifically to study collocational priming. Hodgson (1991) found priming for 'phrasal associates' (e.g. 'private-property', 'vacant-building', 'arm-chair') in a lexical decision task, but not in a naming task. However, since he does not give any information regarding how his test items were identified as phrases, and in particular does not provide any frequency data for the items, his results have little to tell us about the relationship between high frequency collocation and priming. Moreover, as Hutchison (2003, p. 789) points out, Hodgson combines his results for phrasal associates with those from five other types of relations (synonyms, antonyms, conceptual associates, co-ordinates, and super/subordinates), reporting only the overall priming effects for all six types of pairs and not the significance of each type alone. Even if corpus data were available, therefore, the study would not be able to tell us about the effects of collocational priming alone.

Williams (1996) finds significant priming for word pairs which were graded by native speaker informants as 'highly familiar' in conjuntive phrases (i.e. 'X and Y'). As in the Hodgson study, however, insufficient frequency data is provided for the nature of any link between corpus and priming to be stated. Though Williams reports a post-hoc frequency analysis of his items using the one million-word Lancaster-Oslo/Bergen corpus, he is only able to tell us that the pairs were found in the same sentence, on average, 2.94 times (range = 0-11), and within a span of +/-2, on average, 2.06 times (range (0-6). The small size of the corpus employed and the variability of the results (with some 'collocations' not being attested together at all) mean that it is again difficult to draw any strong conclusions regarding the effects of frequent co-occurrence. Moreover, all of Williams' collocations were attested to be strong associates. It is therefore possible that the priming found is limited to those collocations which are sufficiently salient to be normative associates. A potentially helpful result in this context was the finding that associates that were posited to be collocations produced a greater priming effect (18ms) than did associates which were not collocations (3ms), hinting at an independent effect of collocation over and above

that of association alone. However, the difference between the two sets, though large, was not significant. This was because, whereas the eight most strongly associated collocations produced very strong priming effects, the other eight associates showed minimal effects. Williams suggests that priming may be obtainable only for those collocates with "the highest psychological salience" (J. N. Williams, 1996, p. 133). Concluding that collocates which are sufficiently psychologically salient to produce priming are likely also to be strong normative associates, he abandons any attempt to separate the two effects (1996, p. 134).

Unlike Hodgson and Williams, McKoon and Ratcliff (1992 - experiment 3) used corpus-derived frequency data in compiling collocations for their priming experiment. They compared priming between associated primes with that between collocating pairs. Collocates were identified as such on the basis of MI scores calculated from frequencies of occurrence in a 6 million word AP newswire corpus. McKoon and Ratcliff also considered the priming effects of collocates with low MI scores. The study found that the greatest facilitation in comparison with an unrelated prime was between associated primes (49ms); significant priming was also found in the collocating condition (21ms); collocates with low MI also speeded recognition (17ms) though this improvement on the unrelated condition was not significant.

While acknowledging that the small size of their corpus must bring the reliability of their findings into some doubt, McKoon and Ratcliff tentatively conclude that "co-occurrence statistics calculated from large corpora have potential applicability as predictors of priming effects" (1992, p. 1164). However, a re-evaluation of their 'high' MI collocations against BNC data, confirms their worries regarding reliability. Indeed, the collocations in this study are no stronger (and if anything, rather weaker) than those found in Shelton and Martin's study (MI: *min* = -3.26; max = 9.64; *Mdn* = 3.43; t-score: *min* = -8.58; *max* = 25.86; *Mdn* = 3.90; log-likelihood: *min* = -35.79; *max* = 2969; *Mdn* = 57.85). Moreover, as in Williams' study, many word pairs used in McKoon and Ratcliff's collocating condition were – as the authors acknowledge - also strong associates of the targets. Again then, it is not clear whether the priming effects found are attributable to collocations in general or are restricted to normative associates.

In sum, while some psycholinguists have suggested that collocational priming exists, there is as yet no strong evidence for this. Previous studies have failed to provide sufficiently robust frequency data to evaluate any claim of a link between frequency of occurrence in a corpus and priming, and have failed to demonstrate that such priming extends beyond the highly salient examples that appear on normative association listings. The studies presented here aim to make up for these shortcomings by exploring the extent to which frequency-based collocational preferences identified in the BNC are reflected in the primings of native speakers of British English, regardless of whether they are strong normative associates.

Study One

Introduction

Study One investigates whether very strong collocations exhibit priming by comparing the time taken to recognise target nouns when they follow primes with which they are frequently collocated (as attested in BNC data) with the time taken to recognise them when they follow words with which they do not commonly co-occur.

Materials

On the basis of frequency data from the BNC, two sets of 40 prime-target word pairs were created. One set (the 'collocate condition') comprised pre-modifier primes (adjectives and noun-modifiers) and target nouns which appear directly adjacent to each other in the corpus with sufficient frequency to qualify as strong collocations. The second set (the 'unrelated condition') was created by re-arranging these primes and targets to form new pairs which are not found directly adjacent to each other in the corpus. Since priming seems most likely to take place when the appearance of the first word strongly predicts the appearance of the second, mutual information – which measures just this predictability – was used to gauge strength of collocation. Though we saw in Section 4.4 that conditional probability provides a slightly more accurate prediction of normative word association than MI, the latter measure was preferred as the primary criterion for item selection because it is so widely used in the corpus-linguistic literature. An analysis based on MI is therefore likely to be more meaningful to the corpus linguistic community as a whole.

Since mutual information can become unreliable for low frequency items, pairs were only included if they were attested at least ten times in ten different texts within the BNC. These criteria should also help to ensure that all collocations selected are known to the majority of speakers, rather than reflecting the usage of specialist groups. All prime-target pairs in the collocate condition had a mutual information score of at least 8.16 (*max* = 12.41, *Mdn* = 9.28) and a t-score of at least 3.49 (*max* = 42.66, *Mdn* = 10.78). Only words of between four and seven letters in length were included in the lists, and no very low frequency words were included. All primes are attested in the BNC between 1,000 and 20,000 times; all targets are attested between 1,000 and 37,000 times.

For each prime word, two non-words of between four and seven letters were also generated using the ARC Nonword Database (Rastle, Harrington, & Coltheat, 2002). Each priming word was paired with two nonwords. This provided 160 prime-target pairs in all: 80 word-word pairs, of which 40 were unrelated and 40 collocating pairs, with each target and each prime word appearing once in each condition; and 80 word-nonword pairs, with each prime appearing twice. This list was divided into two blocks, in which each target appeared once and each prime appeared four times – once in the related, once in the collocate and twice in the nonword conditions. The two blocks are shown in Appendix Ai.

Participants

22 students at the University of Nottingham. All were native speakers of British English.

Procedure

Participants were tested individually in a quiet room. Presentation of the stimuli and recording of the reaction times were controlled by Psychology Software Tools' *E-Prime* software and items were displayed on a CRT monitor. On each trial, a fixation point ('+') was presented, centred on the screen, for 2,000ms. This was replaced with the priming word, which was presented in lowercase letters for 600ms. The prime was then immediately replaced by the target, also in lowercase letters. The target stayed on the screen until the participant made a response or for 2,000ms, whichever was sooner. Participants were instructed to press 'W' on the computer keyboard if a target was a

word and 'N' if it was a nonword. They were told to make this decision as quickly as possible. Reaction times were measured from target onset to response. Participants received 30 practice trials, sampled from across the conditions. There was then a break, at which point participants were invited to begin the trial proper in their own time. They were then presented with one block of 80 trials before a second break and then the second block of 80 trials. Order of presentation within blocks was pseudo-random. An initially randomised order was adjusted to ensure that for each of the four conditions (prime-related target; prime-unrelated target; two x prime-non-word) there was an equal number of primes appearing for the first, second, third and fourth time in the block. This was to offset any possible effect of repetition priming between primes. The order of presentation of the two blocks was counterbalanced across participants.

Results and discussion

Reaction times of less that 250ms and greater and 1,250ms (2.98% of the total) were excluded from analysis. Mean accuracy by participant, collapsed across conditions was 96%. There was no effect of condition upon accuracy.

Reaction times for correct responses (which were not normally distributed within conditions) were faster for collocating (*Mdn* = 546ms) than for unrelated targets (*Mdn* = 567ms). This difference was significant both in the analysis by participants (T = 28.00, p (one-tailed) < .001, r = -.68) and in the analysis by items (U = 573.50, p (one-tailed) < .05, r = -.24). Collocating primes appear therefore to substantially facilitate the recognition of the target words.

As was discussed above, it is already well known that normatively associated words show a significant priming effect. I have argued that, though normative association and priming are linked, the existing evidence for the former cannot count as evidence for the latter since, on the one hand, many associates are not collocations and, on the other hand, only the most salient of collocations are prominent in word association norms. The present study overcomes the first of these problems: all of the items used here are collocations. The issue of whether priming works for collocations which are not sufficiently salient to appear prominently in word association norms has not so far been addressed, however. The collocating items used in this study were selected purely on the basis of corpus evidence, without reference to association norms. A

post-hoc search for these collocations in the Edinburgh Association Thesaurus (EAT) <http://www.eat.rl.ac.uk/> found that 21 of the 40 collocating pairs in our study are attested as associates, with 10 of these being supplied by at least 5% of respondents. 13 collocations are not attested as associates, and 6 pairs are not classifiable because the prime is not listed in the thesaurus (these classifications are indicated in Appendix Ai). None of the non-collocating pairs were found to be associates.

It should be noted that a pair's failure to appear in the Edinburgh norms does not show that the words are not strong associates for any of our participants. The Edinburgh database was not collected from the same population of speakers as took part in this experiment, and – as was discussed above – the methodology used to elicit this norming set is likely to have found only the strongest associates. However, it seems likely that word pairs not appearing here will at least not be amongst the most salient associates for the majority of speakers of English. The putatively 'non-associated' pairings were:

death-penalty
deep-sigh
foreign-affairs
huge-amounts
peace-talks
private-sector
wild-flowers
east-coast
fatal-error
liquid-assets
officer-corps
silent-movies
silk-shirt

In light of this information, we can re-analyse the data to ask whether priming is restricted to collocations which are sufficiently salient to feature in the association norms. To this end, items were re-organised into three groups: unrelated pairs, collocations which are attested as associates, and collocations which are not attested

as associates. Median reaction times for correct responses in each condition are shown averaged across participants in Table 8.

Table 8: median lexical decision times for accurate responses

	unrelated	associated collocates	non-associated collocates
median RT (ms)	567	534	512

Significant facilitation was found in comparison to the unrelated condition for both associated (T = 30.00, p (one-tailed) < .001, r = .64) and non-associated collocations (T = 43.00, p (one-tailed) < .005, r = .58). However, probably as a result of the small number of items involved, neither of these advantages is statistically significant in an analysis by items (unrelated vs. associated collocates: U = 326.50, p (one-tailed) > .05, r = -0.18; unrelated vs. non-associated collocates: U = 206.00, p (one-tailed) > .05, r = -0.15). While the small number of items – and the imperfect nature of the EAT as a check for psychological association in our participant group - must weaken our conclusion somewhat then, we can tentatively conclude that priming may be taking place here even when collocations are not sufficiently salient to feature on normative association lists.

Study Two

Introduction

Study One showed that a priming effect can be found between extremely strong collocates (MI > 8). It also suggested that collocational priming may exist independently of whether collocates are salient enough to be listed as psychological associates, though some reservations remain about this conclusion. The present study aims to extend these findings by considering a) whether similar effects can be found for less extreme examples of significant collocation; b) whether this effect is indeed independent of association; and c) whether the effect is independent of mere semantic congruence.

The thinking behind question c) is that the collocating word pairs used in the previous study (e.g., *death-penalty; elder-brother*) form meaningful phrases, whereas the control pairs are often difficult to make sense of (e.g. *east-penalty; leather-brother*). It seems possible that the second part of a word pair which has a plausible meaning may

be recognised faster than the second part of a pair which is highly implausible, regardless of whether the pair forms a significant collocation, since semantic congruence alone will provide the subject with evidence that the target is indeed a word. It may be, then, that the priming effects found in Study One are due to mere semantic congruence, rather than to frequency of collocation.

To answer these three questions, the present study measures priming between three levels of collocation:

- moderately strong collocations which are not strong normative associates;
- strong collocations which are not strong normative associates;
- strong collocations which are strong normative associates.

It uses for non-collocating controls only word combinations which are attested more than once in the BNC, and so which are presumably semantically plausible.

The inclusion of moderate strength collocations is intended to address question a); more rigorous tests of association are introduced (see below) to address question b); and the use of attested controls is designed to address question c)

Materials

16 collocating prime-target pairs were created for each of three conditions:

- Level 1 pairs are moderately strong collocations, having MI scores of 4-5 (*Mdn* = 4.47) and t-scores of 4-8 (*Mdn* = 5.52), and are not strong normative associates;
- Level 2 pairs are strong collocations, having MI scores of more than 6 (*Mdn* = 7.65) and t-scores of more than 7.5 (*Mdn* = 10.95), and are not strong normative associates;
- Level 3 pairs are strong collocations, having MI scores of more than 5.5 (*Mdn* = 7.01) and t-scores of more than 6 (*Mdn* = 10.63), and are strong normative associates.

Frequency measures were again derived from BNC frequency data. To determine whether pairs were strong normative associates or not, two methods were used. First, as in Study One, the EAT was consulted. Pairs at levels 1 and 2 were deemed not to be strong associates only if neither the target word nor any word form derivationally or inflectionally related to the target was listed as an associate of the prime in the Thesaurus; similarly, neither the priming word nor any word form derivationally or inflectionally related to the prime was listed as an associate of the target. Pairs at level 3 were judged to be strong associates only if the target word was listed as either the first or second strongest associate of the prime in the Thesaurus and had a minimum association score of 10% (i.e. was supplied by 10 out of 100 respondents).

It was noted above that, because the EAT was elicited from a different population from that taking part in the present study (university students between 1968 and 1971), word pairs which are not attested as associates in the EAT may nevertheless be strongly associated for our participants (university students in 2007). Similarly, some pairs which are prominent on the EAT may not be strong associates for these participants. Moreover, because the EAT elicited only a single response from each participant, it is possible that some highly salient collocations do not feature on its listings. To ensure that the putatively 'non-associated' pairs in the current study were indeed not strong associates for our participants, and that the putatively associated pairs were, a second test for association was used in the present study. Two groups of 22 subjects (who did not take part in the main study) were each presented with 40 stimulus words (a different stimulus list for each group, giving a total of 80 stimuli) and asked to write down the first three words which came to mind on reading each stimulus. The subjects in these groups were taken from the same pool as those participating in the main priming study (i.e. 2nd year undergraduate native speakers of British English enrolled in a modern English language course at the University of Nottingham), and so should provide a good indication of the likely associates of the main study participants. Moreover, by eliciting three associates for each stimulus, we may move a little beyond the very strongest associates. Pairs at Levels 1 and 2 were deemed not to be strong associates only if neither the target word nor any word forms derivationally-related to the target was supplied as an associate of the prime. Pairs at Level 3 were judged to be strong associates only if the target was supplied as an

associate of the prime by at least two respondents (the median association score was 43% - i.e. the association was given by 9.5 out of 22 subjects).

The 48 target nouns from the three collocation lists were also matched with 48 control primes. The control prime-target pairs were intended to be semantically plausible combinations which did not co-occur in the corpus with sufficient frequency to be considered collocations. These pairs occurred directly adjacently to each other in the between BNC two and four times, with MI scores of less than 2.5 and t-scores of less than 1.5. All pairs co-occurred within a +/- 4-word span of each other fewer than 10 times in the 100 million word corpus. Though they are not sufficiently frequent to count as collocations, therefore, the fact that they were all attested more than once in the corpus suggests that they are not likely to be semantically anomalous.

No very common or very rare words were used as targets or as primes: all words used occurred in the BNC between 3,000 and 30,000 times; placing them well outside the top 300 word forms in the corpus and well within the top 3,500 (Leech, Rayson, & Wilson, 2001). All words were one or two syllables (four to seven letters) in length.

The collocating and control primes were combined into two counterbalanced lists – referred to below as Set 1 and Set 2 - such that eight collocating pairs from each level were included in each list and targets which were matched with their collocating prime in one list were matched with their control prime in the other. No prime or target word was used more than once in either list. A single set of 48 prime-non-word pairs was also added to both lists. Non-words were items of four to seven letters, generated using the ARC Nonword Database (Rastle et al., 2002). Primes were items which appeared in the BNC between 3,000 and 30,000 times and which were attested to be used as pre-modifiers but which had not been used elsewhere in the experiment. The final materials are shown in Appendix Aii.

Participants

40 undergraduate students at the University of Nottingham participating in a course in modern English language. All were native speakers of British English.

Procedure

Participants were randomly assigned to one of two groups. The first group was tested on Set 1 only, the second group on Set 2 only. Participants were tested individually in a quiet room. Presentation of the stimuli and recording of the reaction times were controlled by Psychology Software Tools' *E-Prime* software and items were displayed on a CRT monitor. On each trial, a fixation point ('+') was presented, centred on the screen, for 1,500ms. This was replaced with a priming word, which was presented in lowercase letters for 600ms. The prime was then immediately replaced by the target, in uppercase letters. The target stayed on the screen until the participant made a response. Following the response, the screen went blank for 1,000ms before the onset of the next trial. Participants were instructed to press the right button on a button-box if the string was a word and the left button if it was not**.** They were told to make this decision as quickly as possible. Reaction times were measured from target onset to response. Participants received 10 practice trials, there was then a break, at which point participants were invited to begin the trial proper in their own time. They were then presented with the appropriate list of 96 trial items, presented in random order.

Results and discussion

Reaction times of less that 250ms and greater and 1250ms (1.33% of the total) were excluded from analysis. Mean accuracy by participant, collapsed across conditions was 97%. There was no effect of condition upon accuracy. Average reaction times for collocations and non-collocations at each of the three levels are shown in Table 9. Reaction times were normally distributed within conditions for Levels 1 and 2 but not for Level 3. Paired samples t-tests (for Levels 1 and 2) and a Wilcoxon signed rank test (for Level 3) revealed no significant differences at the $p < .05$ level between reaction times for the collocating and non-collocating conditions at any level.

Table 9: average reactions times in each condition

	collocations	**non-collocations**
Level 1 Mean RT (ms)	514	519
Level 2 Mean RT (ms)	516	530
Level 3 Median RT (ms)	479	490

Study One appeared to show that target words are recognised more quickly when they follow collocating primes than when they follow non-collocating primes. However,

that study compared collocating prime-target pairs only with semantically-incongruous pairs (e.g. *death-penalty* vs. *east-penalty*). This left open the possibility that the advantage seen was due not to the relatively higher frequency of the collocating pairs but to the fact that they were semantically congruous, whereas the control items were, in general, not. The present results suggest that this may indeed have been the case. High frequency collocations – regardless of whether they are likely to be strong associates - appear not to demonstrate any statistically robust priming effect in comparison to low frequency but semantically plausible word pairs.

Some concerns might be raised by our failure to find any priming between pairs which are likely to be strong psychological associates (i.e. Level 3 pairs). As we have seen, priming between such pairs has been extensively documented. However, this finding appears to have been based entirely on comparisons of the sort seen in Study One – i.e. comparison between associates and semantically incongruous word pairs. All of the associative priming studies with which I am familiar and which report their method of creating 'unrelated pairs' have followed the lead of Meyer and Schvanevendlt's original paper (1971) by interchanging words from the associated items such that there are "no obvious associations within the resulting pairs" (1971, p. 228). These studies tend not to list the items used in the control condition, but Meyer and Schvaneveldt provide the illustrative examples of BREAD-BUTTER and DOCTOR-NURSE being re-paired to BREAD-DOCTOR and NURSE-BUTTER. As in Study One, then, what the associative priming literature demonstrates is an advantage of associated pairs over pairs which are highly incongruous. The findings in this literature are therefore entirely consistent with the results reported here.

We can conclude, then, that though it is possible to observe collocational priming in experimental set-ups which use semantically incongruous control items (such as that in Study One), the effect is not replicated when more plausible controls are used. This applies even to very high frequency collocations which are not normative associates. Since much of the interest of collocation is in the distinction between collocating word pairs and pairs which are plausible but uncommon, these considerations suggest that the traditional priming paradigm may not be a helpful way to study collocations in the mind.

Study Three

Introduction

A number of points remain unclear from the studies presented so far. Study One suggested that priming may occur between both associated and non-associated collocations. However, the small number of 'non-associated' collocations used in that study left this conclusion in need of further support. Moreover, Study One used only very high frequency collocations. It is still not clear, therefore, if priming can also be found between more moderate-strength collocations or if it is particular to these extreme cases. Study Two attempted to address these issues, but failed to find any evidence of priming. I have suggested that the failure to find any priming in this study was due to the fact that semantically-plausible control items were used, but this interpretation stands in need of direct experimental confirmation. Furthermore, if it is true that using semantically-plausible controls attenuates priming effect, this suggests that semantically-plausible non-collocations may themselves exhibit priming, in comparison to less plausible word pairings. The existence of priming between such items would be of considerable interest, suggesting that priming is at least in part a result of participants' being able to construct a semantically-plausible context for a word pair, rather than simply a matter of co-occurrence frequencies. This possibility also requires further investigation.

The present study aims to clarify these issues. The experiment reported here is similar to that in Study Two, but differs in that it uses non-attested word pairs as controls and adds an additional level of attested, but low frequency, pairs as test items. The use of non-attested controls aims to test whether the failure to find priming in Study Two was due to our use of attested controls in that study. Since the procedure and collocating items used in the current study remain the same as those in Study Two, if we find any evidence of priming here, we can conclude that our previous failure to find priming was indeed the result of using attested controls. It will also enable us to ask again whether priming exists for moderate-strength collocations and for collocations which are not normative associates, as well as for very high frequency, associated collocations. The addition of a level of attested, but low frequency, word pairs aims to test whether semantically-plausible pairs which are not collocations themselves demonstrate priming.

Materials

The present study will compare facilitation between four types of word combinations, relative to that between non-attested pairs:

- Level 0 pairs are low frequency pairs, being attested in the BNC between two and four times and having MI scores of less than 2 and t-scores of less than 1.5;
- Level 1 pairs are moderately strong collocations, having MI scores of 4-5 (*Mdn* = 4.47) and t-scores of 4-8 (*Mdn* = 5.52), and are not strong normative associates;
- Level 2 pairs are strong collocations, having MI scores of more than 6 (*Mdn* = 7.65) and t-scores of more than 7.5 (*Mdn* = 10.95), and are not strong normative associates;
- Level 3 pairs are strong collocations, having MI scores of more than 5.5 (*Mdn* = 7.01) and t-scores of more than 6 (*Mdn* = 10.63), and are strong normative associates.

As in Study Two, there are 16 test pairs for each level. The test items for levels 1-3 are identical to those used in Study Two. Control items were created by re-pairing adjective-noun combinations from the test items such that the new pairs are not attested as occurring adjacent to each other in the BNC. The final lists used in this study are shown in Appendix Aiii.

Participants

32 students at the University of Nottingham. All were native speakers of British English.

Procedure

The procedure for this experiment was similar to that in Study Two. The only difference was that, since the extra experimental level meant that a larger number of items were used, items were presented in two blocks, with a self-paced break between blocks. Each block contained an equal number of items from each level, and the order of presentation of the blocks was counterbalanced between participants.

Results and discussion

Reaction times of less that 250ms and greater and 1,250ms (4.3% of the total) were excluded from analysis. Mean accuracy by participant, collapsed across conditions was 96%. There was no effect of condition upon accuracy. Average reaction times for collocations and non-collocations at each of the three levels are shown in Table 12. Reaction times were not normally distributed within conditions.

Table 12: average reactions times in each condition

	collocations	**non-collocations**
Level 0 Median RT (ms)	522.23	528.82
Level 1 Median RT (ms)	511.97	525.07
Level 2 Median RT (ms)	520.50	521.32
Level 3 Median RT (ms)	506.75	529.07

Though median reaction times were somewhat lower for collocating than for non-collocating pairs at all levels, this difference was statistically significant only at level 3: i.e. between normatively associated word pairs, where a strong and highly significant facilitation effect was found (analysis by participants: $T = 72.0$, $p < .001$, $r = -.45$; analysis by items: $T = 0$, $p < .001$, $r = -.62$).

This result clarifies a number of points which had remained ambiguous after Studies One and Two. Firstly, it appears that high frequency collocations which are not strong normative associates (i.e. items at Levels 1 and 2) do not prime each other in the same way as do pairs which are strongly associated. The priming seen in Study One may therefore have been an effect restricted to the few items used in that study. Secondly, it appears that our failure in Study Two to find priming between strongly associated pairs was due to the nature of the control items used. It seems that priming can be found between associated pairs only when relatively implausible pairs are used as controls. Finally, low frequency but semantically plausible pairs do not appear themselves to exhibit priming in these circumstances. This result is unsurprising in light of the fact that priming was also not detected between much higher frequency pairs, unless they were psychological associates.

Study Four

Introduction

A possible limitation of the studies presented so far is that they do not preclude the intervention of 'strategic' processes on the part of subjects. That is to say, it is possible that subjects attempted to find relationships between primes and targets and adopted task-specific strategies in light of their hypotheses. There have been two main suggestions in the priming literature as to how such an effect might work (Neely, 1991, pp. 299-317). One is that, on encountering a prime, subjects generate a list of probable targets. When the actual target appears, it is first checked against the list of likely candidates. If a match is found, then a word decision can be made more rapidly than under normal circumstances. Words which are not included on this list, however, are recognised more slowly than usual since the list-checking process needs to be completed before normal processes of word recognition can begin. A second model suggests that subjects might adopt a strategy of 'post-lexical checking'. Since targets which are related to the prime must of necessity be words, subjects could base their word-nonword decision partly on the presence or absence of a relationship between target and prime. If a high proportion of words are related to the target, this will be an effective time-saving strategy. However, recognition of non-related words will also be slower than under normal circumstances, since the absence of a relationship to the prime will initially mislead subjects.

Three main types of evidence have been thought to indicate the presence of such 'strategic' processes (Shelton & Martin, 1992). The first is the 'relatedness proportion effect'. This is the finding that a higher proportion of related pairs amongst the stimuli - and thus a prime-target relationship which is more obvious to subjects - leads to increased priming effects. This suggests that the more likely subjects are to become aware of the relationship under consideration, the greater the priming, a finding which points to a degree of conscious control. The second piece of evidence for an effect of strategic processes is so-called 'backward priming', where facilitation is seen between pairs in which the prime is an associate of the target but the target is not an associate of the prime (e.g. prime: *cut*; target: *crew*). This effect is thought to show that post-lexical checking must be taking place. Finally, some studies have found not only facilitation for targets following associated primes, but also inhibition for targets following non-associated primes. This effect is explicable on the strategic models

outlined in the previous paragraph, but should not take place if priming is purely automatic.

It has been suggested that only automatic processes reflect the long-lasting organisation of the lexicon, whereas strategic processes are merely ad-hoc products of the experimental task, controlled by higher-order mental faculties, rather than by lexical organisation (Lucas, 2000, p. 619). Moreover, psychologists with an interest in collocation have suggested that links between collocates are likely to be implicit – i.e. not always accessible to conscious awareness, but demonstrable in performance (Ellis, 2002b). It is possible, therefore, that if strategic processes were in operation during the studies reported here, they may have obscured the operation of collocational priming. In particular, if our subjects attempted consciously to predict target words, this may have biased strongly associated collocations, which are by definition pairs which bring each other consciously to mind.

A second limitation of the studies described so far is that they have dealt only with one grammatical type: modifier-noun combinations. Pairs of this type have been used because they are very common in natural language and because they are relatively 'fixed' as directly adjacent pairs. That is to say, modifier-noun collocates almost always occur in the same order and immediately next to each other; this is not in general true of other collocation types, such as verb-noun combinations. It was thought that this relative consistency would increase the chances of finding a priming effect. However, it may also be that the consistency of grammatical form encourages participants to try to predict the target words, an effect which might – as the previous paragraphs have discussed – obscure primings between collocations which are not highly salient.

The present study attempts to overcome these limitations by using a methodology which is thought to access only automatic priming, and by using a more diverse set of collocations.

Two principle word recognition methodologies have been developed to identify purely automatic priming. The first works by abandoning the traditional arrangement in which items are presented explicitly as pairs, with a passively-observed prime

followed by a target to which a response is required. This obvious pairing-up, the logic goes, serves to emphasise the relationships between items, so making it easier for subjects to form response strategies. McNamara & Altarriba (1988) found that by presenting words one at a time, and requiring a response to each, such that primes and targets are less explicitly paired, they were able to obtain a 'mediated priming' effect – that is, priming of a target which is related to the prime only via a third word (e.g. *lion* primes *stripes* via the mediating associate *tiger*). It has been argued that such priming must be the result of automatic processes, since it is unlikely that strategic processes could make the necessary connections quickly enough. Shelton and Martin (1992) combined this single-presentation methodology with the use of a stimulus list in which only a small proportion of stimuli were related and also obtained results which they claimed were indicative of automatic priming: mediated priming was found without any inhibition between unrelated pairs and no backward priming effect was found.

The second methodology which is claimed to tap automatic priming makes use of very short stimulus onset asynchronies (SOAs), such that subjects do not have time to form conscious expectations about the target word. The effect of SOA on the automatic/strategic nature of priming was demonstrated by Neely (1977), who obtained a pattern of results which strongly suggest that trials with longer SOAs elicit strategic priming, while trials with shorter SOAs elicit automatic priming: as SOA decreases, so inhibition between unrelated items in comparison to a neutral condition decreases, while facilitation between related items remains constant; in cases where subjects were led to consciously expect targets related to a category other than the prime (e.g. primes from the category BODY PART are usually followed by targets from the category BUILDING PART), at long SOAs, facilitation was found between such pairs, while inhibition was found between unexpected but semantically related pairs (such as BODY PART-*heart*); at shorter SOAs, there was no facilitation or inhibition between the expected/unrelated pairs but there was significant facilitation between unexpected but semantically related pairs. A more recent elaboration on this method of eliciting automatic priming is to combine very short SOAs with 'pattern masks' (e.g., ######) before and/or after the prime. When such a technique is used, subjects are not usually consciously aware of the presence of the prime, so presumably cannot make use of conscious strategies. Several studies have found evidence of

semantic and/or associative priming under these conditions, and have concluded that the priming involved must be automatic (de Groot & Nas, 1991; Perea & Rosa, 2002; Sereno, 1991).

Of the two methodologies, it seems likely that the masked prime approach will be the better suited to studying collocational priming. All of the methodologies used in the current set of studies have the weakness that they present words as individual items, an approach which may discourage normal collocational processing (see *Summary and conclusions*, below, for further discussion of this point). However, by encouraging an explicit response from participants to every word seen, the single-presentation paradigm emphasises this possible isolating effect to a still greater extent. The current study will therefore use a masked prime methodology to look for automatic priming between collocations.

Materials

32 associated prime-target pairs were selected from those shown to demonstrate automatic priming in the studies of Sereno (1991) and Shelton and Martin (1992). Pairs were selected from the lists used in those studies on two conditions: 1) the target was confirmed as the primary associate of the prime in the Edinburgh Association Thesaurus (EAT); 2) frequently collocating primes could be identified which were not attested as (forward or backward) associates of the target in the EAT.

For each of the 32 targets, a frequently collocating priming word was identified which was not a strong normative associate. These were words which appeared frequently in the BNC within a span of four words to the left of the target. Collocates were taken not to be strongly associated with a prime only if, within the EAT, the target was not attested as an associate of the collocate and the collocate was not attested as an associate of the prime (the possible shortcomings of EAT as the sole test of association have been pointed out above, and the implications of this for the present study will be discussed further below).

This method of selection yielded a rather different set of word combinations from those used in Studies One to Three. Because priming words were selected if they frequently appeared anywhere within a four word span to the left of the target, rather

than only directly prior to it, and because no part of speech criteria were used, they include not only modifier-noun combinations such as those used above (e.g. *six-foot; triple-jump; big-apple*) but also combinations of other parts of speech (e.g., *turned-sour; worked-hard*) and non-adjacent collocations (e.g. *middle-night; parked-street; stretch-river*). As was discussed above, this more diverse set of items should further discourage strategic processing.

The frequencies of prime-target pairs in the associating and collocating conditions were quantified according to raw frequency of co-occurrence, t-score and mutual information. The calculation of association measures was carried out in two different ways: according to the number of times the prime was found within four words to the left of the target, and according to the number of times the prime was found within a span of four words to the left or right of the target. Since we are interested in how well the target predicts the prime, I would argue that the former is likely to provide the more relevant information. The range of BNC-based collocation strengths for pairs in the two conditions on each measure is shown in Table 13.

Table 13: frequency data for items in the associate and collocate conditions

		occurrences within 4 words to the left			**occurrences within 4 words to left or right**		
		raw freq.	**t-score**	**MI**	**raw freq.**	**t-score**	**MI**
associates	**max**	545	19.65	11.19	755	19.75	11.23
	min	1	-4.37	-1.82	2	-4.19	-1.77
	median	38	5.09	3.30	74	5.14	3.34
collocations	**max**	561	23.22	7.81	574	23.04	7.85
	min	13	3.57	3.58	14	3.57	3.24
	median	34	5.68	5.19	47	5.85	5.10

Associates are, as we would expect, often strong collocates, and on many of the measures have a higher maximum score than do the collocations. However, association is often not the result of collocation, so the list of associates also contains many pairs which are not collocations. For this reason, the minimum association scores on every measure are much lower for associates than they are for collocations, which never drop below the traditional thresholds of t-score = 2 and MI = 3. Wicoxon signed-rank tests (see Table 14) reveal no significant differences between the frequencies or t-scores of associates vs. collocations, while the MI scores of the

collocations were found to be significantly higher that those of the associates. If collocational priming occurs independently of association, therefore, items in the collocation condition should exhibit at least as much – and possibly more – priming than those in the associate condition.

Table 14: Wilcoxon's signed ranks test comparing the frequencies of items in the associates vs pure collocates conditions

	occurrences within 4 words to the left			**occurrences within 4 words to left or right**		
	raw freq.	**t-score**	**MI**	**raw freq.**	**t-score**	**MI**
Wilcoxon signed ranks test	$T = 224.5$, $p > .05$, $r = .13$	$T = 182.5$, $p > .05$, $r = .27$	$T = 159.0$, $p < .05$, $r = .35$	$T = 223.5$, $p > .05$, $r = .13$	$T = 176.0$, $p > .05$, $r = .29$	$T = 157.0$, $p < .05$, $r = .35$

An unrelated prime was also assigned to each target. These were words which, according to the Edinburgh Association Thesaurus, are not associated with the prime and which were either never or very infrequently found in the BNC within a 4-word span to left or right of the target. 32 nonwords were also generated using the ARC Nonword Database (Rastle et al., 2002). Non-words were matched with 32 further primes.

Four counterbalanced sets of stimuli were created in which each target word appeared once and no prime appeared more than once. Each set contained eight associated primes, eight collocate primes, eight unrelated primes and eight neutral primes (i.e., an asterisk). This last condition was included to test for inhibition. If any priming effects found are automatic, rather than strategic, we should not find inhibition of targets following the neutral prime. The non-word list was the same for all sets. The four sets of word target stimuli are shown in Appendix Aiv.

Participants

20 undergraduate and postgraduate students from the School of English Studies at the University of Nottingham. All were native speakers of British English.

Procedure

Participants were tested individually in a quiet room. Presentation of the stimuli and recording of the reaction times were controlled by Psychology Software Tools' *E-Prime* software and items were displayed on a CRT monitor. On each trial, a forward

mask composed of a row of seven hash marks (#######) was presented, centred on the screen, for 500ms. This was replaced with the priming word, which was presented in lowercase letters for 60ms. The prime was immediately replaced by the target, in uppercase letters. The target stayed on the screen until the participant made a response or for 2,000ms, whichever was sooner. Participants were instructed to press 'W' on the computer keyboard if a target was a word and 'N' if it was a non-word. They were told to make this decision as quickly as possible. Reaction times were measured from target onset to response. Participants were not told of the presence of lowercase words. Participants received 30 practice trials, sampled from across the conditions, prior to the 64 experimental trials.

Results and discussion

Mean accuracy by participant, collapsed across conditions was 98%. There was no effect of condition upon accuracy. Reaction times of less than 250ms and greater than 1,250ms were excluded from the analysis. Median reaction times for correct responses in each experimental condition are shown in Table 15. Reaction times were not normally distributed within groups.

Table 15: average reaction times across participants for accurate responses in each experimental condition.

	RT (ms) associates	RT (ms) collocations	RT (ms) neutral	RT (ms) unrelated
Median	544	568	563	569

Associated primes significantly facilitated reaction times to target words relative to the unrelated condition. This difference is significant both in the analysis by participants ($T = 7.57$, p (one-tailed) $< .05$, $r = -.31$) and in the analysis by items ($U = 353$, p (one-tailed) $< .05$, $r = -.27$). This confirms previous findings regarding automatic associative priming and demonstrates that the methodology is capable of detected such priming.

Reaction times in the neutral condition were not significantly different from those in the unrelated condition, either in the analysis by participants ($T = 7.71$, p (one-tailed) $= .47$) or in the analysis by items ($U = 443$, p (one-tailed) $= .18$). While it is difficult to draw strong conclusions from non-significant results in a study with relatively few

participants (N=20), this finding suggests that no consistent inhibition occurred in the unrelated condition, indicating that the methodology is likely to have tapped purely automatic processes. This suggestion is further supported by the finding that associated primes show some evidence of having facilitated recognition of targets not only in comparison to unrelated primes, but also in relation to neutral primes, though this advantage is only significant in the analysis by participants (T = 53.00, p (one-tailed) < .05, r = -.31); the analysis by items by items not reaching significance at the p < .05 level (U = 399, p (one-tailed) = .066). While not conclusive, these findings suggest that the priming found in the associated condition is likely to have been due to automatic facilitation effects between associated words, rather than inhibition between unrelated words, as would have been the case if strategic processes had been involved.

As in Study 2, collocating primes did not facilitate reaction times to target words relative to the unrelated condition, either in the analysis by participants (T = 9, p (one-tailed) = .42) or in the analysis by items (t(62) = .293, p = .77 (uniquely, results by items in the unrelated and collocating conditions were normally distributed, so a parametric test was used here)). There is also some evidence that targets following these 'pure collocate' primes were recognised more slowly than were targets following associates, though this difference is significant only in the analysis by items (U = 316.5, p (two-tailed) = .006, r = -.33), with the analysis by participants falling just short of the p < .05 threshold (T= 6.63, p (two-tailed) = .052).

In sum, these results replicate the pattern seen in Study Three: priming is seen between word pairs which are strong normative associates, but not between high frequency collocations which are not likely to be strong associates. This difference appears to be the result of automatic, rather than purely strategic processes. The distinction between associated and non-associated collocations is further reinforced by the finding that the former also facilitate recognition of target words in comparison to the latter, though the failure to find a significant result in the analysis by items means that this conclusion remains provisional. On a methodological point, it is also interesting to note that the facilitation shown between associated pairs showed relative to non-associated collocations stands in contrast to our earlier finding that association priming is only observable relative to semantically incongruous word pairs. It may be that this limitation applied only to methodologies which tap strategic processes.

It may be seen as a shortcoming of the present study that it has used the EAT as the sole criterion for judging whether word-pairs are likely to be strong associates for the study participants. The possible weaknesses of this approach have been discussed above. However, in defence of the present arrangement, we can note, firstly, that items in the associate condition were also found to be strong associates in (at least one of) the studies of Sereno (1991) and Shelton and Martin (1992). The fact that they are also attested as very strong associates in the EAT suggests that their status is likely to be rather stable across different groups of native speakers. Secondly, the fact that a significant difference in reaction times was found between the associate and collocation conditions suggests that there is a real difference between the two sets of items. Since the only distinction made in generating the items was whether they were attested as associates or not, this result appears to confirm that this criterion revealed a genuine difference between the two groups.

Summary and conclusions: co-occurrence frequency and collocational priming

The studies reported in this section aimed to determine whether the relationship between high frequency collocates could be validly described in terms of the psycholinguistic notion of 'priming'. Our results suggest that recognition of target words is facilitated by collocating primes only in special cases: i.e. where the collocation is sufficiently salient for its components to be linked in word association norms. It had been hoped at the outset of these studies that, as well as providing a model of how collocation works in the mind, the priming paradigm might offer us a better way of examining the relationship between corpus data and mental representations, and of studying the development of collocational knowledge in learners, than could be achieved through word association methodologies. However, it appears that the priming approach used here is not more sensitive to mentally represented collocations than is word association. Given the greater difficulty of performing priming studies, therefore, word association would appear to be the better approach.

It must be noted as a weakness of the present studies that the highly artificial nature of the tasks used may have undermined normal mechanisms of collocation processing. In particular, by presenting collocates as two individual words, divorced from any meaningful context, processing advantages which might be found in more natural situations may have been obscured. Moreover, the artificial nature of the word-recognition task, which does not correspond to any real-life language situation, may also have prevented normal processing effects. Further research may therefore wish to study the processing of collocations in larger contexts and with less intrusive measurement techniques. Such work will require a great deal of theoretical sophistication, however. In particular, once collocations are embedded in larger environments, it will be necessary to take account not only of the mutual predictability of the two words which make up a collocation, but also their probabilistic relationships with other (lexical, grammatical, and discoursal) items in the surrounding context. Such measures have yet, to the best of my knowledge, to be developed.

4.6 Summary and conclusions: the psychological reality of high frequency collocations

This chapter has aimed to investigate the relationship between frequencies of co-occurrence of words in a corpus and the mental representation of collocations in the mind. I agued that, though a number of prominent corpus linguists have claimed that high frequency of collocation in a corpus is likely to indicate some form of holistic representation, or mental association between words, the link from corpus to mind is likely to be at best an indirect one, and that further work is needed to investigate the exact nature of this relationship. Various frequency-based methods of identifying collocations were described, and the ability of these methods to predict psychological associations between words was evaluated. It was found that all methods are significant predictors of association, with the directional 'conditional probability' score doing best, and chi-square and z-score also performing well overall. However, the 'noise' inherent in word association databases (arising from the facts that not all association in based on collocation and that only a subset of associated collocations will be reported in the norms) means that the reported correlations probably rather underestimate the true predictive ability of these measures. A second set of studies

examined Hoey's theory of 'lexical priming' in the hope that this might allow a more accurate assessment of the psychological status of high frequency items. However, priming was found only to exist between those collocations which were sufficiently salient also to be registered as psychological associates. This model does not, therefore, appear to offer any advantage over word association tests. It remains possible, however, that other methodologies for investigating processing will be able to identify priming between a wider range of collocations.

Chapter 5
The acquisition of collocations by adult second language learners

5.1 Introduction

Section 3.3 described Ellis's (2001) model of how child first language learners acquire collocations through an associative process of 'chunking'. This mechanism is suggested to be fundamental to the acquisition of all levels of the language system, and renders native speakers sensitive to transition probabilities between sequences of language at all levels of abstraction, with collocation between words being the archetypal example. Section 3.3 also described Wray's (2002) claim that adult second language learning usually fails to follow this process. Wray's model proposes that various cognitive and social factors lead adult learners to analyse the language they meet into word-length units, such that information about the wider syntagmatic context of words is not retained. For this reason, Wray claims, adult learners' knowledge of formulaic language tends to remain relatively weak, even when their knowledge of individual words and of syntax is quite advanced. While it is possible for learners to build up a stock of formulaic language through conscious effort, the fact that this is a post-hoc and attentional process, rather than one of implicit learning from input, means that their consequent feel for formulaicity is unlikely to be entirely nativelike. On this model, Schmitt's recommendation that collocation be learned through "massive exposure to the L2" (Schmitt, forthcoming) does not appear to be a viable pedagogical approach. Rather, learners must build up their collocational associations through a process of proceduralization involving the automatic planning and assembly of utterances (Wray, 2002, pp. 201, 211).

The present chapter will aim to evaluate Wray's model. Section 5.2 will review the relevant literature on the acquisition of formulaic language. We will see that, though this research has frequently pointed out the shortcomings of non-native knowledge with respect to knowledge of formulas, it fails to provide strong evidence either for or against Wray's model because it fails to relate knowledge with learners' likely input.

Sections 5.3 and 5.4 will then present two original studies which aim to test the model. The study in Section 5.3 uses a lab-based training paradigm, in which learner input can be very tightly controlled, in order to evaluate the effects of input on learning. While this approach enables close control over input, it suffers from problems of contextual validity and is not able to evaluate long-term learning. The study in Section 5.4 aims to make up for these shortcomings by comparing learners' use of collocations in natural production with their likely life-long input as estimated by frequencies of occurrence in a large corpus. This achieves greater contextual validity and enables us to evaluate longer-term learning, but does so at the expense of sacrificing tight control over input. These two studies are therefore intended to complement each other, and taken together should give a robust overall view of the impact of input on collocation learning

5.2 Previous research on the acquisition of formulaic language by adult second language learners

Introduction

Previous research into the acquisition of formulaic language by adult second language learners can be divided into two main strands: studies of learners' performance in pen-and-paper tests and studies of advanced learners' spoken and written productions. This section will discuss each of these strands in turn and assess the degree to which they provide support for Wray's thesis that adult learners tend not to learn the collocations they meet in input.

Pen-and-paper tests of formulaic sequence knowledge

A number of studies have assessed learner's knowledge of formulaic language through translation tasks, gap-fills, and intuition judgements. An early study taking this approach is that of Scarcella (1979). She gave 30 advanced adult L1 Spanish learners of English, enrolled in an adult school in the US, a series of situational contexts and asked them to complete each with "a short expression (four words or less) which is frequently used in the given situation" (1979, p. 87 original emphases). Scarcella reports that the test scores were "very low" (1979, p. 81): the average score was 38%, and a similar result (30%) was obtained from a second group of 30 advanced university ESL students. She interprets these results as indicating that

"many common routines are not easily 'picked up' by adult second language learners" (1979, p. 81).

While this conclusion matches Wray's, Scarcella's results should be treated with some caution, since they are based on a rather small sample of items (N=15), many of which appear problematic. Scarcella reports that in norming, all items had elicited 100% uniform responses from a set of 20 native speakers. However, while the necessary completion for some items is quite clear (e.g., "David sneezes. His friend, Sharon, politely says, '___'"), it seems surprising that others produced such predictable responses (e.g., "Robert is a waiter. He is taking an order. When he finishes taking the order, he checks to be sure that his customer doesn't want to order more. Robert asks his customer, '___'"; "Mel accidentally spills coffee on his friends white jacket. Mel says, '___'"). Moreover, other items appear to require phrases which may be highly salient for natives because of their situational distinctiveness, but which are likely to be of rather low frequency for many learners (e.g. "Gary is at a gas station. He wants to buy a full tank of gas. Gary tells the gas station attendant, '___'" – presumably requiring *fill her up*). It is not clear, therefore, just how 'common' these routines will have been in learners' input.

Bahns and Eldaw (1993) used a German-English translation task to assess the knowledge of 15 verb + noun collocations amongst 34 advanced L1 German learners who were studying English as their university major in Germany. They found that a disproportionate number of lexical errors involved words that were part of a collocation: though collocations made up only 23.1% of lexical words in the target translations, 48.2% of lexical errors involved a collocation. Bahns and Eldaw interpret this result as indicating that a learner's knowledge of collocation does not "expand in parallel with his knowledge of general vocabulary" (Bahns & Eldaw, 1993, p. 108). This conclusion, if correct, would appear to support Wray's thesis that adult learners focus on individual words at the expense of collocations. However, the direct comparison which Bahns and Eldaw draw between vocabulary and collocation knowledge is probably not a valid one. Collocations are, in general, much less frequent than words. We would therefore expect a large disparity between knowledge of the two sets, even if learners were equally adept at acquiring both. As in Scarcella's study, no attempt is made to control or describe the frequencies of either the

vocabulary or the collocations tested (which are again few in number), so any direct comparison between vocabulary and collocation learning is difficult to draw.

Farghal and Obeidat (1995) tested the knowledge of 22 English collocations amongst L1 Arabic learners studying English as their major at a Saudi university. 34 learners were tested through a gap-fill task and 23 through an Arabic-English translation task. On the gap-fill task, learners achieved only 18.3% correct responses, and on the translation task a still poorer 5.5%. On both tasks, the chief error type was the use of a synonym. This accounted for 41% of errors on the gap-fill and 35.4% on the translation task. The prominence of this error type, the researchers suggest, is due to teachers' "tendency to teach words individually rather than collocationally", such that the learner "solely relies on the open choice principle" (1995, p. 321). However, there appear to be serious problems with the test items used in this study, which must make us question the value of its results. As with the previous studies reviewed, no indication is given as to the frequencies of the collocations tested (indeed, no indication is given at all as to how the test items were chosen). Moreover, though the researchers claim that the items had been trialled on two native speakers, some sentence contexts appear not to give sufficient clues to the intended answer: "Some people like salty soup, but others like ____ soup", for example, is intended to elicit (without further clues) "bland", while "John is the one with the plain shirt, whereas George is the one with the ____ shirt" requires the answer "striped"). In short, this test does not appear to be sufficiently well-designed to allow any meaningful conclusions to be drawn.

Granger (1998) evaluated the collocational knowledge of 56 L1 French learners of English by asking them to indicate, from a list of 15 adjectives, the acceptable collocates of 11 *–ly* amplifiers. Informants were also asked to note any adjectives which they felt to be more frequently associated with the amplifier than the others. Results were compared with those elicited from 56 native speakers on the same task. Non-natives both marked fewer combinations as being particularly strong than natives (280 vs. 384 for natives) and marked a wider range of types as being acceptable. Granger interprets the former results as indicating that the non-native sense for collocation is weak, and the latter as indicating that it can be misguided (1998, pp. 152-153). However, it must be noted that these results indicate only that the non-

native sense for collocation is weak/misguided in comparison to that of the native, which is hardly surprising. They do not show – as Wray's thesis requires - that this sense is less well-developed than it should be given the amount of exposure to English which these learners have had.

Hoffman and Lehman (2000) also compared native and non-native intuitions, examining how well each matched frequency-based findings from a corpus. They extracted 55 word pairs which were found to be strongly associated in the BNC (as measured by log-likelihood). Most pairs were adjective-noun or noun-noun combinations, though the listing also included other parts of speech. The researchers prepared a questionnaire in which each node was presented without its partner, and asked 16 native and 16 non-native-speaker informants to supply the collocates. All participants spoke "fluent English" and has at least a Swiss high-school degree (2000, p. 21) It was found that, on average, native speakers supplied the 'correct' collocate in 70% of cases. Non-native speakers did less well, achieving an average accuracy of only 34%. While all but one native speaker supplied at least 50% correct answers, and about half managed more than three-quarters, only two non-natives scored more than 50%, the rest falling between 20% and 50%. The authors note, however, that given the relatively low frequency of the nodes used here (such that, they estimate, these collocations are likely to be encountered only five times a year by the average native speaker), even this degree of accuracy is surprisingly high. Of the 'incorrect' answers supplied, native speakers were more likely than non-natives to give answers which were at least attested, but with a lower log-likelihood value. However, while non-natives were more prone to providing semantically-plausible but unattested collocates (*silly-pretences; seasonally-dependent*), Hoffmann and Lehman report that their choices often overlapped with those of natives (though these relationships are not quantified). A further overlap between native and non-native responses is found in the fact that those items which non-natives found most difficult were often also problematic for native speakers. Thus, of the six items for which no non-native provided the 'right' response, only one was answered correctly by more than 50% of native informants.

Siyanova and Schmitt (2008) also compared the sensitivity of native and non-native speakers to patterns of co-occurrence frequency in a corpus. They compiled a listing

of 31 ‘typical’ and 31 ‘atypical’ adjective-noun combinations taken from non-native speaker essays. Typical combinations met at least two of four criteria: a minimum of 21 appearances in the BNC; an MI score of greater than three in the same corpus; appearance in one of two different collocational dictionaries (these dictionaries counting as two separate criteria). Atypical combinations were pairs which did not appear in either the BNC or the collocation dictionaries. The researchers asked 60 native speakers and 60 advanced non-natives to rate each combination on a six-point scale according to how common they felt the combination was in English. Both native and non-native respondents gave significantly higher ratings to the typical than to the non-typical combinations. However, the natives were rather more emphatic in their judgements. Where non-natives gave typical combinations a mean score of 4.51, the mean native rating was 5.50; non-natives gave atypical combinations a mean of 3.32, compared to 2.51 for natives. Similarly, while the ratings of both groups correlated significantly with the frequency of combinations in the BNC, the relationship was stronger for native informants (non-native $r = .44$; native $r = .58$). The researchers also separated typical combinations into ‘high’ (> 100 appearances in BNC) vs. ‘medium’ (21-100 appearances in BNC) frequency groups. While native raters gave reliably higher scores to the former group, non-natives did not. As with Hoffman and Lehman, then, these results suggest that non-native learners do have a reliable sense of collocational frequency, but that it is less strong than that of natives.

Siyanova and Schmitt also examine the effects of extended exposure on collocation knowledge. They divided the non-native group into those who had never been to an English-speaking country, those who had spent 12 months or less, and those who had spent over 12 months in such a country. They found that those who had spent more time in a country tended to give higher scores to more frequent collocations and lower scores to less frequent. Moreover, the group who had spent the longest time in an English-speaking country were the only non-natives to reliably distinguish medium- from high frequency combinations. These results suggest that extended exposure to the L2 can result in a more acute sense of collocational frequency, a conclusion that does not fit well with Wray’s thesis.

In a rare longitudinal study, Schmitt et al (2004) tested 94 non-native learners of English on their knowledge of 20 formulaic sequences at the beginning and end of

two- and three-month presessional courses in academic English at a British university. Both productive and receptive knowledge of these sequences were tested through sentence completion tasks. On the productive test, learners were asked to complete a phrase which was prompted with its initial letters, while on the receptive test, they were asked to choose one of a series of options to complete a sentence. Learners were found to have rather good knowledge of these sequences at the outset of their courses, achieving mean scores of 13/20 on the productive and 17/20 on the receptive versions of the tests. Schmitt et al comment that many of the items tested are unlikely to have been a focus of explicit teaching, and that this high level of knowledge may therefore indicate incidental learning of such sequences. The learners also showed a significant improvement in their knowledge of sequences (all of which appeared at least once during classes) by the end of their courses (productive scores rising to 17/20 and receptive to 19/20). Starting knowledge of formulaic sequences correlated only modestly with scores on a separate vocabulary test, and there was no significant correlation between gain in sequence knowledge and either starting vocabulary knowledge or gain in vocabulary knowledge. The link between formulaic language and vocabulary in general is, the researchers conclude, not a straightforward one. There was also no significant correlation between gain in formulaic sequence knowledge and scores on tests of language aptitude and motivation. In a follow-up study, Dörnyei et al (2004) clarify this surprising finding, showing that degree of social integration into the L1 community has an overriding impact on the learning of formulaic language, which is only modulated by aptitude and frequency.

In sum, though many pen-and-paper tests have claimed to show non-native learners to be poor at learning formulaic language (Bahns & Eldaw, 1993; Farghal & Obeidat, 1995; Granger, 1998; Scarcella, 1979), the methodology behind these studies has often been rather weak, using small samples of occasionally rather suspect test items. From our perspective , these studies are weakened, in particular, by their failure to provide any frequency information about the target items. In the absence of such information, it is impossible to tell whether learners do not know the items because – as Wray's claims - they are particularly bad at learning collocations or simply because they are so infrequently encountered. The only studies which do provide such information (Hoffman & Lehmann, 2000; Siyanova & Schmitt, 2008) seem to show non-natives to have a reasonable grasp of collocational relationships, relative to their

likely input. This grasp is consistently weaker than that of native speakers, but this, of course, is a result that could be replicated for most areas of linguistic competence and could easily be accounted for by the smaller body of input learners are likely to have acquired, without needing to invoke any fundamental difference in learning strategy. It does not seem, therefore, that these studies provide good evidence either for or against Wray's claim that adult L2 learners fail to retain the collocations they meet.

Studies of advanced non-native language

Formulaic language in advanced non-native speech

A number of studies have used 'learner corpus' methodologies to study the occurrence of formulaic language in advanced non-native speech. An early example is DeCock et al (1998), who examined 'recurrent word combinations' in a 60,000 word corpus of informal English language interviews with advanced (L1 French) learners. These were compared with the recurrent combinations found in an 80,000 word corpus of similar interviews with native speakers of English. The researchers automatically extracted from the corpora all continuous two-, three-, four-, and five-word combinations occurring with frequencies of greater than nine, four, three and two respectively. They found as much use of such sequences in the non-native as in the native corpus. Indeed, non-natives were found to use significantly more of the longest (i.e. four- and five-word) combinations that their native counterparts. They also found that non-natives tended to repeat combinations more often, though the differences here were not large: the average log type-token ratio across different lengths of combination being 72.5 for natives and 70.6 for non-natives. DeCock et al also found that native and non-native combinations differed somewhat in character. The learners, they noted, made more use of hesitation phenomena, such as repetitions (*the the; I I*) and filled pauses (*and er*), while frequent native items such as *you know, sort of, I mean* were less prominent in non-native speech. Looking specifically at a defined sub-type of recurrent combinations – those used to mark vagueness – they found significant underuse by non-natives, and also misuse, in that combinations were used in different syntactic and pragmatic contexts than those seen in the native corpus.

Oppenheim (2000) studied the use of 'recurrent sequences' (as identified by human raters) in a set of three-minute speeches given from notes by six advanced non-native speakers from East Asia studying at a university in the USA. She found that her

subjects made extensive use of recurrent sequences, with a mean of 66.4% of words taking part in such sequences. Interestingly, it seems that these learners used recurrent sequences quite consciously. All reported that they made an effort to learn phrases and collocations and that, when preparing their speeches, "pieces of phrases, entire phrases, or strings of phrases came to mind, as opposed to individual words or pieces of words" (2000, p. 235). Oppenheim found only limited variation between individuals in their use of such sequences, and a high level of homogeneity within individual speakers' language across two topics and two deliveries of each speech. She also found that different speakers tended to rely on different types of recurrent sequence, and that this reflected to a certain extent conscious strategies. Thus, speakers who reported that they concentrated on organisation in their speeches were found to use a high proportion of organisational sequences, while those who placed more focus on a fluent, nativelike delivery, made greater use of immediate repetition, a device that appears to aid fluency. Like DeCock et al, Oppenheim also found that sequences were often not nativelike. Indeed, she notes that sequences were "almost exclusively… idiosyncratic" (2000, p. 235).

Foster (2001) looked at formula use in a 20,000 word corpus of spoken data produced by 32 native and 32 non-native speakers of English completing three different interactive tasks (a personal information exchange, a narrative, and a discussion), either with or without time allowed for planning. She asked a panel of seven native speakers to identify language which appeared to be produced as a 'chunk', rather than word by word, or which was part of a sentence 'stem' which had required morphological adjustment or lexical addition. Language identified by five or more informants was taken to be 'lexicalised'. She found that a higher proportion of native than of non-native speech was lexicalised. Interestingly, when native speakers were given time to plan for the tasks, their reliance on lexicalised language decreased (from 32.29% to 25.08%). Non-natives did not show a similar tendency (16.87% in unplanned and 17.23% in planned conditions). This suggests that natives, but not non-natives, employed a higher level of lexicalised language to help cope with the pressures of coming up with both content and language simultaneously. Foster notes that much of this language consisted of time-filling phrases such as *I don't know, I mean, sort of*.

Foster also found that the diversity of native speaker phrases increased in the planned condition. Whereas in the unplanned condition, 32.4% of their lexicalised language was made of phrases repeated seven or more times each, in the planned condition only 20.8% were such highly-repeated phrases. At the other end of the spectrum, the percentage of lexicalised language comprising phrases used only once increased from 31.9% in the unplanned to 55.6% in the planned conditions. Non-natives showed much less diversity overall in their phraseology, and contrary to the native speaker pattern their usage become less diverse in the planned condition (unplanned: 42.5% repeated at least seven times, 24.9% used once; planned: 55.3% repeated at least seven times, 16% used once).

Foster also looked at the accuracy, complexity, and fluency (measured in number/length of pauses) of speech across conditions. For both natives and non-natives, complexity and fluency increased in the planned conditions, and for non-natives accuracy also increased (this was not measured for natives). She concludes that, under the less pressurised conditions of the planned task, native speakers used a "more fluent, open-choice, rule-based style of language" (2001, p. 89) than in the unplanned condition, which elicited a greater reliance on lexicalised language, less complexity and less fluency. For non-natives, on the other hand, the difference between the two conditions rested solely in their producing more complex, accurate and fluent language, with no corresponding change in reliance on lexicalised language. This, Foster suggests, indicates that they relied in both conditions on a rule-based approach to language, requiring either pausing or planning time to execute fully and accurately.

Adolphs and Durow (2004) studied the use of three-word sequences produced in informal English language interviews by two L1 Mandarin students enrolled on masters degrees at a UK university. Two interviews were analysed for each student: one recorded when they were attending a pre-sessional English language course, and one recorded seven months later, when they were some way into their course of study. The two students were selected for analysis from a larger group on the grounds that they represented extremes of 'social integration' into the local community – one ('Beth') having joined social groups and made many native-speaking friends, while the other ('Ann') spent time mainly with co-nationals. Interviews ranged in length

from 3,046 to 7,162 words (excluding the interviewer's turns). In the first part of their analysis, Adolphs and Durow identified the ten most frequent three-word sequences in each interview. They found that, for each speaker, the total percentage of production made up of the ten most frequent sequences rose slightly (from 2.38 to 3.53% for Ann and from 1.34% to 1.48% for Beth) from the first to the second interview. In contrast, the total contribution to production of all recurrent three-word sequences produced more than twice decreased slightly (from 20.98% to 18.93% for Beth and from 12.66% to 9.55% for Ann). Repetition of sequences therefore decreased somewhat overall, but there was a simultaneous increase in the degree to which production relied on a small group of favoured items. Perhaps the most interesting finding from this analysis lies in the nature of the top-ten sequences themselves. For both speakers (but especially for Beth), sequences in the first interview consisted largely of hesitation markers (*just I er; I I I; yeah just er*), while in the second interview these were almost entirely supplanted by more meaningful items (*a lot of; it's very nice; I got some*). This would appear to indicate the learners' moving away from the 'idiosyncratic' sequences identified by DeCock et al and Oppenheim and towards more nativelike usage.

Similar conclusions can be drawn from the second part of Adolphs and Durow's analysis. Here, they looked specifically at how closely the phraseology of lexical items matched native speaker usage. Identifying the 15 most frequent lexical words in each interview, they searched CANCODE, a five million word corpus of native speech, for three-word contexts in which these words were repeatedly used (i.e. more than once per million words). They then determined what percentage of the learners' use of these words matched these sequences. While the numbers involved were too small for valid inferential analyses, the authors found a substantial overall increase in the percentage of Beth's usage which matched frequent native phraseology (from 42.28% to 59.13%) but a small drop for Ann (from 55.72% to 52.99%). However, looking only at the words which appeared in both first and second interviews, both learners showed convergence with native phraseology for the majority of items. Adolphs and Durow conclude that, while Beth's phraseology seems to have improved overall (partly as a result, the implication appears to be, of her higher level of social integration, and consequent exposure to native speech), Ann improved her usage of just those lexical items which she used most frequently.

Formulaic language in advanced non-native writing

An early study of formulas in advanced non-native writing is that of Yorio (1989), who found "extensive use of conventionalized language" (which he appears to identify intuitively) in his analysis of the writing of 25 ESL students who had been resident in the Unites States for 5-7 years. Yorio notes, however, that the learners had "no formal control" over the formulas they used. He found errors of grammar (**take advantages of; *are to blamed for*) and of lexical choice (**made a great job; *on the meantime*), mixed idioms (**give up their freedom of mobility*, meaning 'give up their freedom of movement'), phrases used with the wrong meaning (*in this way,* meaning 'for this reason'; *in addition to*, meaning 'in order to') and what he terms "attempted idioms"' (*at the end of the road,* meaning 'ultimately'; *they feel suspended upon their heads, the Damocles' sword*). Yorio also looked at the use of phrasal and prepositional verbs. Comparing the usage in his learner corpus with that in a similar corpus produced by 15 native writers, he found that natives and non-natives used these forms to a similar extent (they constituted 19.5% of conjugated verbs for native speakers, 14% for non-natives). However, natives used a far higher proportion of 'idiomatic' two word verbs (e.g. *bring up*) than non-natives. Idiomatic forms comprised 36% of all two word verbs for native speakers, but only 6.5% for non-natives. Moreover, non-natives showed a tendency to use two word verbs incorrectly, getting them right in only 59% of cases.

Yorio also compared writing produced under matched conditions by immigrant students (L1 Spanish) resident in the US for five to six years and by English majors (also L1 Spanish) at a university in Argentina who had never been part of an English-speaking community. Yorio found that the latter group produced more grammatically accurate language than the former and made more use of idioms. He also felt that their writing was "more authentic" than that of the immigrant group. Yorio speculates that the authentic nature of their language is the product of a greater use of frequent collocations. While this finding is suggestive, it suffers for not being quantified. In support of his judgement that the Argentinian group's writing was more authentic, Yorio writes that "[t]his impression of greater idiomaticity was apparent to me and to other colleagues whose native speaker impressions I sought"(Yorio, 1989, p. 65). His assertion that they made greater use of collocations is similarly subjective: "After reading the two sets of compositions…it became clear to me that…the compositions

that appeared more native-like contained many more 'English phrases'" (Yorio, 1989, p. 66)

A more rigorous approach is taken by Granger (1998), who compared the use of two "productive speech formulas" in a 250,000 word corpus of essays by advanced (L1 French) learners of English with a similar corpus of native speaker writing. The constructions examined were the passive structure:

> *it* + modal + passive verb (of saying/thinking) + *that*-clause
> (e.g. *it is said that; it can be claimed that*)

and the active structure:

> *I* or *we/one/you* (generalized pronoun) + (modal) + active verb (of saying/thinking) + *that*-clause
> (e.g. *I claim that; we can say that*)

She found that, while the passive structure was used with approximately equal frequency by native and non-native writers, non-natives massively overused the active structure compared to native speaker norms. In particular, certain instantiations of this form were used far more frequently by non-natives: in 20,000 words, non-natives produced the construction with *say* 75 times, compared with four uses by native speakers, and they produced the sequence with *think* 72 times, compared with three by native speakers (1998, p. 155). This lack of diversity in non-native phraseology tallies with the findings of DeCock et al and Foster, reported above. Granger suggests that learners' limited expressive repertoires may lead them to 'cling on' to certain fixed phrases – often L1 cognates - with which they feel confident; using them as (in Dechert's words) "islands of reliability" (Granger, 1998, p. 156). She also notes in this context a possible transfer effect from the first language – the overused expressions tended to be those with direct L1 translation equivalents.

Granger also studied two-word collocations in the same corpora. Looking at the use of intensifying adverbs ending in *–ly* combined with adjectives (e.g. ***perfectly*** *natural;* ***closely*** *linked*), she found that 'maximizers' (e.g. *absolutely; entirely; totally*) were

used with roughly the same frequency (in terms of both types and tokens) by natives and non-natives, but that native writers used far more (types and tokens of) 'boosters' (e.g. *deeply; strongly, highly*). As with the active and passive frames described above, learners tended to overuse a few favourite intensifiers. Granger again notes an apparent influence of the L1 here. *Completely* and *totally*, which have very high frequency direct translation equivalents in French (*complètement* and *totalement*) were significantly overused by these learners, while *highly*, whose literal French equivalent (*hautement*) is infrequent and reserved for formal language, was significantly underused. Moreover, the non-natives tended to adopt what Granger calls 'stereotyped' maximizer + adjective combinations (i.e. formulaic items like *acutely aware, keenly felt, painfully clear*) only when they had a direct translation equivalent or were "lexically congruent". That is to say, restricted collocations were adopted only when similar to L1 phrases.

Lorenz (1999) has also investigated the use of intensifier-adjective collocations in advanced English learners' writing. He compared four corpora of "expository-argumentative" texts: 155,000 words produced by L1 German 16-18 year olds in the *Bundeswettbewerb Fremdsprachen*, the German nation-wide foreign language competition; 145,000 words produced in writing classes by university students of English; 126,000 words of general-topic argumentative essays produced by 15-18 year old British students; and 92,000 words of argumentative essays produced by British undergraduates. Lorenz's study provides further evidence for the 'islands of reliability' hypothesis, finding that the non-natives both overuse "a limited number of high frequency stock items" and that their overall repertoire of collocations (as measured by a 'type-token ratio') is much lower than that of natives (Lorenz, 1999, pp. 168-170).

Lorenz attempts to quantify the 'idiomaticity' of collocation use in terms of the mutual information scores of intensifier-adjective combinations. He finds that the average mutual information score of the 920 combinations in his combined non-native corpora (MI = 7.41) is about 20% lower than that of the 626 combinations in his native corpora (MI = 9.22). Collocations which score highly on mutual information tend, we have seen, to be infrequent, but strongly-associated pairs. On Lorenz's analysis, it appears that native speakers use more of these than do non-natives, who

instead “show a preference for attestedly viable, recurrent combinations” (Lorenz, 1999, p. 181). On these grounds, Lorenz makes the bold claim that mutual information “is no more and no less than a statistical representation of a stylistic quality as elusive as ‘idiomaticity’” (Lorenz, 1999, p. 184).

Hyland (2008) compares 4-word clusters (defined as chunks appearing at least 20 times per million words, and in at least 10% of texts) found in a 730,000 word corpus of research articles in electrical engineering, business studies, applied linguistics and microbiology with those in a 1.9 million word corpus of PhD dissertations and a 825,000 word corpus of MA theses in the same disciplines written by university students in Hong Kong. He finds the two corpora of student writing to contain both a greater concentration and a wider variety of clusters than was found in the research article corpus. Clusters constituted 5.1% of the MA corpus, 3.8% of the PhD corpus and 3.1% of the research article corpus. The MA corpus included a total of 149 different clusters, the PhD corpus 95, and the research article corpus 71. The student genres, Hyland observes, appear to be “more phrasal” than published writing, suggesting a “considerably higher reliance on prefabricated patterns among the less experience writers” (2008, p. 50). Hyland also finds differences between corpora in the actual structures and in their typical structures and functions. He warns, however, that these results need not indicate any “deficiencies” in the student writing. The differences in number and type of cluster between corpora could, he suggests, reflect the differing goals and audiences of the three text types.

A number of researchers have looked specifically at the use of restricted collocations as they are defined on the semantic-syntactic criteria of ‘Russian school’ phraseologists (see Section 2.2). Howarth (1998) claims to find evidence of the underuse of verb + noun collocations and idioms in non-native English academic writing. He defines collocations as combinations in which there is some restriction on the substitutability of elements, and idioms as combinations with entirely figurative meanings. In two corpora of native writing (a 58,000 word compilation of 29 social science texts and a 180,000 word collection comprising “papers on law, chapters from a books on language studies, and a complete book on social policy” (Howarth, 1998, p. 165)), the percentage of verb-noun combinations which were restricted collocations or idioms was 31% and 40% respectively. The figure for a non-native corpus (25,000

words produced by students on a masters course in English Language Teaching), however, was only 25%. These figures lead Howarth to conclude that "native speakers employ about 50% more restricted collocations and idioms (of a particular structural pattern) than learners do, on average" (1998, p. 177). It is worth noting, however, that much of the difference depends on native writers' greater use of idioms, rather than collocations. The differences between collocation use in the non-native corpus and the first of the native corpus (24% vs. 28%) is actually smaller than that between the two native-speaker corpora (28% vs. 35%).

Kaszubski (2000) looks at intermediate and advanced English learners' use of six high frequency verbs (*be, do, have, make, take, give*) in free combinations, restricted collocations, and 'frozen uses'. She compares argumentative essays from a range of corpora produced by intermediate Polish and Spanish learners, advanced Belgian-French and Polish learners, native college students, and native professional writers. She reports that variation between the behaviour of different verbs makes it difficult to make absolute claims about the degree to which writers use restricted collocation in general, but that there appear to be three broad groups, comprising 1) intermediate learners, 2) advanced learners and native college students, and 3) native professional writers, with usage of free combinations decreasing from 1 to 3. The trend is far from emphatic though, and is even less so when considering the proportion (rather than number) of combinations which are free or restricted. One pattern which does emerge quite strongly is that of learners' overuse of a few favoured collocations, generally either high frequency register-neutral items or items similar to L1 phrases.

The most comprehensive analyses of phraseologically-defined collocations in learner writing to date is that of Nesselhauf (2005). Like Howarth, Nesselhauf looks specifically at verb + noun combinations, defining collocations as combinations in which there is some arbitrary restriction on what nouns can appear with (a given sense of) the verb. She analyses some 2,000 collocations taken from a 150,000 word corpus of argumentative essays written by advanced German and Austrian learners of English (part of the International Corpus of Learner English). Nesselhauf finds evidence for extensive erroneous use of collocations (though the issue of whether collocations are more problematic than non-collocations is not satisfactorily resolved (Durrant, 2007)). However, she also claims that her data show extensive use of collocations which have

been produced as memorized "chunks". This is evidenced both in the large number of native-like collocations used and in the fact that inappropriate usage is often the result, not of combining words in an unconventional way, but of using conventional word pairs which are not appropriate (Nesselhauf, 2005, p. 247). This suggests that the difficulty learners have is not so much that of learning which words go together as of learning how to employ the chunks they know.

Nesselhauf also considers how learners' use of collocations varies depending on the conditions under which texts are composed and on the length of learners' experience with English. She finds that use of restricted collocations was somewhat lower in writing produced under pressure of time (with 12.6 collocations per 1000 words in timed and 14.3 in untimed conditions). This she contrasts with Cowie's finding that native speakers made greater use of collocations when writing under time pressure, a result which appeared to indicate that writers resort to prefabricated language in increase fluency (Cowie, 1992). These diverging results tally well with those found for native vs. non-native speech by Foster (Foster, 2001 - see above).

To gauge to the effect of learners' length of experience with English on collocation, Nesselhauf divides learners into four groups according to the number of years they have studied English (5-8 years; 9-10 years; 11-12 years; 12-17 years). She finds that the number of collocations produced decreases slightly as experience increases, while the percentage or errors made remains roughly the same. A similar analysis which divides learners according to the length of time they had spent in an English-speaking country (never/less than one month; 1-6 months; at least 7 months), also finds the number of collocations used decreasing with experience, but does show some improvement in accuracy over time (2005, pp. 234-236).

Summary and conclusions: formulas in advanced non-native language

All of the studies reviewed here agree that advanced non-native learners do use formulaic language (in some cases quite self-consciously (Oppenheim, 2000)). Indeed, certain types of formulaic language appear to be used more extensively in non-native than in comparable native productions (De Cock et al., 1998; Granger, 1998; Hyland, 2008; Lorenz, 1999). Some researchers have suggested that this may indicate over-reliance on a small range of favourite phrases, especially on items that are frequent or

are cognate to L1 forms (De Cock et al., 1998; Foster, 2001; Granger, 1998; Kaszubski, 2000; Lorenz, 1999; Nesselhauf, 2005). At the same time, certain categories of formulaic language appear to be underused, compared to native norms (Foster, 2001; Granger, 1998; Howarth, 1998). It also seems that non-natives, unlike natives, do not make greater use of formulaic language when working under increased pressure, either in speech or writing (Foster, 2001; Nesselhauf, 2005). Non-native speech is marked by extensive use of recurrent dysfluency markers (such as filled pauses and hesitation markers) (De Cock et al., 1998; Oppenheim, 2000), although it seems that extensive interaction with native speakers enables them to overcome this (Adolphs & Durow, 2004). With regard to writing, in contrast to this last finding, neither extent not accuracy of collocation has been shown to increase with time spent in an English-speaking country (Kaszubski, 2000; Nesselhauf, 2005; Yorio, 1980).

With regard to Wray's thesis that adult learners do not acquire collocations from their input, these studies suffer from a similar problem to the pen-and-paper studies: they do not provide any information about how frequent formulas are likely to have been in input. Those studies which have used frequency criteria to define formulas in learner language (Adolphs & Durow, 2004; De Cock et al., 1998; Hyland, 2008; Lorenz, 1999) have looked at how frequent phrases are in the learners' own production, rather than in their likely input. This is a good approach identifying the phrases which are likely to be formulaic *for the learners*, but it does not tell us anything about the relationship between learning and input. The second part of Adolphs and Durow's (2004) paper uses phrases identified as frequent in a native corpus, and appears to identify convergence between input and learner knowledge. However, the small sample (two learners) leaves this result in need of corroboration.

Formula learning and learner input

The aim of the present chapter is to evaluate Wray's claim that adult second language learners tend not to acquire the collocations they meet in input. To assess this claim properly, it is essential to have some idea of what input learners are likely to have received. From this perspective, it is a major shortcoming of the studies reviewed above that the vast majority look only at the product of learning, without taking any account of likely input. The only studies which enable us to get an idea of how knowledge might have been affected by input are those of Hoffman and Lehmann

(2000), Siyanova and Schmitt (2008), and Adolphs and Durow (2004). None of these studies offer direct support for Wray's thesis, since the first two indicate that non-native learners appear to be sensitive to co-occurrence frequencies (though less sensitive than natives) and the last suggests convergence between learner knowledge and patterns in the input. However, neither can they be said to disprove the model. It is possible that the deficit between native and non-native performance seen in Hoffman and Lehmann (2000) and Siyanova and Schmitt (2008) is partly the result of a different learning approach; while the small sample of Adolph and Durow's study prohibits generalisations.

Further research is clearly needed to evaluate the extent to which adult learners acquire the collocations they meet in input. An ideal study in this context would need to relate learners' knowledge of collocation to their whole history of interaction with the language. Such a project is, of course, impossible in practice, and probably even in principle (recall Hoey's observation that "the personal 'corpus' that provides a language user with their lexical primings is by definition irretrievable, unstudiable and unique" (2005, p. 14)). It is possible to make various approximations to this ideal, however. The studies reported in what follows are attempts at this.

The first study (Section 5.4) uses a lab-based training methodology. On this approach, the input which learners receive can be very tightly controlled and small resultant gains in knowledge can be tested. However, it has the disadvantage of lacking contextual validity. One problem is that any approach to learning demonstrated in the lab may not apply equally in other settings. Another is that any learning demonstrated over the necessarily short course of an experiment may not be durable, and so may not feed fully into the longer-term learning process. With these problems in mind, the second study (Section 5.5) aims to relate learners' use of language, as evidenced in corpora of learner language, to their likely long-term input, as evidenced in a corpus of the target language. This approach sacrifices much of the control over input which a lab-based approach enables - the target-language corpus used can only provide a very rough approximation to likely learner input. Moreover, learner corpora may not fully reflect the collocational associations which learners have formed. However, the corpus-based approach gains much in terms of contextual validity: it enables us to consider the cumulative effects of likely long-term exposure in a normal learning

environment and their reflection in normal language use. It is hoped that these two methodologies will complement each other. Used in tandem, any convergent results from the tightly controlled, but less contextually-valid lab-based approach and the less well-controlled, but highly contextually-valid corpus-based approach should provide us with a robust composite view of the collocation learning process.

5.3 Do adult second language learners remember the collocations they meet in input?

Introduction

The main aim of this study is to test, through a lab-based training experiment, Wray's (2002, p. 209) claim that adult non-native learners of English do not retain information about the collocations they meet in their input. A secondary aim is to determine the effectiveness of different types of repeated exposure with which teachers might wish to present their students to facilitate collocation learning. The study will control the input adult L2 learners receive of target word pairs, and then test their retention of those pairs. Participants will undergo a short training session in which they are exposed to a number of target adjective-noun combinations embedded in sentences. They will then undergo a cued recall test to see whether memory for target nouns is facilitated by the presence of their paired target adjectives. Any such facilitation will provide evidence that an association has been formed between the paired words in training.

The study will look at learning under three different conditions. In the first condition, participants are exposed to word pairs in a sentence context one time only. Presumably, a single exposure to a word combination is unlikely to have a lasting impact on a learner's language system. Combinations which are encountered once and never met again are not collocations in the sense in which that term is used in this thesis. Collocations are, rather, those combinations which the language user meets repeatedly over time and which for this reason come to be retained as permanent features of the speaker's linguistic knowledge. However as Goldberg (2007) has noted, if the effects of repetition over time are ever to be felt, some memory trace must be left by even a single exposure to a stretch of language. Without such a trace, the

learning process could never get started. This first condition aims to test for such a trace.

The second and third conditions examine the effects of different types of repetition on learning. Most individual collocations are – compared to individual words – relatively rare, and for the native speaker repeated exposure to important collocations will be provided only through an extended period of immersion in the language. Since most adult learners do not have the luxury of such extensive input, teachers may wish to short-cut this process somewhat by providing 'artificially-enriched' input in which learners encounter target collocations repeatedly in a short space of time. At least two different types of repetition could be envisaged here. The first is verbatim repetition of a single linguistic context. That is, the learner could engage with one piece of language a number of times over. Repeated exposure of this sort to a single stretch of language can be seen as a form of fluency-building activity. A learner's initial contact with a piece of language is likely to involve a number of pressing cognitive demands – recognising the words, decoding the syntax, creating a plausible semantic context and deriving a meaning – which may inhibit any actual learning. A second exposure to the same stretch of language, with these issues at least partially resolved in the learner's mind, may enable the learner to focus more on consolidating and building fluency with the language (Schmidt, 1992, p. 361; Segalowitz & Hulstijn, 2005, p. 381). This is the rationale behind such recommended collocation-consolidating activities as '4-3-2 minute talks', in which learners are asked to repeatedly repeat a particular talk in increasingly shorter lengths of time (Hill, Lewis, & Lewis, 2000, pp. 90-91).

A second type of repetition is the repeated use of a target collocation in different sentence contexts. In this scenario, the learner's cognitive burden may remain relatively high at the second encounter. However, the fact that the learner meets two stretches of language in which only the collocation remains constant will presumably make that collocation much more salient for the learner than it would otherwise be. This may be a distinct advantage over the first type of repetition, in which there is nothing to direct the learner's attention to the target collocations, rather than to any other aspect of the sentences they encounter. The second and third experimental conditions in this study aim to examine the effects of each of these types of repetition.

For all three conditions, the testing phase of the experiment consists of a naming task. For each item in the test, learners will first be shown the adjective part of one of the adjective-noun pairs. Immediately afterwards, they will be shown the first two letters of the noun part, followed by dashes for the missing letters. They will be asked to say the noun aloud if they recognise it. If nouns are recognised more reliably when they follow an adjective with which they were paired during training, this will be taken as evidence that some memory has been retained of the two words' co-occurrence. It should be noted that this test is not designed to assess all aspects of collocation knowledge. Rather, it aims to determine whether learners have established a formal association between the two words involved.

Materials

To create target adjective-noun pairs for learning, a number of nouns were first selected according to the following criteria:

- All nouns appear in the British National Corpus (BNC) with a lemmatised frequency of between 50 and 100 occurrences per million words. This places them within the top 2,150 most frequent words in the corpus but outside of the top 1,100 (Leech et al., 2001). This criterion aims to ensure that subjects are likely to have some familiarity with the words while avoiding the ceiling effects associated with very high frequency forms.
- All nouns are four or five letters in length. In the testing phase of the experiment, subjects will be asked to complete words from two-letter stems (e.g. *EV_ _ _*; for *event*); words of similar lengths were therefore used since words of very different lengths would be likely to make this task more difficult for some words than for others.
- The word-completion task may also be affected by the number of other words sharing the same stem as the target (e.g. *RO_ _* could be completed by a large number of alternative words – *road, rock, role, roll, room, rope,* etc – whereas *ER _ _ _* offers few alternatives to *error*). To control for this, I determined how many nouns with lemma frequencies of at least 15 per million in the BNC shared the same stem as the target noun; nouns which shared a stem with fewer than 10 or more than 30 other items were not included. I also checked how

many words both shared a stem with the target noun and contained the same number of letters (such that they could actually be substituted for the target in the task); nouns with fewer than three or more than nine such possible substitutes were also excluded.

Second, target adjectives were selected according to the following criteria:

- All adjectives appear in the BNC with a frequency of between 50 and 100 occurrences per million words. As with the nouns, this aims to ensure that the words are known to subjects without being over-frequent.
- Since strong pre-existing collocational associations of the adjectives may affect the testing phase, any adjectives which were likely to have such associations were excluded; i.e. any adjectives which are followed by one particular noun in 5% or more of their occurrences in the BNC were excluded.

The selected nouns and adjectives were then combined into 20 target word pairs which fulfilled the following conditions:

- All pairs appear with zero or low frequency (i.e. one or two occurrences) in the BNC. This condition aims to ensure that subjects are unlikely to have formed any collocational association between the words prior to training.
- All pairs were judged by myself to be meaningfully combinable in plausible contexts.

A number of different sentences were then created containing each target pair. Twelve native speakers of English were asked to rate each of these sentences on a six-point Lickert scale according to how 'natural' they were (1 = 'very unnatural', 6 = 'completely natural'). Only sentences receiving a mean rating of five or above were retained. From the retained sentences, 40 were selected for use in the final materials: two sentences for each of 20 target word pairs. Additionally, 40 matched 'control sentences' were also created. These were identical to the training sentences except for the target word pair. In each case, the target noun was kept, but the adjective was either deleted or – if deletion made the sentence unnatural or nonsensical - replaced by a different adjective (none of which had been used in any of the target sentences). For

example, the training sentence for the target pair *busy route* was *Extra buses were introduced on the busy route into the city*, and this was matched with the control sentence *Extra buses were introduced on the route into the city*. Likewise, the training sentence *Hot chocolate is an excellent drink on a cold evening* was matched with the control *Hot chocolate is a wonderful drink on a cold evening* (target pair = *excellent drink*).

To create the training materials, these sentences were divided into two sets, with each set containing 20 training sentences (one for each of the 20 target pairs), plus 20 matched control sentences. Each of the two sentence sets was further divided into two counterbalanced experimental lists, each containing 10 target sentences and 10 control sentences, with nouns which appear in a training sentence in one list appearing in their control sentence in the other, and vice-versa. For both lists, the 10 training sentences included six with five-letter target nouns and four with four-letter target nouns. Two groups of 20 'filler sentences' were also created – one for each of the two main training sets – and added to each list. Each filler sentence included a noun of four or five letters from the same 50-100 occurrences/million band as the target nouns. In this way, the final materials consisted of two sets of two counterbalanced lists, with each list containing:

- 10 sentences containing a target word pair;
- 10 sentences containing only the noun part of a target word pair;
- 20 sentences containing other nouns.

None of the target nouns or adjectives were used in any sentence other than their training or control sentence. No lexical words were used in any of the sentences which shared a two-letter stem with any of the target nouns. The four experimental lists are shown in Appendix B.

Participants

The participants were 84 non-native speakers of English (56 female, 28 male). All were undertaking taught postgraduate courses at the University of Nottingham at the time of the experiment. The mean age of participants was 25.1 (max = 41, min = 19).

Participants came from the following L1 backgrounds: Mandarin (26), Thai (6), Malay (5), Serbian (5), Arabic (4), Cantonese (4), Hindi (3), Kiswahili (3), Russian (3), Spanish (3), Igbo (2), Indonesian (2), Japanese (2), Marathi (2), Telugu (2), Gujarati (1), Hungarian (1), Italian (1), Kazakh (1), Kinyarwandan (1), Melayalam (1), Persian (1), Portuguese (1), Singhalese (1), Slovene (1), Vietnamese (1), Yoruba (1). Although no standardized measure of L2 proficiency was available for each student at the time of the study, the University of Nottingham has an entry requirement of 6.0 IELTS or 550 TOEFL (paper version), and so the students can be assumed to be reasonably proficient in English.

Procedure

Participants were assigned in equal numbers to one of the three training conditions: single exposure, verbatim repetition, and varied repetition. Within each condition, participants were in turn assigned in equal numbers to one of the two counterbalanced experimental lists. In this way, half of the participants in each condition saw 10 of the 20 target collocations. They saw the other 10 target nouns in control sentences without their paired adjectives. For the other half of the participants, this situation was reversed: the target nouns which the first group had seen alone were presented with their adjective pairs, while the nouns which the first group had seen with their pairs were seen in control sentences.

In the single exposure condition, participants only saw sentences from the first of the two sets of training materials. That is, each participant saw 40 sentences: 10 in which target nouns appeared together with their paired adjectives, 10 in which target nouns appeared without their paired adjectives, and 20 filler sentences. Sentences were presented to participants on a computer screen in random order. Before each sentence, participants were presented with a fixation point (‘+’) for two seconds. This was then replaced with the sentence, which remained onscreen for seven seconds. Participants were instructed to read the sentence aloud into a headset-mounted microphone. After seven seconds, the sentence disappeared and they were invited to press a button on the computer keyboard to continue to the next item. The training phase lasted approximately seven to eight minutes.

In the verbatim repetition condition, participants were again exposed only to sentences from the first of the two sets of training materials. This time, however, each sentence was presented twice. The training began identically to that in the single exposure condition, with sentences being presented in random order for seven seconds each and participants instructed to read the sentences aloud. On completion of this phase, participants were told that they would be asked to repeat the process at a faster rate. The same sentences were then re-presented in the same way and in a new random order, but this time participants were only given three seconds to read each sentence. This condition was intended to focus students on fluent production of the sentences. The training phase lasted approximately 11-12 minutes.

In the varied repetition condition, participants were exposed to sentences from both sets of training materials. That is, each participant saw 80 different sentences, including 20 in which target nouns appeared together with their paired adjectives (two different sentences for each of 10 different targets), 20 in which target nouns appeared without their paired adjectives (again, two sentences for each of 10 different targets), and 40 filler sentences. The training again began identically to that in the previous two conditions, with sentences from the first training set being presented in random order for seven seconds each and participants reading each sentence into the microphone. On completion of this phase, participants were told that they were half way through and invited to re-commence in their own time. In the second phase, sentences from the second training set were presented, again in random order and for seven seconds each, with participants reading each sentence aloud into the microphone. The training phase lasted approximately 14-15 minutes.

On completion of the training, participants moved directly to the testing phase of the experiment. Testing took the form of a naming task (based loosely on the tasks used in Schooler and Anderson (1997)), in which subjects first saw a fixation point ('+') for 1.5 seconds, followed by the adjective part of one target adjective-noun pair, presented in lower-case letters (e.g. 'warm'), which also remained onscreen for 1.5 seconds. This was immediately followed by the stem of the noun from the same pair, in upper-case letters (e.g. 'FL_ _'; for the noun *flat*). The stem remained onscreen for five seconds. Participants were told that the upper-case word stem would be a word from one of the sentences they had just read and were instructed to say the word into

the headset-mounted microphone if they thought they knew what it was. Participants were not informed of any connection between the adjective and the target noun. The test administrator noted whether a correct or incorrect response had been given. The test started with four practice items (using nouns from the filler sentences), followed by a main test consisting of 20 items – one for each target word pair – presented in random order. The same test items were presented to all participants.

Results and discussion

Since participants saw all target nouns either once (in the single repetition condition) or twice (in the other conditions) each, recall for all nouns should have been the same, other things being equal. However, if some memory was retained from the training phase of the pairings between the target adjectives and nouns, then the adjective prime should have provided participants with an additional memory cue for those nouns they saw together with their adjective partners. Thus, we can expect some level of recall for all nouns, but if participants formed an association during training between the two parts of the target pairs, their recall should have been better for those nouns which they had seen with their adjective pairs than for those nouns which they had seen only in control sentences.

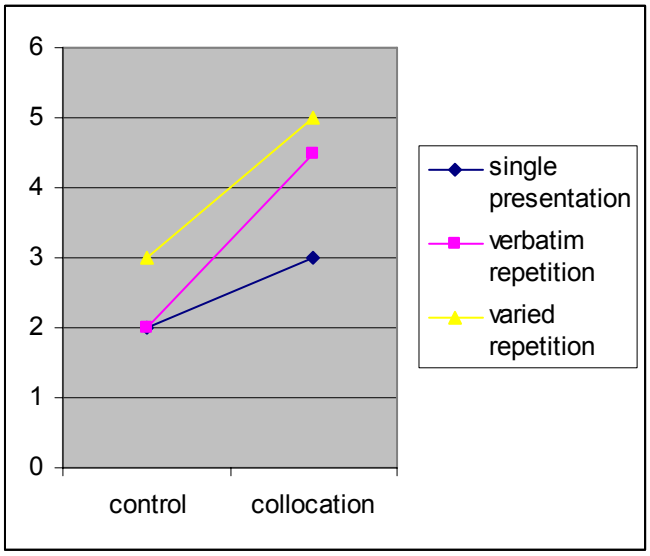
Table 1 shows the minimum, maximum, and median number of correct identifications of target nouns in control and collocation conditions for each of the three training types (median scores are presented because scores were not normally distributed within conditions). These median scores are also presented visually in Figure 1.

Table 1: Retention of collocation information across three conditions

	min. recall (/10)		max recall (/10)		median recall (/10)			
	control	coll.	control	coll.	control	coll.	Wilcoxon Signed Ranks Test[1]	Wilcoxon Signed Ranks Test[2]
Single presentation (N=28)	0	0	6	7	2	3	$T = 55.5$, $p < .05$ $r^3 = -.25$	$T = 42.0$, $p < .05$ $r = -.26$
Verbatim repetition (N=28)	0	2	5	9	2	4.5	$T = 8.0$, $p < .001$ $r = -.56$	$T = 9.5$, $p < .001$ $r = -.57$
Varied repetition (N=28)	0	2	7	8	3	5	$T = 30.0$, $p < .001$ $r = -.48$	$T = 2.0$, $p < .001$ $r = -.58$

1. One-tailed significance, averaged across participants
2. One-tailed significance, averaged across items
3. Effect size

Figure 1: retention of collocation information across three conditions



In all three training conditions, nouns which were seen together with their paired adjectives during the training phase were remembered significantly more frequently than those which were not. The strength of this effect was, as we would expect, weakest for those participants who had received only a single exposure to the target collocations. The effect size was -.25, which falls below Cohen's benchmark of .3 for a medium effect (Field, 2005), and so must be considered a small effect. However, some memory of the co-occurrence of the adjective-noun pairs met in training was clearly retained, in spite of the facts that each pair was seen only once in an eight-

minute training phase and that participants had not been told that they would be asked to recall anything about the language they read. We can conclude, therefore, that adult second language learners do – in contrast to Wray's claims – retain some memory of which words go together in the language they meet. Since this retention appears to occur implicitly – i.e. without the conscious intention of the learner – this suggests that adult L2 learners are likely to gather information about the collocations in their input, regardless of any intentional study techniques or strategies. This suggests that any shortcomings in non-natives' grasp of collocational links between words may be a product of insufficient exposure to the target language, rather than of a distinctively 'word-based' approach to learning.

Unsurprisingly, both of the repetition training conditions yielded superior levels of recall in comparison to the single presentation condition. Verbatim repetition also appears to have some advantage over varied repetition. Although there is little effect size difference when results are averaged across items (.57 vs. .58), a wider difference opens up when results are averaged across participants (.56 vs. .48). Cohen's criterion for a large effect is .5 (Field, 2005), and both repetition conditions either closely approach or exceed this, so the effect of minimal repetition of collocation input (only two repetitions) in facilitating collocation recall can be considered large. Thus both the verbatim and varied conditions in this study appear to be effective means of establishing initial collocation memory traces, with verbatim repetition being slightly more effective.

This superiority of verbatim repetition over varied repetition is shown more clearly when comparing the 'gain' in recall across the two conditions (where gain is defined as the recall for control nouns subtracted from the recall for collocating nouns). Gain in the verbatim repetition condition ($Mdn = 2$) is confirmed to be significantly greater than that in the varied repetition conditions ($Mdn = 1.5$, $U = 123.5$, p (two-tailed) $< .05$, $r = -.33$; note that the use of median scores for this non-parametric data means that the average of all gain scores is not exactly equal to the difference between average control and average target scores presented in Table 1). It seems, then, that the fluency-oriented repetition of a single sentential context yielded better collocation learning than exposure to alternative contexts, with a medium effect size (i.e. > .3). Although this study was not designed to explain this advantage, we can speculate that

the cognitive ease of reading an identical sentence the second time around somehow makes it easier to form a collocational memory trace. Alternatively, perhaps the timed nature of the fluency-based verbatim condition somehow increased the participants' attention on the language, leading to better results, even though less time was spent on the input (7 minutes + 3 minutes for the verbatim condition; 7 + 7 minutes for the varied condition). Further research will be required to disambiguate these possibilities and to further our understanding of the most effective methods of facilitating collocation knowledge.

These results raise a number of other important issues for future research. In this study, testing took place immediately after training. However, attrition of lexical knowledge is a widely reported phenomenon. It would be very interesting to discover how durable the reported memory traces are. It is especially important for pedagogical reasons to establish the length of time before the initial trace 'disappears', as subsequent exposures need to be received in time to build upon the previous knowledge, otherwise the learner would be forever 'starting over' in the effort to establish collocation knowledge. Detecting a level of retention below which learners could be counted as 'starting over', however, would probably require more subtle tests of association than are used in the present study (the 'savings' methodology employed in lexical attrition studies (e.g., Hansen, Umeda, & McKinney, 2002), for example, might be used). It would also be interesting to establish how many exposures are typically required for stable, long-term associations to become established. Research from reading indicates that new words need to be seen around 8-10 times in order to be learned (Schmitt, forthcoming), and it would be surprising if collocation knowledge could be acquired in any fewer exposures. Research establishing how many exposures are required over what period of time for collocation learning to take place would certainly be useful for pedagogy.

Also, this study dealt with implicit learning only. As explicit attention is widely acknowledged to facilitate lexical learning, it can only be assumed that an explicit focus on target collocations would dramatically improve their acquisition. Moreover, the learners in this study were required only to engage with the training sentences at a 'formal' level; i.e. by reading the sentences aloud. As was noted in Section 3.3, Ellis (2005) maintains that meaning can play an important role in collocation learning. It

would therefore be interesting to see what effect different types of meaning-focused training tasks would have on acquisition. Additionally, the knowledge assessed is only knowledge that there is a formal connection between the two words involved in each collocation. No assessment has been made here of how well the meaning and use of the collocations have been learned. Future research should also include a consideration of these other aspects of collocation knowledge.

Finally, this study has looked only at how learners come to establish associations between words that they are already assumed to know. It is possible that somewhat different processes will be involved for collocations of previously unknown words. It would be interesting for future research to address this issue.

As has been noted, the research reported here is limited in that is deals only with learners' short-term retention of the forms of one type of collocation – directly adjacent modifier-noun combinations – in a rather artificial learning environment. Any strong pedagogical recommendations must therefore await further research with a wider variety of items, in a number of different settings, and over longer periods of time. It is also important that we establish how meaning interacts with this purely formal learning. This would constitute a substantial research programme. However, the results of such a programme have potential applications in a number of areas. They would, for example, give teachers important clues as to how materials should be designed (e.g., replication of our current results would suggest a need for materials in which the target collocations are met several times within a relatively short period of time); what the pay-off would likely be from different types of classroom activity (e.g., replication of our current results would suggest that fluency-based exercises would facilitate advances in collocational knowledge, as well as in reading fluency); and what the likely size of any gains from input are likely to be (replication of the small gains seen here would suggest that learners will need substantial levels of exposure to build up a native-like knowledge). Given these potential benefits, such an ambitious research programme seems worth the substantial time and effort it would involve.

Summary and conclusions: recall for collocations

This study has shown that adult non-native learners of English do retain, at least in the short-term, and under laboratory conditions, some information about what words

appear together in their input. This suggests that adult L2 learners may not, as Wray has claimed, focus their learning entirely on individual words. Rather, as Ellis's L1 model predicts, learners automatically retain a memory of collocational chunks from the language to which they are exposed. This suggests that they will learn the collocations they repeatedly meet. Any deficit in learners' knowledge of collocation may therefore be the result of insufficient exposure to the language than of a fundamentally different approach to learning.

We have also seen that the fluency-oriented repetition of individual sentence contexts has a greater impact on collocation learning than does exposure to the same collocations in different contexts. Teachers wishing to foster their students' collocation learning may therefore wish to give special emphasis to activities in which learners have the opportunity to encounter the same language several times, enabling them to focus on building up fluency with particular strings of language without the 'distractions' of dealing with new contexts and meanings.

5.4 The use of frequent collocations in native and nonnative writing

Introduction

The study reported in Section 5.3 has suggested that adult second language learners can acquire collocational associations from exposure to the L2. However, as was discussed above, it is possible that the learning demonstrated in a lab setting may not be found under more normal conditions. It is also not clear that memory for collocations will be retained over sufficient lengths of time for stable associations to form. The present study aims to examine collocation learning in more natural conditions and over a longer time span by comparing learners' use of collocations with their likely long-term input.

In particular, it will ask to what extent advanced non-native speakers of English use in their writing collocations which are frequent in the language. Kjellmer (1990) has claimed that even quite advanced learners tend not to use much formulaic language, and that this is a major reason why otherwise competent non-natives can sound unidiomatic. Rather than constructing their language phrasally, as native speakers

often do, non-natives piece their language together word-by-word, in ways that they can only hope will prove acceptable; as Kjellmer puts it, their "building material is individual bricks, rather than prefabricated sections" (1990, p. 124). If this is right, it suggests that learners are – as Wray's model predicts - failing to retain (or, at least, to use) the formulaic language to which they are exposed.

The present study will examine Kjellmer's claim by comparing the extent to which native and non-native writers use collocations which have a high frequency of occurrence in the British National Corpus (BNC). It will be assumed that frequencies of occurrence in the BNC will approximately reflect those which both native and non-native speakers are likely to have encountered in their input. The BNC will not, of course, be an exact match for the language experience of either group. Such loss of control over learners' input is, however, an inevitable consequence of looking at normal learning over an extended period of time.

If adult learners tend, as Wray (2002, p. 209) claims, to forget the collocations they meet, we would only expect them to pick up and use those word pairs which they had intentionally learned. Since this is likely to constitute only a small proportion of the frequent collocations in the input, we would expect them to make far less use of such pairs overall than native speakers. In other words, their writing is likely to follow the non-formulaic style described by Kjellmer. If, on the other hand, non-native learners do remember the collocations to which they have been exposed, we would expect them to use many more such collocations. Since their exposure to the language will have been so much smaller than that of mature native speakers, however, we might expect their repertoire to consist largely of those pairs which are most frequent in the language, since they may not have had sufficient input for associations between lower-frequency pairs to form.

Materials

This study will compare the use of high frequency collocations in several comparable sets of native and non-native writing. The first set of non-native texts are research assignments produced as project work for courses in English for Academic Purposes (EAP). This text type was chosen because it is one of the few varieties of extended non-native writing. It was thought necessary to use such extended pieces because the

study will rely on analysing the extent of collocation use in individual texts and it was suspected that statistically robust trends may only emerge in longer stretches of writing, where larger numbers of collocations could be identified. The essays were written by two groups of learners: postgraduate students on pre-sessional EAP courses at a British university; and first-year undergraduates on in-sessional EAP courses at an English-medium university in Turkey[3]. To explore whether the analysis could also work for less extended texts, a set of shorter essays was also analysed. These comprised short compositions written by pre-sessional students at a British university and short 'argumentative' essays from the Bulgarian sub-corpus of the International Corpus of Learner English (ICLE) (Granger, Dagneaux, & Meunier, 2002).

Identifying native texts that are equivalent in type to non-native writing is, as other researchers have noted, highly problematic (Granger et al., 2002, p. 40; Lorenz, 1999, p. 14). The long non-native texts under analysis here do not have readily available native-speaker equivalents: EAP research projects are different in type from normal academic research projects, since they are produced in a class focusing primarily on generic writing and academic skills, without specialist topic-based input, and are intended to be read by an English teacher, rather than by a subject lecturer. In lieu of strictly parallel corpora, therefore, two sets of native writing were analysed which were taken to resemble the EAP projects in different and complementary ways: postgraduate writing (assignments from students on the MA degree in Applied Linguistics at the University of Nottingham), and essays from the current affairs magazine *Prospect*. The former are similar in form to the EAP projects, but more specialised in topic, since they are written with the support of content-based courses and are intended for an expert readership. The latter are argumentative essays of a similar length to the academic papers. Though distinct in style from academic writing, they are similar to the non-native texts in that they are of similar length, are formal in style, present an argument, and are intended for a general lay audience rather than for specialists.

As a comparison for the shorter non-native texts, two sources were again used. One was argumentative essays written under timed conditions by British undergraduates on

[3] Part of this corpus was provided by Robin Turner.

the topic, ‘A single Europe: A loss of sovereignty for Britain’. These essays were collected by Granger and her colleagues for the Louvain Corpus of Native English Essays (LOCNESS) (Granger et al., 2002, p. 41) with the specific intention of paralleling texts in ICLE. While these texts are similar in type to the shorter non-native texts, the fact that they are all written on a single topic introduces a risk of skewed data. To incorporate a broader range of topics, opinion articles from two UK newspapers (*The Guardian* and *The Observer*) were also analysed. These short, argumentative pieces are perhaps the closest readily-available parallel to the short compositions produced by the learners.

A total of 96 texts were analysed: 24 long native speaker texts (hereafter referred to as ‘NS Long’), 24 long non-native texts (‘NNS Long’), 24 short native speaker texts (‘NS Short’) and 24 short non-native texts (‘NNS Short’). Table 2 describes the four sets of texts in detail.

Table 2: summary of texts analysed in the study

type	sub-type	description	number of texts	number of writers	total words	mean words/text	writers' L1
NS Long	Prospect	essays from the 'international' section of the current affairs journal *Prospect*	12	12	41304	3442	English
	Academic	academic essays written by students on the MA programme in Applied Linguistics at the University of Nottingham. 2 essays each were taken from 6 different MA courses	12	7	37429	3119	English
NNS Long	British EAP Project	research projects written by non-native students as part of their final assessment for a pre-sessional course in EAP at Durham University. Essay topics are taken from a variety of subject areas, reflecting the academic interests of the students (7 business finance/management; 3 law; 1 classics; 1 political science)	12	12	39145	3262	7 Mandarin 1 Arabic 1 French 1 Greek 1 Korean 1 Russian
	Turkish EAP Project	academic essays written by non-native students for an in-sessional course in EAP during the first year of their degree at Bilkent University, an English medium institution in Turkey. 6 come from a course based around the themes of the nature-nurture debate and philosophical concepts of personal identity; 6 were from a course based on the philosophy of happiness.	12	12	33217	2768	Turkish
NS Short	Opinion articles	opinion articles from *The Guardian* and *The Observer* newspapers	12	12	8401	700	English
	LOCESS essays	timed essays (1 hour) on the topic of European integration written by British undergraduates	12	12	6734	561	English
NNS Short	British short essays	short compositions written by postgraduate students on a pre-sessional course in English for Academic Purposes at Durham University. 6 compositions were on the topic of *Consumerism*; 6 were on the topic of *Education*.	12	6 (all of these writers are also represented in 'British EAP Project')	7936	661	5 Mandarin 1 Russian
	Bulgarian subcorpus of ICLE	short argumentative essays written by students at Sofia University "St Kliment Ohridski". All writers were reported to have spent two years studying English at university level.	12	12	6860	572	Bulgarian

Procedure

Identification of word combinations

The present analysis, like that in Section 5.3, was limited to directly adjacent premodifier-noun word pairs (including both adjective-noun and noun-noun combinations). Modifier-noun combinations were chosen because they were found to be particularly common in the texts analysed, and so provided a rich source of data.

All such pairs were manually extracted from the texts. No attempt was made to filter out pairs which might be considered words in their own right (*e.g. prime minister; martial arts*); such pairings are taken simply to represent one extreme on the scale of collocational fixity.

Combinations were not included if they contained one of the following elements:

- proper nouns (identified by capitalization);
- acronyms defined in the paper (e.g. 'CCT' for 'cross-cultural training')
- pronouns;
- possessives;
- semi-determiners – as listed in Biber et al. (1999), i.e.: *same, other, former, latter, last, next, certain, such*;
- numbers/ordinals.

Since the study aims to draw conclusions regarding the performance of the writers themselves, quotations were not included in the analysis.

To keep the calculation of association measures relatively straightforward, only directly adjacent word pairs were included in the analysis (see Section 4.3 for a discussion of the problems of comparing associations measures for collocations with differing spans). Thus, where more than one adjective modifies a noun (e.g. *beautiful green eyes*), only the final adjective-noun pair (*green eyes*) is included. Where a premodifying noun is itself premodified, only pairs in the group where the modifier can be read as modifying the succeeding noun itself are included: e.g. from the phrase *national security adviser*, two collocations are extracted, *national security* and

security advisor; in *local power plant workers, power plant* and *plant workers* are recorded, but not *local power* since *local* doesn't modify *power.*

This procedure retrieved a total of 10,839 word combinations from the 96 texts. The total number of combinations for each text type and the average numbers of combinations retrieved for each text are shown in Table 3. Since different text types were of characteristically different lengths, Table 3 also shows these averages normalised to combinations per 1,000 words of text.

Table 3: summary of combinations retrieved

type	sub-type	total combinations retrieved	average combinations/text	average combinations/1000 words
NS Long	Prospect	2845	204.25	59.34
	Academic	1500	196.42	62.97
NNS Long	British EAP Project	2451	237.08	72.68
	Turkish EAP Project	2357	125.00	45.16
NS Short	Opinion articles	513	40.42	57.73
	LOCNESS essays	296	24.67	43.96
NNS Short	British short essays	485	42.75	64.64
	Bulgarian subcorpus of ICLE	392	32.67	57.14

Calculation of collocational frequency

Three different measures were taken of the frequency of these collocation in the BNC: raw frequency[4], t-score, and mutual information (MI) (see Section 4.3). As we saw in Chapter 4, t-score and MI are both widely used measures in lexicography, but tend to emphasise rather different sets of collocations. In particular, whereas rankings based on t-scores tend to highlight very frequent collocations (and so are similar to rankings based on raw frequency), MI tends to give prominence to word pairs which may be less common, but whose component words are not often found apart (Stubbs, 1995). Thus, pairs like *good example, long way*, and *hard work* attain high t-scores but low MI scores, while pairs like *ultimate arbiter, immortal souls* and *tectonic plates* attain

[4] The program for extracting frequency data about the target word combinations (i.e. the frequency of each word and of each word pair) from the BNC was developed by Jakup Marecek of the University of Nottingham School of Computer Science and Information Technology. This program did not use lemmatisation or part of speech information.

the reverse. With this in mind, both association measures were used with the intention that they might provide different types of information about pattern of use.

It has been suggested that a t-score of 2 or above and/or a MI score of 3 or above may be indicative of collocation (e.g., Hunston, 2002; Stubbs, 1995). The present study will take these values as minimum conditions for collocation. However, simply dividing combinations into 'collocations' vs. 'non-collocations' on this basis would not be satisfactory, since this would disguise the evident difference between combinations which narrowly pass the threshold (e.g. *remarkable book; sweet child*) and much stronger collocations (*ethnic minorities; global warming*). Combinations will therefore be classified across a scale of collocational strength. This approach of using association measures to grade collocations, rather than simply dividing items into collocates vs. non-collocates accords with the view that association measures are best used to provide ranked lists of collocational strength, rather than to demarcate clear categories (see Section 4.3). Moreover, by looking at the spread of collocational strength we can get a much more fine-grained view of the data than would be possible on the basis of a simple division of combinations into 'collocations' and 'non-collocations'.

The extracted collocations were divided into 7 bands of t-score, as follows:

t = 2-3.99; t = 4-5.99; t = 6-7.99; t = 8-9.99; t = 10-14.99; t = 15-19.99; $t \geq 20$

Piloting showed this banding to provide a maximally fine differentiation whilst maintaining a reasonably high number of instances for each level. Similarly, the MI scores were divided into the following bands:

MI = 3-3.99; MI = 4-4.99; MI = 5-5.99; MI = 6-6.99; MI = 7-7.99; MI = 8-8.99; MI = 9-9.99; MI ≥ 10

Because association measures are thought to be unreliable for low frequency collocations, and because corpora cannot provide stable evidence for infrequent events (Stubbs, 2001), combinations appearing in the BNC fewer than five times were not assigned t-scores or MI scores (see Results section).

Group vs. individual scores

Previous analyses of native vs. non-native writing (Section 5.2) have worked by pooling the writing of large number of learners and large numbers of natives into separate 'native' vs. 'non-native' corpora and comparing the two as wholes. This approach may be problematic in that it is not clear to what extent results will mask variability between different learners (a point acknowledged by Howarth (1998, p. 177)). If there are regular and stable norms in the extent to which natives and non-natives make use of formulas, this is not a problem. However, such regularities have not been established for either group. Without established norms, and given that variability seems to be the rule in most second language learning and use, the significance of the averaged-out figures which these studies present is not clear.

The present analysis aims to overcome this problem by recording results individually for each text and then comparing the four groups of texts using standard inferential statistics, taking each text as an individual case. The difference between this and previous approaches can be understood with an example. The first part of the analysis looks at the proportion of combinations which are rare in English (appearing fewer than five times in the BNC). To describe this, a whole-corpus approach would simply find one set of figures for each of the four sets of texts. On the approach taken here, a separate figure is instead calculated for each of the 96 texts. An average is then taken for the 24 texts within each type. The advantage of this approach is that we record not only an average figure for each text type, but also the degree of variation between texts. This enables us to use inferential statistics to find whether texts of one type contain a significantly higher percentage of infrequent collocations than those of another. Significant scores on these tests will indicate relative homogeneity within groups and meaningful differences between them.

Results and discussion

Low frequency combinations

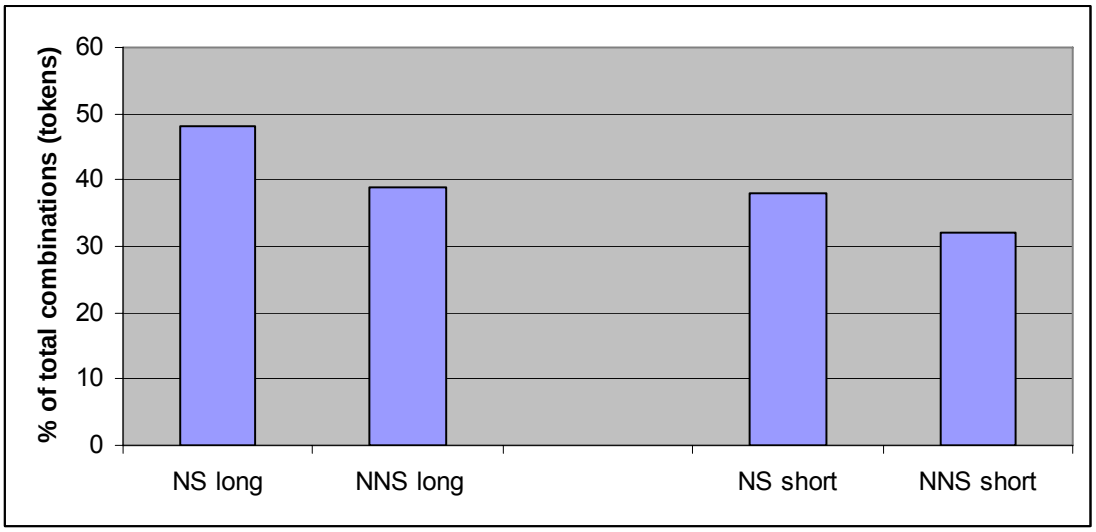
As a first stage of analysis, we can ask to what extent native and non-native writers make use of combinations which are rare in British English. An obvious way of analysing the prevalence of such pairs would be to look at the average number used in a given length of text. While this sort of analysis would give an indication of the

'density' of use of rare collocations, it has the disadvantage of confounding the extent to which writers use such collocations with the extent to which they use premodifier–noun constructions in general. Thus, as Table 3 indicates, texts from the group 'British EAP Project' use this construction to a much greater extent than do those from the group 'Turkish EAP Project'. Given this, it is not clear whether a finding that rare combinations are more common in the former than in the latter (as indeed they are) is due to a greater degree of reliance on such collocations or is merely a product of their greater use of modifier-noun constructions overall. This problem can be overcome by looking not at the total number of collocations used, but rather at the percentage of premodifier-noun combinations which are strong collocations. This analysis should give a more valid representation of the degree to which writers rely on conventional collocations.

Figure 2 shows the percentage of combinations used in each set of texts which appear fewer than 5 times (or which fail to appear at all) in the BNC. The mean percentage of combinations falling into this category in the long native texts is 48%, while for the long non-native texts the figure is 38%, a substantial and statistically significant difference (NS M = 48.19, SE = 2.14, NNS M = 38.87, SE = 1.52, t(46)= 3.55, p (two-tailed) < .001, r = .46). The shorter texts use in general a lower proportion of low frequency combinations, but show a similar pattern - i.e., low frequency items are more prevalent in native than in non-native texts, though in this case the difference is not statistically significant (NS M = 38.14, SE = 3.39, NNS M = 31.95, SE = 2.63, t(46)= 1.42, p (two-tailed) > .05, r = .21).

These results suggest a certain conservatism in the use of collocations by non-native writers. They are far less likely than natives to coin novel modifier-noun combinations, preferring instead to rely on pairings which are likely to have been attested in their input.

Figure 2: mean percentage of combinations which appear < 5 times in BNC



Strong collocations

T-score analysis

The main focus of this study is on the use of 'strong' collocations. As a first method of quantifying this, the percentage of pre-modifier – noun combinations falling into each t-score band was calculated for each text. Figures 3 and 4 summarise the results of this analysis, showing the median percentage of collocation tokens found at each level for long and short texts respectively (median percentages are used here because the distribution of percentages is not normal within all bands). Since a large number of combinations either appeared in the BNC fewer than 5 times or attained a t-sore of less than 2, the bandings do not sum to 100%.

Figure 3: median % of collocation (tokens) found at different levels of t-score for long texts

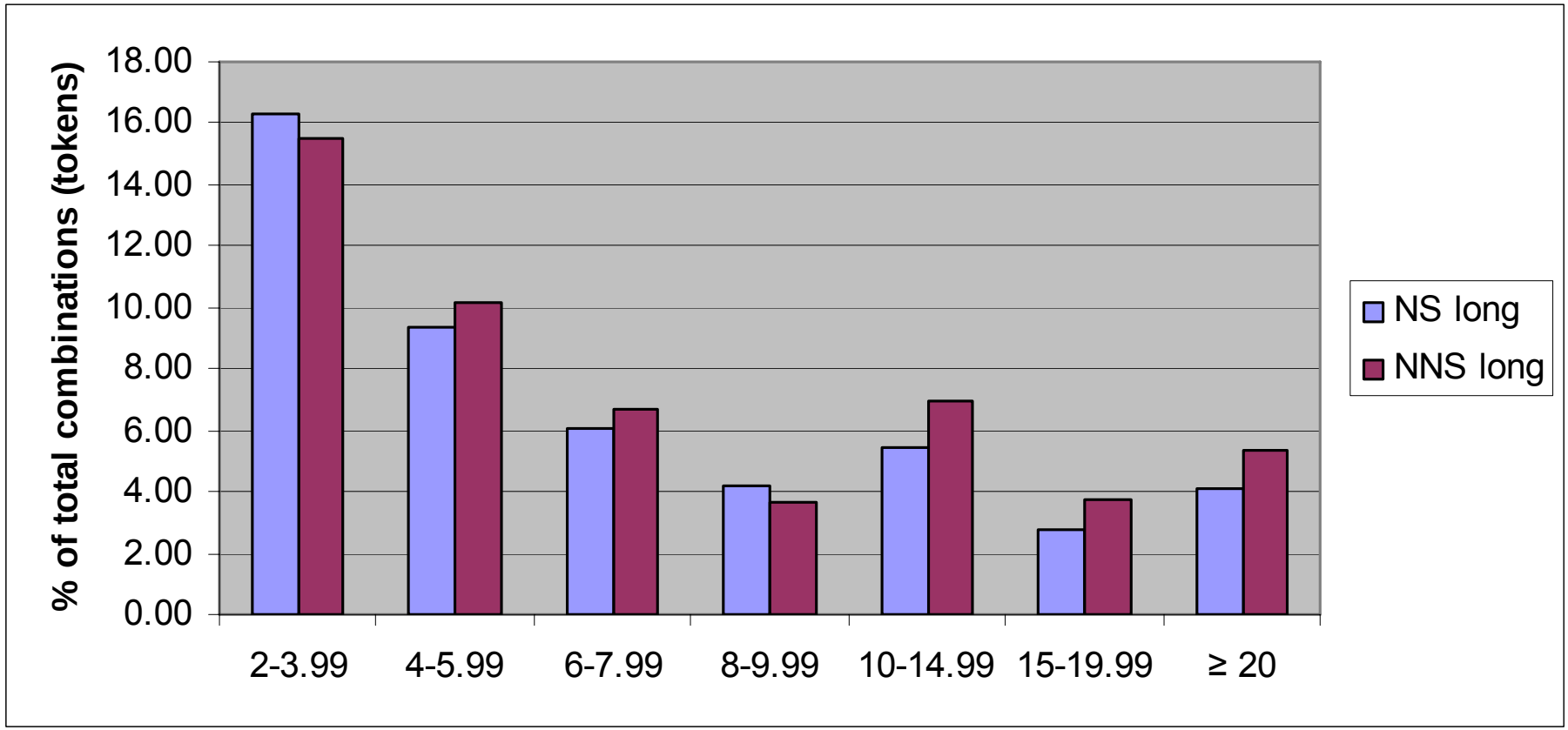
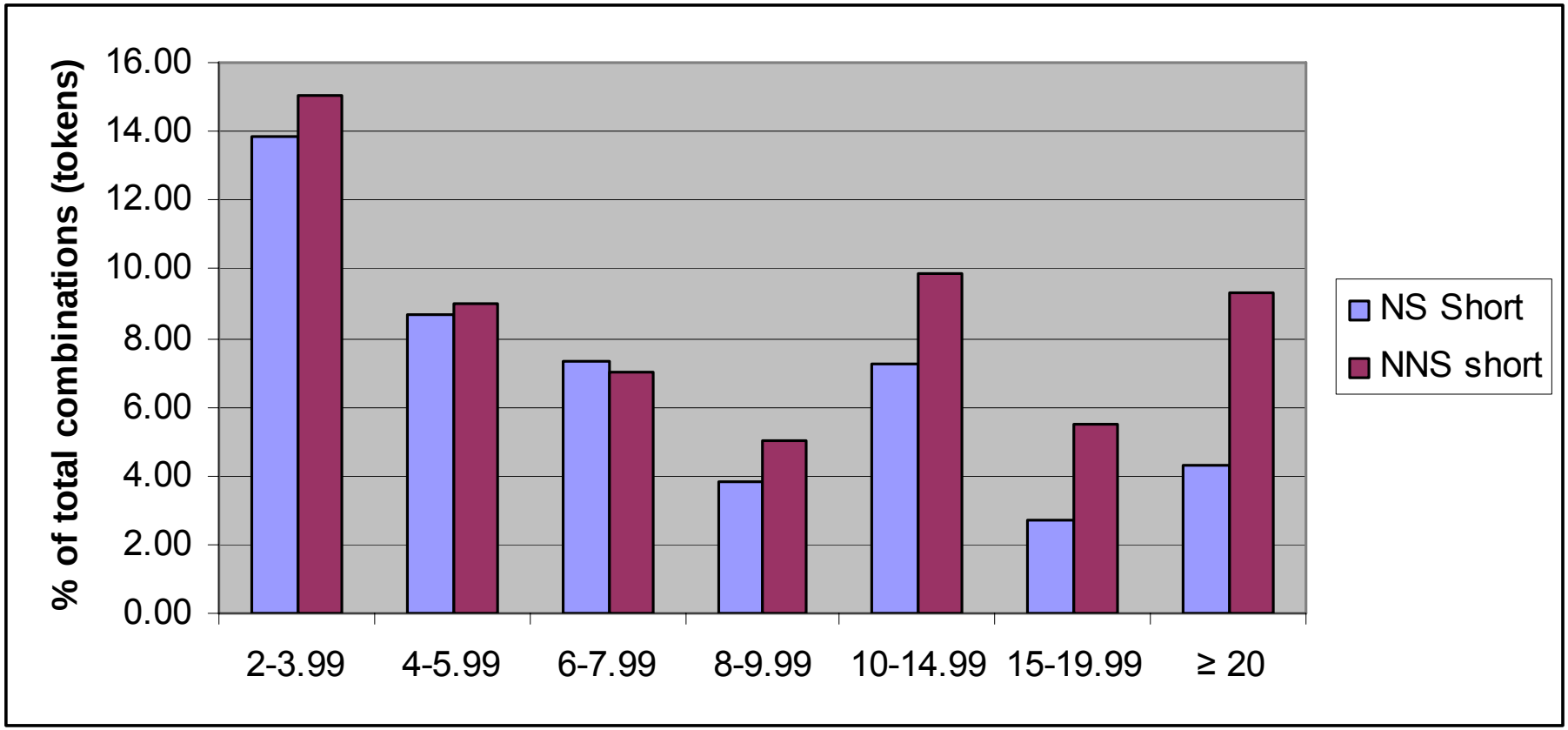


Figure 4: median % of collocation (tokens) found at different levels of t-score for short texts



Looking first at differences between the longer native and non-native texts, it would appear from Figure 3 that non-native writers take a rather higher proportion of their collocations from the highest bands ($t \geq 10$) than natives. At lower levels, usage appears similar between the two groups of texts. Collapsing the bands into broader 'high' ($t \geq 10$) vs. 'low' ($t < 10$) groupings, enables us to confirm this trend. Non-natives take, on average, 20% of their collocations from the 'high' band, compared to only 14% for natives. According to an independent samples t-test, this difference is significant at the $p < .005$ level (NS $M = 13.52$, $SE = 1.42$, NNS $M = 20.16$, $SE = 1.55$, $t(46) = -3.153$, p (two-tailed) $< .005$, $r = -.42$). At the other end of the scale, there is no significant difference between the two sets of texts in their use of the lower strength collocations (NS $M = 35.69$, $SE = 1.27$, NNS $M = 37.14$, $t(46) = -0.796$, p (two-tailed) $> .05$, $r = .12$).

We saw in Section 5.2 that some researchers have claimed that non-native writing is characterised by the repeated use of a small repertoire of collocations (Granger, 1998; Kaszubski, 2000; Lorenz, 1999). That the non-native texts in our data make greater use of repetition than the natives can be confirmed by calculating a collocational type-token ratio (calculated as the mean number of collocation types per 100 collocation tokens) for each text. The median ratio for long native texts is 90, compared with 63 for non-natives. The median ratio for short native texts is 96, compared with 90 for non-natives (note that type-token ratios are typically higher for shorter texts (Richards, 1987)). It may be then, that the non-native writers' comparative 'overuse' of strong collocations comes about because they rely on repeating a few favoured formulas. To check whether this is the case, we can recalculate our data using collocation types rather than collocation tokens. Such an analysis can be interpreted as telling us about the repertoire of collocations demonstrated by each writer.

Using these data to re-examine the differences described above, we find that the pattern of non-native overuse is indeed weakened somewhat. In this case, non-natives continued to take a higher proportion of their collocations from the $t \geq 10$ band than natives, but the difference is now much smaller and marginally nonsignificant (NS $Mdn = 11.70$, NNS $Mdn = 14.26$, $U = 200.00$, p (two-tailed) $= .07$, $r = -.26$; non-parametric tests are used because results were not normally distributed within the long non-native texts). Any non-native overuse of the strongest collocations may therefore

be the result of the repeated use of favoured items. However, even when repetition is removed from the data, it is fairly clear that non-natives make no less use of strong collocations than natives.

Turning now to the shorter texts, Figure 4 seems to indicate a pattern similar to that seen for natives vs. non-natives as a whole – i.e. relative overuse by non-natives at the higher levels. Again collapsing the results into high ($t \geq 10$) vs. low ($t < 10$) bands, we find significant overuse of high scoring combinations by non-native speakers (NS *Mdn* = 18.34, NNS *Mdn* = 26.60, *U* = 190.00, *p* (two-tailed) < 0.05, *r* = -.29; non-parametric tests are used because results for short native speaker texts were not normally distributed). Again, the difference is weakened if we look at collocation types rather than tokens (NS *M* = 18.95, *SE* = 2.44, NNS *M* = 22.81, *SE* = 1.51, *t*(46) = -1.345, *p* (two-tailed) > 0.05, *r* = .19).

It appears, then, that non-native writers in general use make at least as much use of collocations with high t-scores (i.e., $t \geq 10$) as natives. We have seen that t-score correlates very strongly with raw-frequency (the chief difference between the two measures being that the former relegates some items composed of very high frequency words). Taking these results together with those from the analysis of low frequency items, therefore, we can say that non-native writing is characterised by extensive use of collocations which are likely to have been attested in learners' input, and in particular of those word pairs which are likely to have occurred with high frequency. This contrasts strongly with Kjellmer's (1990) notion that non-natives fail to use formulaic language, and is difficult to reconcile with Wray's notion that non-natives do not recall the collocations they have seen.

Mutual information analysis

Mutual information is known to emphasise a rather different set of collocations from t-scores, so I also carried out a similar analysis using the MI procedure. Figures 5 and 6 summarise the results of this analysis, showing the median percentage of collocation tokens found at each level for long and short texts respectively.

Figure 5: median % of collocations (tokens) found at different levels of MI for long texts

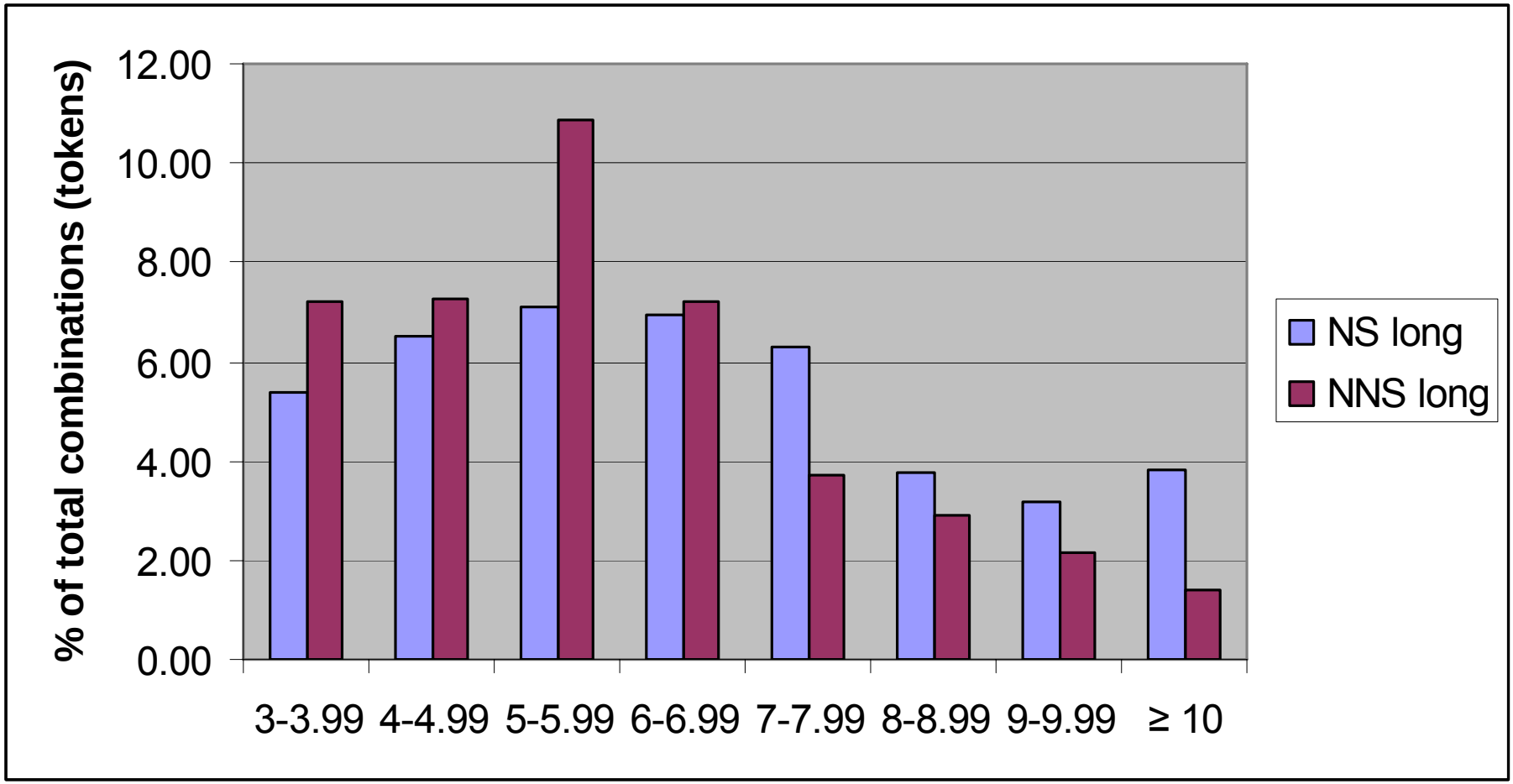
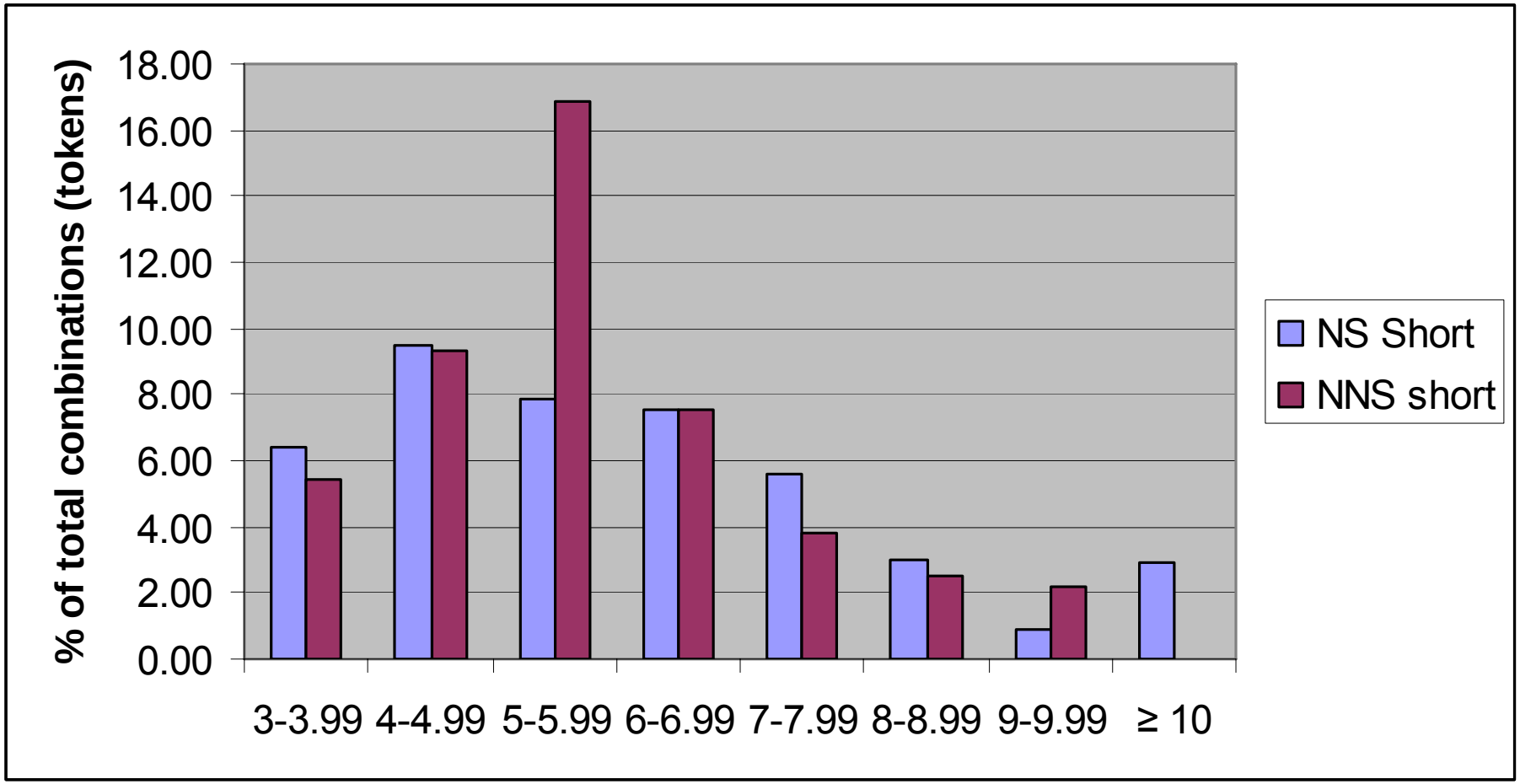


Figure 6: median % of collocations (tokens) found at different levels of MI for short texts



Again we can start by looking at the differences between the longer native and non-native texts. Reversing the results seen for the t-score analysis, Figures 5 appears to indicate that non-native writers relied to a lesser extent on very strong collocations than did natives. In particular, non-natives show a consistent pattern of 'underuse' at all levels MI ≥ 7. An independent samples t-test shows the difference between native and non-native use of MI ≥ 7 collocation tokens not to be significant (NS M = 17.48, SE = 1.30, NNS M = 14.95, SE = 1.46, $t(46) = 1.289$, p (two-tailed) > .05, r = .19). However, if we look at the percentage of collocation *types* taken from these levels, the difference becomes highly significant (NS M = 15.47, SE = 1.00, NNS M = 11.07, $t(46) = 3.386$, p (two-tailed) < .001, r = .45). As before then, non-native use of the stronger collocations seems to have been boosted by repetition. Taking a slightly more exclusive band of strong collocations (MI ≥ 8), the difference between the two sets of texts is more emphatic: non-natives show significant underuse of items from these bands in both the analysis by tokens (NS Mdn = 11.08, NNS Mdn = 8.32, U = 184.50, p (two-tailed) < .05, r = -.31; non-parametric tests are used because results for long non-native speaker texts were not normally distributed) and that by types (NS M = 9.75, SE = 0.71, NNS M = 5.85, SE = 0.58, $t(46) = 4.236$, p (two-tailed) < .001, r = .53).

The shorter texts exhibit a similar, though slightly less robust, pattern. Non-natives show a nonsignificant underuse of strong collocation tokens (MI ≥ 8) in comparison to native norms (NS Mdn = 11.62, NNS Mdn = 6.29, U = 218.5, p (two-tailed) > .05, r = -.21; non-parametric tests are used because results for short non-native speaker texts were not normally distributed), but this difference reaches significance in the analysis of types (NS M = 11.43, SE = 1.30, NNS M = 7.88, SE = 1.01, $t(46) = 2.159$, p (two-tailed) < .05, r = .30).

These findings stand in interesting contrast to those for low frequency and high t-score collocations. Word pairs with high MI scores are characteristically of lower frequency than those attaining high t-scores, but tend to be strongly associated with each other, in that where one part of the collocation is found, the other is very likely to be nearby. Examples from the texts analysed here are *densely populated, bated breath,* and *preconceived notions.* Whatever their approach to learning, it is unsurprising that learners should be slow to pick up such low frequency pairs. It is also worth noting

that such striking combinations seem likely to be highly salient for native speakers. It may well be items of this sort which mark out language as particularly idiomatic (Lorenz, 1999, p. 184). If this is right, their absence here would explain Kjellmer's intuition that non-native language lacks phrasal authenticity. The problem is not that language is missing high frequency phrases, but rather that it makes too little use of these lower-frequency but strongly-associated items which are indicative of native-like production.

Summary and conclusions: collocations in non-native writing

This study has aimed to describe the extent to which non-native writers make use of word combinations, and particularly strong collocations, in comparison to native speaker norms, by using methodologies which take advantage of frequency information, and which take account of individual variability between texts. Three main findings have emerged. Firstly, native writers use more low frequency combinations than non-natives. This trend appears to be fairly consistent across texts, even though it was statistically significant only in the comparison of longer texts. Secondly, non-native writers make at least as much use of collocations with very high t-scores as do natives. Since non-natives also tend to repeat certain favoured collocations, if we consider collocation tokens, rather than types, they show a significant overuse of these strong collocations in comparison to native norms. Thirdly, non-native writers significantly underuse collocations with high mutual information scores in comparison with native norms. Again, the repetition of favoured items bolsters the non-native count somewhat, so the difference is more marked on an analysis of collocation types. All of these regularities were less marked in shorter texts, but even here we found sufficient consistency of usage for the same tendencies to emerge, if not always with statistical significance.

I argued above that, if Wray's model were correct, we would expect non-native writers to make much less use of high frequency collocations than natives. I also argued that, if, in contrast to Wray's thesis, second language learners retain and use those sequences of language which are frequent in their input, we would expect non-native writers to rely especially on the most high frequency collocations. The pattern of results presented here supports the latter position, suggesting that the retention of collocational information shown by adult learners in Section 5.3 does carry over into

the normal, long-term second language acquisition process. Taken together, then, this and the previous study suggest that Wray's thesis that adult second language learners fail to acquire appropriate collocations because they take a more word-oriented approach to learning than child L1 learners is not correct.

At the same time, these results enable us to account for Kjellmer's sense that there is something inauthentic about the phraseology of second language learners. The problem is not that they fail to use formulaic language altogether, but rather that they avoid those items which are of relatively low frequency, but which are strongly associated. Since these items are probably highly salient for natives, their absence may give a strong impression in unnaturalness. However, it is not necessary to posit any radically different L2 learning mechanism to explain this absence; their characteristically low frequency of occurrence simply means that such collocations are likely to be acquired later than other parts of nativelike phraseology.

5.5 Summary and conclusions: collocation learning from input

This chapter set out to examine Wray's (2002) claim that adult language learners do not acquire the collocations to which they are exposed because their mature, literate cognitive systems, together with various situational pressures, push them to focus their attention on individual words, rather than meaningful chunks. I argued that this claim cannot be properly evaluated by simply focusing on the end results of adult learning, since it is not clear whether any shortcomings in collocational knowledge are a product of an alternative learning approach or of insufficient input. The studies in this chapter have therefore attempted to link adults' knowledge of collocation with the input they have received.

Linking input with knowledge presents the researcher with an observer's paradox: the more tightly we control the input learners receive, the more likely we are to distort the natural learning situation. With this in mind, I approached the question using two complementary experimental paradigms. In the first, lab-based, approach, it was possible to exercise very tight control over input, but the learning process may not have been entirely natural, and only short-term learning could be traced. In the second,

corpus-based, approach, a more natural and longer-term learning situation was studied, but it was only possible to make a rough estimate of learners' likely input.

The results of both studies suggest that adult learners do acquire at least some of the collocations to which they are exposed. The lab-based study (Section 5.3) showed that learners who were asked to perform an entirely formal task (reading a sentence aloud) with no knowledge that they were expected to learn anything from this exposure, retained information about which words appeared together. Limited repetition of the task increased this retention dramatically. This suggests that collocation learning may be an automatic process, which will continue regardless of any strategies adopted by the learner, but that techniques such as fluency-oriented re-reading may hasten this learning. The corpus-based study (Section 5.4) found that advanced non-native speakers of English make at least as much use of high frequency collocations as natives. However, their failure to use lower-frequency, strongly associated word pairs may create a superficial impression that they are avoiding formulaic language. This pattern of results is, I argued, compatible with a model whereby learners extract the most frequent collocations from the input they meet. It does not seem to be consistent with the idea that learners fail to remember the collocations they encounter.

Taken together, these results suggest that adult second language learners are capable of learning collocations implicitly from input. However, this does not mean that they typically do so. The distinctive pattern of collocation use found in Section 5.4 suggests that these learners had some way to go in their collocation learning. The most likely reason for the problem seems likely to be a lack of sufficient input. Adult second language learners typically have far less exposure to the target language than native speakers. This relative sparsity of input may mean that for lower frequency collocations the gap between repeated exposures is too great for the necessary representations to become entrenched. If this is the case, then learners will require special instruction in collocation, either through explicit teaching or artificially enriched input. If such input is to be provided, of course, teachers will need to know what collocations their learners need to learn. The next chapter will explore the possibility of identifying key collocations for learning in one particular area of language – English for academic purposes.

Chapter 6
Constructing a pedagogical listing of academic collocations

6.1 Introduction

Language teachers have long made use of word lists as a means of focusing students' vocabulary learning. It is known that the vast majority of language use is composed of a relatively small number of very high frequency words – Nation (2001, p. 11) reports that the 2,000 word families of West's (1953) *General Service List* (*GSL*) account for around 80% of naturally occurring text in general English – so focusing on these high frequency items seems likely to pay substantial dividends for novice learners. Within English for Academic Purposes (EAP), there has been much interest in constructing lists of generic 'academic' vocabulary - 'sub-technical' (Yang, 1986) words which are common across academic disciplines, but which may cause problems for learners because they are neither sufficiently frequent in the language as a whole to be learnt implicitly nor part of the technical lexicon which is likely to be explicitly taught as part of subject courses (Nation, 2001, pp. 189-191). The most commonly-used listing of academic vocabulary today is Coxhead's *Academic Word List* (*AWL*), a collection of 570 word families which are claimed to account for approximately 10% of the words found in academic texts (2000).

The present chapter considers whether it is feasible to extend the academic word list approach into the realm of formulaic language by constructing a listing of 'academic collocations' – i.e. collocations that will be of use to students from across a wide range of academic disciplines. We saw in Chapter 4 that frequency information from a corpus does appear to give useful information about the collocations which are likely to be psychologically real for members of a language community, suggesting that data of this kind may enable us to pick out sets of collocations which will be useful targets for learners. Chapter 5 showed that, though adult second language learners can acquire some collocational associations from input, they are also likely to benefit from more targeted input. We saw both that the profile of collocations used by advanced non-

native speakers is characteristically different from that of natives, suggesting that the input to which such learners have had access may not be sufficient for effective acquisition, and that targeted repetition of collocations can dramatically increase learning. Taken together, these conclusions argue for an explicit teaching focus on collocation, and so suggest that a frequency-based listing of the collocations to which learners should be directed may be a useful resource.

However, there are also a number of problems for the idea of an academic collocation list. One is that it is not yet clear how academic collocations should be defined or identified. Academic word lists have relied on identifying words which occur with high frequency in academic texts, but which do not appear on listings of basic vocabulary. However, in the absence of any listings of 'basic collocations' or any standards for what should count as 'high frequency' in this context, it is not clear how this method should be applied for collocations. More fundamentally, we do not yet know whether any identifiable set of collocations exists which would be of genuine use to students from across the full range of academic disciplines. Research has often emphasised the topic- and genre-specificity of collocation, suggesting that many collocations may be so domain-specific that the idea of generic 'academic collocations' will not be a viable one. A final issue is that a listing of two-word collocations may not be sufficient for learners' needs – much of the phraseology that learners need to acquire are likely to be associations of more than two words, and it is possible that a two-word listing would give a misleading impression of the phrasal learning task. It is not yet clear, however, how such longer items can be identified.

The present chapter aims to explore these issues by constructing and evaluating two different collocation lists. Section 6.2 provides some background by describing previous work on academic word lists and academic collocation and discussing some of the issues that will be involved in the construction of an academic collocation list. Section 6.3 describes the design and compilation of a large academic corpus which will be used in attempting to identify academic collocations. Section 6.4 introduces two alternative ways in which an academic collocation list could be generated, while Section 6.5 evaluates the products of each approach. Finally, Section 6.6 discusses the limitations inherent to listings of two-word collocations and considers some ways in

which future research might move beyond word pairs to incorporate larger combinations.

6.2 Academic word lists and academic collocations

Academic word lists

Academic vocabulary can be distinguished from, on the one hand, 'basic' and, on the other, 'technical' vocabulary (Nation, 2001, pp. 11-12). Basic vocabulary consists of words that are frequent throughout the language. These are items of the sort appearing on West's (1953) *General Service List* of 2,000 important words in English. These high frequency items account for a large percentage of the language we meet on a daily basis and, as such, are taken to be of high priority for elementary learners. Technical vocabulary, in contrast, consists of words which are closely associated with particular subject areas. These tend to be found with moderate or high frequency in a narrow range of texts, but are rare elsewhere. Academic vocabulary falls somewhere between these two types. These are items which are neither sufficiently frequent in general to be part of basic vocabulary nor tied to specific disciplines, as technical vocabulary is. Though their relatively high frequency in academic writing means that such items will be important for EAP learners, it has been argued that their 'intermediate' status may cause difficulties: such items are neither sufficiently frequent to be learned implicitly or to form part of the basic education students can be assumed to have already encountered, nor sufficiently central to students' subject areas to be taught by subject teachers or to stand out as particularly salient in the language they encounter (Nation, 2001, pp. 189-191).

There have been a number of attempts at compiling listings of generic academic vocabulary. The general approach has been to identify those words which are a) frequent across a wide range of academic disciplines, but b) not part of basic vocabulary. Early attempts, (e.g, Campion & Elley, 1971; Praninskas, 1972) were compiled by counting words manually, and so were limited to relatively small samples of language. However, developments in automated analysis have enabled far larger corpora to be investigated. An early attempt along these lines was that of Yang (1986), whose primary interest was in technical vocabulary, but who also provided formulae

for identifying what he calls 'sub-technical' items. The most influential listing to date, however, is Coxhead's *Academic Word List* (2000).

Coxhead's listing is based on the analysis of a 3.5 million word corpus of academic articles, textbooks and lab manuals, sampled equally from across the areas of arts, commerce, law, and science. Words are included in Coxhead's listing if they:

- do not appear in West's (1953) *General Service List;*
- occur at least 10 times in each of the four subject areas;
- occur in at least 15 of 28 sub-areas;
- occur at least 100 times in the corpus as a whole.

An important part of Coxhead's methodology is the grouping of words into 'families' of related forms. That is to say, the various inflectional forms of a word, and forms differing only in regular productive affixes, are taken to constitute a single item. Thus, the headword *concept* incorporates the forms *conception, concepts, conceptual, conceptualisation, conceptualise, conceptualised, conceptualises, conceptualising*, and *conceptually*. Any occurrence of one of these forms goes towards the count for the item as a whole. In this, Coxhead follows the example of both West (1953) and Xue and Nation (1984). While a certain amount of information is no doubt lost in this way, since it is not clear which form of a word learners are most likely to need, Coxhead justifies the move on the grounds that such groupings form important units in the mental lexicon and that, when one form has been learned, comprehension of the others is facilitated (Coxhead, 2000, pp. 217-218).

Coxhead's research yielded 570 academic word families (made up of 3,110 individual word forms) which, she claims, account for 10% of words encountered in academic writing. The list appears to be slightly more useful for some subject areas than for others: 12% of the commerce sub-corpus is covered, in comparison to 9.4% of law, 9.3% of arts and 9.1% of science. However, even the lowest of these figures remains an impressive return and Coxhead stresses that almost all words (94%) are found in more than 20 of the 28 subject areas included in the corpus. A note of caution is sounded by Hyland and Tse (2007), however. They report that in their corpus of academic writing - from engineering, science, and social science - many AWL items

have only a limited spread across disciplines, that many have the vast majority of their occurrences in one subject grouping only, and that few of the most frequent words within any one area are common in any of the others. They also claim that many AWL items have different meanings and different characteristic collocations in different areas. Based on these considerations, Hyland and Tse call into question the very notion of a substantial, cross-disciplinary, academic vocabulary. The basic thesis that academic writing has a sufficiently homogenous core vocabulary to make an academic word list viable remains, then, an open question which stands in need of further investigation. It is hoped that the present investigation into academic collocations will make some contribution to this debate.

Academic collocations

The central claim of the formulaic language movement is that the mental lexicon does not consist solely of words, but also includes larger chunks of language. If we accept this claim, purely word-based lists begin to look inadequate as a guide to vocabulary learning. Coxhead (2008, p. 152) has acknowledged this problem, noting that the absence from the AWL of any information about phrasal patterning may lead students to ignore such patterns and focus their learning exclusively on individual words. More fundamentally, Hyland and Tse (2007) have claimed that by ignoring the distinctive collocational patterns found in different disciplines, the AWL may overestimate the homogeneity of vocabulary use across subject areas, and so misrepresent the learning task. They see divergent collocational patterning as one factor undermining the notion of generic 'academic' vocabulary.

These considerations suggest a need to extend listings of academic vocabulary beyond the single-word level. There are, however, problems with the idea of an academic collocation list. Much of the appeal of traditional word lists derives from their ability to 'cover' a large percentage of the vocabulary learners need with a small number of items. Collocations, however, are on average more numerous, rarer, and more tied to specific areas of discourse than words. This means that no listing of collocations will be able to emulate the impressive coverage statistics of word lists. It remains to be seen whether such a listing would remain attractive to teachers and learners. Moreover, for academic vocabulary, as Hyland and Tse's (2007) critique suggests, it is not even obvious that collocation use will be sufficiently consistent across disciplines for a

worthwhile generic listing to be identifiable. Research is therefore needed to determine whether collocation lists in general – and an academic collocation list in particular – are viable projects.

While a number of detailed studies have been made of the 'technical' collocations found in specific academic disciplines (e.g., Cortes, 2004; Gledhill, 2000; Marco, 2000; G. C. Williams, 1998; Yang, 1986), there are as yet few studies looking at collocations in academic language as a whole. Biber and his colleagues (Biber, 2006; Biber, Conrad, & Cortes, 2004; Biber et al., 1999) looked at the use of 'lexical bundles' – frequently recurring fixed sequences of words – in a range of different types of 'university language'; a heading that includes not only academic discourse types such as textbooks and lessons but also 'non-academic' language such as that of 'service encounters' and 'institutional writing'. This research has yielded some interesting results; however, the corpus used, which includes only 760,000 words of academic writing in total, is – as Biber acknowledges (2006, pp. 163-164) – rather too small to support robust claims about the usages of particular disciplines, and so to determine which clusters are used consistently across fields. Moreover, the limitation of the analysis to fixed multi-word sequences is likely to leave out much that is of collocational interest. Collocation, as it has been studied in this thesis, often involves relationships between words which may be separated by other, non-fixed, or semi-fixed words, and which may differ in their position relative to one another. Compare:

> *he made a **powerful argument;***
> *he made a **powerful**, but ultimately unconvincing, **argument;***
> *his **argument** was a **powerful** one*.

Such collocations are of great interest, but will be missed by the lexical bundle approach.

Ellis et al (2007) also take a 'lexical bundle' approach to identifying academic phrases. They compare a 2.1 million word corpus of academic speech and a 2.1 million word corpus of academic writing with a 2.9 million word corpus of non-academic speech and a 1.9 million word corpus of non-academic writing to find three- four- and five-word bundles that are significantly more frequent in the former than in the latter. Ellis

et al make some progress on the issue of determining how well spread these bundles are across academic disciplines. They divide their two academic corpora into nine 'genres':

- Spoken: HumArts, SocSci, BioSci, PhysSci/Engin, Other
- Written: HumArts, SocSci, NatSci & Med, Technol/Engin

Phrases are then graded according to the number of different genres in which they attain a frequency of four per million words. While this approach is helpful, it still has important shortcomings. One problem is that it is not clear on what basis the different genres were arrived at or whether these divisions truly reflect the full variation in phrase use across the corpora (see Section 6.3 for a fuller discussion of this point). A second issue is that the reproduction of some areas in both spoken and written corpora will mean that higher scores may not indicate greater spread across disciplines, but rather consistent use across writing and speaking within a few disciplines.

In sum, though the phenomenon of formulaic language suggests that pedagogical listings of academic words may at best fail to provide much of the information learners need, and at worst systematically misrepresent the vocabulary learning task, the feasibility of a phraseological listing remains relatively unexplored. Indeed, it is still not clear whether a listing of such items would be of use to learners, or even whether a substantial generic academic phraseology exists. While some progress has been made in identifying and describing 'lexical bundles' in academic language, this has left out of the picture the great many collocations which are positionally flexible. Since such collocations seem likely to constitute a large proportion of what is interesting in academic phraseology, this is an important shortcoming. Moreover, because they have been based on relatively small corpora, previous studies have not been able to ensure that the phraseology they identify is genuinely universal across academic disciplines. The possibility of constructing a listing of positionally-flexible collocations which are generically academic remains, as far as I am aware, unexplored.

6.3 Creating an academic corpus

Introduction

As I was unable to locate any publicly-available corpus of academic English suitable to use as a basis for identifying academic collocation, it was necessary to create my own. The present section details its design and compilation.

Design of the corpus

Two key criteria determined the design of the corpus. The first was size: identifying collocations is thought to require relatively large corpora (Halliday, 1966). At the same time, I was faced with the practical limitations that the corpus was to be compiled by a single researcher with limited resources and within the relatively narrow time-scale permitted by this thesis. In short, the corpus needed to be as large as possible, given limited resources. The second criterion was range. Academic collocations are those which are common across academic disciplines. It was important, therefore, to gather data from a wide variety of subject areas.

Since a large corpus was required, it was decided to limit the investigation to academic writing, the compilation of large spoken corpora being too labour-intensive an undertaking for the current project. Practical considerations aside, an exclusive focus on written collocations is also justified in terms of the wider aims of the research, since the majority of EAP pedagogy continues to focus on written language. Within academic writing, I decided to concentrate exclusively on research articles published in scholarly journals. Other types of academic writing, such as student essays and textbooks were not included. This decision was again driven in part by the need to create a large corpus with limited resources. Collecting large samples of student writing is a problematic and costly business. Even well-funded large-scale projects (such as the corpus of British Academic Written English <http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/>) have been slow in development and have yielded relatively small corpora. Similarly, though textbooks are more accessible, converting them to electronic form is both time-consuming and error-prone. Again, even in large-scale projects (e.g., Biber, 2006), collections from academic textbooks tend to be relatively small in comparison to the multi-million word corpus which we are aiming to compile. Research articles, in contrast, are far

easier to work with, with a huge range of sources being freely available in electronic form to institutional subscribers. Compiling a large and widely-sampled corpus of such articles is therefore a relatively straightforward task.

This exclusive focus on research articles will, of course, limit the representativeness of the corpus as a sample of academic writing in general. Biber (1993) notes the importance of sampling texts from the fullest possible range of registers and text types found in a population in order to capture that population's full range of linguistic variability. Since almost nothing is known about the use of collocations across different academic text types, we have no way of knowing how much variation may be lost by our focus on journal articles. However, I would argue that an entirely article-based corpus may in fact be better suited to our current purposes than one representative of academic writing as a whole. As research articles are (for most disciplines) the most prestigious form of academic writing, and as they are more analogous in their aims and structure to student writing than are other forms of professional academic prose (e.g. textbooks), they would seem to provide the best available model of 'target language' for students of EAP (Hyland, 2008). A corpus based on research articles may therefore be more representative of the language students should be aiming to acquire than a more broadly-based sample would be.

As we have seen, it was important for the aims of the project to capture the language of a wide range of academic disciplines. As a first step towards this, I took the departmental structure of the University of Nottingham to represent an approximate 'map' of the spectrum of academic study. The university is divided into five faculties:

- Arts and Humanities (referred to below as 'Arts');
- Engineering ('Eng');
- Medicine and Health Sciences ('Med');
- Science ('Sci');
- Social Sciences, Law and Education ('SS').

These were taken as a first level of division within the corpus. A target of 25 million words for the whole corpus was divided equally between these faculties. That is, I

aimed to collect five million words for each faculty. The university faculties are further divided into a number of Schools, and most Schools are divided into a number of Divisions, representing particular research foci. Thus, for example, the Faculty of Arts and Humanities hosts the School of English Studies, within which are the four Divisions of Medieval Studies, Modern English Language, Modern English Literature, and Drama. To achieve a spread of texts from across disciplines, I divided the target five million words of each Faculty equally between however many Schools were in the Faculty, and further divided the words assigned to each School equally between its Divisions.

It is important to note that this division may not be an entirely accurate reflection of the full linguistic variation in our target population. One issue is that there will be certain subjects not taught at the University. Another is that the division of subjects between university departments often reflects administrative priorities and accidents of history as much as principled academic divisions (and still less principled linguistic divisions). A corpus based on another university's administrative structure might give more prominence to law, business, or political science for example (which in our division must share five million words with education, and other social sciences), or de-emphasise engineering (which here takes up a full fifth of the corpus, whereas on another university's structure it might be subsumed within science). However, since little is known about how collocations vary across disciplines, we cannot be sure, in advance of any analysis, what a 'comprehensive range' should consist of or how it should be structured. My approach will therefore follow the 'cyclical' process recommended by Biber (1993), in which initial sampling specifications are open to later correction once some preliminary analysis has been undertaken (see 'Restructuring the corpus', below).

Compiling the corpus

For each Division in the corpus 'map', five prominent journal were identified, and the number of words to be collected for each discipline divided equally between them. Prominent journals were identified in one of two ways. The first was through the *ISI Web of Knowledge* database <http://portal.isiknowledge.com/portal.cgi>. This provides listings of journals under disciplinary headings and enables them to be ranked according to the number of times they have been cited in the previous year.

The texts included for each Division in the corpus came from the five most frequently-cited publications to which I was able to gain electronic access. In most cases, the disciplinary headings within ISI corresponded to Divisions within the Nottingham University structure. Exceptions to this occurred where disciplines went under slightly different names (e.g. Nottingham's School of 'Pharmacy' vs. ISI's classification of 'Pharmacology and Pharmacy') or where ISI made distinctions between subdisciplines which were not reflected in the university structure. In such cases, the ISI classification was followed.

Since the ISI database includes only information on science and social science journals, texts for a minority of disciplines (primarily those from the Faculty of Arts and Humanities) could not be identified in this way. For these cases, I referred instead to *Ulrich's Periodicals Directory* <http://www.ulrichsweb.com/ulrichsweb/>. This enabled me to find current journals with available electronic editions for each of the remaining subject areas. Citation information was not available on this database, so journals were selected, firstly, on the basis of the length of time they had been in publication, (with preference given to the most long-established journals) and, secondly, such as to achieve a spread between apparent sub-areas of a discipline (e.g. for American and Canadian studies, I attempted to achieve a spread between journals focusing on literature and journals focusing on history). In the case of German studies and Russian and Slavonic studies, I was not able to identify five electronically-available English language journals. These schools are therefore represented by three journals only. There was some overlap between areas, with certain journals appearing in the databases under more than one subject heading. In such cases, the journal was used only for the first Division compiled.

In each case, the most recent available issue of the journal was used. In selecting contents from a journal, I aimed to give priority to articles describing original research. Texts such as editorials and letters were not included. In some disciplines (e.g. Law), book reviews and extended essays are a prominent part of journals' content. These were therefore also included for such subject areas. Only complete texts were used. Academic language was presumed not to have any native speakers and to exist somewhat independently of national linguistic varieties. No attempt was therefore

made to distinguish between writers from different L1 backgrounds or between journals using British, US, or other forms of English.

Restructuring the corpus

Previous research on academic word lists (e.g., Coxhead, 2000; Ellis et al., 2007; Hyland & Tse, 2007) has usually divided a main academic corpus into several sub-corpora corresponding to different subject areas. The point of this has been to ensure that vocabulary items are found across a variety of disparate disciplines. Thus, Coxhead (2000) requires that words occur at least ten times in each of four sub-corpora (arts, commerce, law, science) and in at least 15 out of 28 different subject areas in order to count as academic vocabulary. In order to capture the full range of variation in an academic corpus, it is important that the divisions used should correspond to natural groupings of subjects in terms of their vocabulary use. That is to say, subject areas which are included within a single sub-corpus should be as similar to each other as possible in the words that they use, while subjects which have less in common should be kept apart. If this criterion is not met, we risk convincing ourselves that some words are more universal than they really are.

To take an example, the Social Sciences, Law and Education faculty of my corpus includes the School of Built Environment. In terms of vocabulary use, however, this subject area has more in common with subjects found in the faculty of Engineering than it does with other schools found within its own faculty (this claim will be substantiated below). By including this school in its faculty grouping, we are therefore likely to create the false impression that certain words which are common in engineering are also common in the social sciences. It is a weakness of previous research into academic vocabulary that this problem has not been seriously addressed. While Coxhead notes the issue, she gives an explicit justification for only one of her categorisations (the inclusion of Psychology and Sociology within the Arts sections), and this choice is based on syntactic, rather than lexical research (2000, p. 220).

Before proceeding to the main analysis then, we need to see whether we can identify more natural lexical groupings of subjects than those of the present faculty-based divisions. To determine this, I will assume that the 31 schools represented in the corpus constitute basic subject units, and ask how these units are best grouped in

terms of their vocabulary use. As a first step to answering this, we need to identify what vocabulary is important in each school. This will be determined using Scott's concept of *keywords*.

Scott (1999), defines keywords as words "whose frequency is unusually high in comparison with some norm". Operationalised, this means that the keywords in a given corpus are those words which appear significantly more frequently than they do in some reference corpus which is taken to represent a relevant 'norm'. A 'significant' (as usual, the word must be treated with some caution, given the unusual nature of linguistic data as samples in inferential analysis) difference in frequencies of occurrence between two corpora is usually identified using the categorical chi-squared or log-likelihood tests. Of the two, log-likelihood is thought to give the better indication of keyness (Dunning, 1993), and this statistic will be employed here. Log-likelihood is calculated by constructing a 2x2 contingency table based on a word's frequencies of occurrence in two corpora. Thus, for example, the word *analysis* appears in our (approximately) 25 million word academic corpus 21,215 times and in the (approximately) 100 million word BNC 13,297 times. We can therefore construct the following contingency table of observed values:

	academic	**BNC**	**total**
words = *analysis*	21,215	13,297	34,512
words ≠ *analysis*	24,978,785	99,986,703	124,965,488
total	25,000,000	100,000,000	125,000,000

These figures can be used to determine the expected occurrences of the word in each corpus on the null hypothesis that there is no difference between the two corpora with the equation:

$$E = \frac{C * W}{T}$$

where C is the number of words in the corpus (i.e. 25m or 100m), W is the total number of appearances of the word across the two corpora (i.e. 34,512) and T is the total number of words in the two corpora combined (i.e. 125m). Using this equation, a contingency table of expected values can be constructed as follows.

	academic	**BNC**	**total**
words = *analysis*	(25m*34,512/125m) = 6902.4	(100m*34,512/125m) = 27,609.6	34,512
words ≠ *analysis*	(25m-6902.4) = 24,993,097.6	(100m – 7419.54) = 99,972,390.4	124,965,488
total	25,000,000	100,000,000	125,000,000

The log-likelihood statistic then compares the observed frequencies with the expected frequencies using the same equation we met in our discussion of association measures in Section 4.3, i.e.:

$$\text{log-likelihood} = 2\sum_{ij} O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

Thus, for the current example, the log-likelihood score is:

$$2*\left[\left(21{,}215*\ln\left(\frac{21{,}215}{6{,}902.4}\right)\right)+\left(13{,}297*\ln\left(\frac{13{,}297}{27{,}609.6}\right)\right)+\left(24{,}978{,}785*\ln\left(\frac{24{,}978{,}785}{24{,}993{,}097.6}\right)\right)+\left(99{,}986{,}703*\ln\left(\frac{99{,}986{,}703}{99{,}972{,}390.4}\right)\right)\right]$$

= 28,222.07

Because of the unusual nature of linguistic data, and because of the risk of inflated family-wise error rates which comes with simultaneously calculating large numbers of significance values, Scott recommends using the conservative significance value of $p < .1x10^{-7}$ in determining listings of keywords (Scott, 2007). This corresponds to a log-likelihood value of 28.33. Since the value found here for *analysis* is well in excess of this, we can conclude this word is 'key' in academic writing, in comparison to the 'norm' represented by the BNC.

To identify important words in the various academic disciplines in our corpus, separate listings of keywords were generated for each of the 31 schools in the corpus using the *Keywords* facility of *WordSmith Tools*, taking the BNC as a reference corpus

and using the program's default value of $p < .1x10^{-7}$ as a criterion for significance. To avoid attributing false importance to low frequency words simply because they fail to occur at all in the BNC, I also required that keywords should appear in the school with a mean frequency of at least 20 occurrences per million words. Another potential problem was that words could be frequent in a school as a whole simply because they were repeated frequently within a particular article (or small group of articles), even if it is absent from the more general discourse of the discipline. To guard against this, I also required that keywords appear in at least 20% of texts in a school. Finally, piloting found that keywords of three letters or fewer were almost exclusively abbreviations or acronyms. For this reason, only words of 4 letters or more were included. In sum, the keywords for a school are those which:

- contain four or more alphabetical characters;
- appear in the school with a mean frequency of at least 20 per million words;
- appear in at least 20% of texts in the school;
- appear in the school significantly more frequently than in the BNC, with the threshold for significance set at $p < .1x10^{-7}$.

This method yielded 31 distinct keyword lists, ranging in length from 466 words (the school of Mathematical Sciences) to 1,111 words (the school of Law). The next step in determining vocabulary groupings was to see how much overlap existed between the keywords of each school. Schools with extensive overlaps between their keyword lists should be grouped together in our later analyses, while schools which have little in common should be kept apart. To this end, the average percentage overlap between the keyword lists of each school was determined (i.e. the percentage of the total unique words found in two lists which are common to both). Table 1 shows the percentage overlaps between all 31 schools. To aid interpretation, overlaps which are more than one standard deviation below the mean (i.e. < 9.16%) are shaded in red; overlaps which are more than one standard deviation above the mean (i.e. > 31.25%) are shaded in blue. Also to aid interpretation, all overlaps are shown in both columns and rows (so that the bottom left half of the matrix mirrors the upper right half), and the sets of cells corresponding to faculties are outlined in bold.

A number of points stand out from this matrix. First, there is a good deal of overlap within the faculty groupings. A large number of blue cells, and no red cells, are found within the faculty boxes, indicating that the overlap within these groups is substantially above the average. This suggests that the faculty-based groupings do correspond reasonably well to vocabulary use. However, the substantial amount of blue appearing outside of these boxes also indicates that that many strong overlaps extend beyond faculty boundaries, undermining the groupings somewhat. Finally, it is clear that subjects in the Arts and Humanities faculty stand somewhat apart from the rest of the corpus. While this faculty is clearly quite coherent internally (all overlaps within the faculty are fairly high), it seems to have little in common with other faculties: in fact, all of the very low (red-shaded) overlaps on the matrix except one involve a school from this faculty, while no school in this faculty forms any strong (blue) tie with any school outside the faculty.

Table 1: overlaps in key vocabulary between schools in the academic corpus

	AmCan	A:Eng	A:Hist	A:Hums	A:Modlang	E:Chem	E:Civ	E:Elec	E:Mech	M:Bio	M:Comm	M:Humdev
AmCan	**100.00**	38.30	30.77	30.22	39.34	6.93	7.77	7.98	8.23	6.99	7.16	5.56
A:Eng	38.30	**100.00**	27.17	39.37	46.44	10.24	10.84	10.00	11.51	10.13	9.83	8.16
A:Hist	30.77	27.17	**100.00**	21.87	30.99	5.54	5.89	4.81	5.32	6.66	7.53	6.10
A:Hums	30.22	39.37	21.87	**100.00**	40.64	9.69	10.51	10.40	11.72	9.11	8.59	7.21
A:Modlang	39.34	46.44	30.99	40.64	**100.00**	12.43	13.07	12.53	14.17	11.96	11.89	10.86
E:chem	6.93	10.24	5.54	9.69	12.43	**100.00**	36.12	41.43	46.02	32.13	26.54	25.97
E:Civ	7.77	10.84	5.89	10.51	13.07	36.12	**100.00**	41.03	52.26	23.79	23.25	21.30
E:Elec	7.98	10.00	4.81	10.40	12.53	41.43	41.03	**100.00**	51.70	22.75	20.21	20.10
E:Mech	8.23	11.51	5.32	11.72	14.17	46.02	52.26	51.70	**100.00**	32.10	28.24	28.71
M:Bio	6.99	10.13	6.66	9.11	11.96	32.13	23.79	22.75	32.10	**100.00**	42.77	41.20
M:Comm	7.16	9.83	7.53	8.59	11.89	26.54	23.25	20.21	28.24	42.77	**100.00**	45.79
M:Humdev	5.56	8.16	6.10	7.21	10.86	25.97	21.30	20.10	28.71	41.20	45.79	**100.00**
M: MedSurg	6.33	8.99	6.58	8.61	12.72	27.72	20.76	21.51	29.09	44.51	46.84	52.17
M:MolMed	7.33	9.49	6.76	8.14	10.91	24.21	18.70	18.48	24.77	49.84	38.45	38.65
M:Nursg	10.95	13.67	9.55	10.26	16.19	20.70	19.27	16.44	22.73	29.43	44.05	34.82
M: Vets	5.23	8.15	6.04	6.77	10.30	26.75	19.88	19.38	26.59	41.87	32.54	39.10
S:Biols	7.34	10.33	6.86	9.58	12.23	30.96	26.46	23.46	31.30	51.80	29.48	29.12
S: BioSci	5.85	8.28	5.34	8.20	10.62	44.98	27.40	27.41	35.15	46.11	32.19	31.67
S:Chem	4.75	6.31	3.41	6.31	7.54	44.30	24.61	30.28	31.64	28.27	19.12	19.68
S:CompSci	10.63	13.88	6.99	13.30	15.75	27.00	42.90	35.19	40.44	22.02	20.47	17.79
S:Maths	5.52	8.29	4.30	9.97	11.32	19.91	29.43	25.79	26.17	14.29	12.65	12.58
S:Pharma	5.15	8.17	4.76	6.89	9.11	27.29	18.84	19.21	25.41	50.21	29.29	30.19
S:Physics	6.84	9.84	4.71	10.43	12.86	40.24	41.88	45.87	44.78	24.17	20.12	18.87
S:Psych	12.53	17.27	10.36	13.84	18.17	22.62	24.45	19.26	25.80	28.10	36.84	25.35
SS:BuiltEnv	17.28	18.99	11.30	18.66	24.09	28.12	39.02	29.39	39.22	20.53	20.30	19.06
SS:Bus	10.98	12.73	9.07	10.54	13.60	20.52	25.14	19.12	23.89	18.59	21.94	17.20
SS:Econ	7.85	10.68	6.49	9.54	10.47	19.08	28.49	20.16	23.03	17.08	18.81	15.26
SS: Edu	13.07	15.39	11.98	12.30	18.17	18.97	19.22	15.82	20.81	21.56	30.38	23.53
SS: Law	14.77	14.42	13.57	13.25	15.67	11.73	14.46	11.08	13.31	12.45	14.42	11.47
SS: Pols	16.12	17.53	15.87	15.26	21.52	17.34	21.34	16.01	20.15	17.15	18.69	15.53
SS: Socs	19.19	19.04	18.02	14.50	23.62	16.15	18.29	13.72	18.55	19.43	26.67	19.37

Table 1 (contd.): overlaps in key vocabulary between schools in the academic corpus

	M: MedSurg	M:MolMed	M:Nursg	M: Vets	S:Biols	S: BioSci	S:Chem	S:CompSci	S:Maths	S:Pharma	S:Physics	S:Psych
AmCan	6.33	7.33	10.95	5.23	7.34	5.85	4.75	10.63	5.52	5.15	6.84	12.53
A:Eng	8.99	9.49	13.67	8.15	10.33	8.28	6.31	13.88	8.29	8.17	9.84	17.27
A:Hist	6.58	6.76	9.55	6.04	6.86	5.34	3.41	6.99	4.30	4.76	4.71	10.36
A:Hums	8.61	8.14	10.26	6.77	9.58	8.20	6.31	13.30	9.97	6.89	10.43	13.84
A:Modlang	12.72	10.91	16.19	10.30	12.23	10.62	7.54	15.75	11.32	9.11	12.86	18.17
E:chem	27.72	24.21	20.70	26.75	30.96	44.98	44.30	27.00	19.91	27.29	40.24	22.62
E:Civ	20.76	18.70	19.27	19.88	26.46	27.40	24.61	42.90	29.43	18.84	41.88	24.45
E:Elec	21.51	18.48	16.44	19.38	23.46	27.41	30.28	35.19	25.79	19.21	45.87	19.26
E:Mech	29.09	24.77	22.73	26.59	31.30	35.15	31.64	40.44	26.17	25.41	44.78	25.80
M:Bio	44.51	49.84	29.43	41.87	51.80	46.11	28.27	22.02	14.29	50.21	24.17	28.10
M:Comm	46.84	38.45	44.05	32.54	29.48	32.19	19.12	20.47	12.65	29.29	20.12	36.84
M:Humdev	52.17	38.65	34.82	39.10	29.12	31.67	19.68	17.79	12.58	30.19	18.87	25.35
M: MedSurg	**100.00**	43.53	34.78	40.91	31.07	31.93	21.21	18.50	11.87	34.07	20.87	27.02
M:MolMed	43.53	**100.00**	26.79	49.28	41.11	34.24	20.98	17.26	11.29	43.97	18.03	25.49
M:Nursg	34.78	26.79	**100.00**	23.95	20.99	21.79	13.34	18.56	10.53	19.41	14.36	40.68
M: Vets	40.91	49.28	23.95	**100.00**	36.60	38.84	22.38	16.01	11.62	39.48	18.88	22.14
S:Biols	31.07	41.11	20.99	36.60	**100.00**	42.09	29.21	24.76	16.46	46.71	25.41	26.04
S: BioSci	31.93	34.24	21.79	38.84	42.09	**100.00**	37.25	22.92	15.89	38.71	27.81	24.08
S:Chem	21.21	20.98	13.34	22.38	29.21	37.25	**100.00**	19.55	15.24	26.71	35.43	15.66
S:CompSci	18.50	17.26	18.56	16.01	24.76	22.92	19.55	**100.00**	36.71	16.45	32.53	26.21
S:Maths	11.87	11.29	10.53	11.62	16.46	15.89	15.24	36.71	**100.00**	11.73	29.10	14.25
S:Pharma	34.07	43.97	19.41	39.48	46.71	38.71	26.71	16.45	11.73	**100.00**	19.78	21.86
S:Physics	20.87	18.03	14.36	18.88	25.41	27.81	35.43	32.53	29.10	19.78	**100.00**	18.86
S:Psych	27.02	25.49	40.68	22.14	26.04	24.08	15.66	26.21	14.25	21.86	18.86	**100.00**
SS:BuiltEnv	19.95	17.40	23.48	17.71	21.90	21.42	17.37	33.73	18.37	15.44	25.02	25.52
SS:Bus	16.62	16.94	25.32	13.79	18.94	18.27	13.43	25.50	15.77	14.68	18.59	29.99
SS:Econ	14.87	14.56	16.97	12.31	18.30	17.21	14.20	29.33	25.61	13.50	21.41	21.68
SS: Edu	22.77	20.58	45.85	17.94	18.84	18.67	11.81	20.94	10.69	15.47	13.89	40.11
SS: Law	11.16	11.06	16.63	9.19	12.12	10.75	7.18	14.94	10.05	9.24	11.33	18.14
SS: Pols	14.90	15.22	22.83	13.12	17.80	16.35	11.62	21.79	13.03	12.38	16.08	25.71
SS: Socs	18.48	18.98	38.52	15.59	17.61	16.49	10.25	19.23	10.30	14.04	13.60	37.65

Table 1 (contd.): overlaps in key vocabulary between schools in the academic corpus

	SS:BuiltEnv	SS:Bus	SS:Econ	SS: Edu	SS: Law	SS: Pols	SS: Socs
A: AmCan	17.28	10.98	7.85	13.07	14.77	16.12	19.19
A:Eng	18.99	12.73	10.68	15.39	14.42	17.53	19.04
A:Hist	11.30	9.07	6.49	11.98	13.57	15.87	18.02
A:Hums	18.66	10.54	9.54	12.30	13.25	15.26	14.50
A:Modlang	24.09	13.60	10.47	18.17	15.67	21.52	23.62
E:Chem	28.12	20.52	19.08	18.97	11.73	17.34	16.15
E:Civ	39.02	25.14	28.49	19.22	14.46	21.34	18.29
E:Elec	29.39	19.12	20.16	15.82	11.08	16.01	13.72
E:Mech	39.22	23.89	23.03	20.81	13.31	20.15	18.55
M:Bio	20.53	18.59	17.08	21.56	12.45	17.15	19.43
M:Comm	20.30	21.94	18.81	30.38	14.42	18.69	26.67
M:Humdev	19.06	17.20	15.26	23.53	11.47	15.53	19.37
M: MedSurg	19.95	16.62	14.87	22.77	11.16	14.90	18.48
M:MolMed	17.40	16.94	14.56	20.58	11.06	15.22	18.98
M:Nursg	23.48	25.32	16.97	45.85	16.63	22.83	38.52
M: Vets	17.71	13.79	12.31	17.94	9.19	13.12	15.59
S:Biols	21.90	18.94	18.30	18.84	12.12	17.80	17.61
S: BioSci	21.42	18.27	17.21	18.67	10.75	16.35	16.49
S:Chem	17.37	13.43	14.20	11.81	7.18	11.62	10.25
S:CompSci	33.73	25.50	29.33	20.94	14.94	21.79	19.23
S:Maths	18.37	15.77	25.61	10.69	10.05	13.03	10.30
S:Pharma	15.44	14.68	13.50	15.47	9.24	12.38	14.04
S:Physics	25.02	18.59	21.41	13.89	11.33	16.08	13.60
S:Psych	25.52	29.99	21.68	40.11	18.14	25.71	37.65
SS:BuiltEnv	**100.00**	24.13	19.10	27.36	16.54	22.14	25.33
SS:Bus	24.13	**100.00**	40.97	27.34	27.94	38.00	30.69
SS:Econ	19.10	40.97	**100.00**	17.77	24.38	30.15	19.20
SS: Edu	27.36	27.34	17.77	**100.00**	18.49	26.24	45.03
SS: Law	16.54	27.94	24.38	18.49	**100.00**	32.75	23.93
SS: Pols	22.14	38.00	30.15	26.24	32.75	**100.00**	33.22
SS: Socs	25.33	30.69	19.20	45.03	23.93	33.22	**100.00**

Key to abbreviations in Table 1

A: AmCan	Arts: American & Canadian studies
A:Eng	Arts: English studies
A:Hist	Arts: History
A:Hums	Arts: Humanities
A:Modlang	Arts: Modern languages & cultures
E:Chem	Eng: Chemical, environmental, & mining engineering
E:Civ	Eng: Civil engineering
E:Elec	Eng: Electrical & electronic engineering
E:Mech	Eng: Mechanical, materials & manufacturing engineering
M:Bio	Med: Biomedical sciences
M:Comm	Med: Community health sciences
M:Humdev	Med: Human development
M: MedSurg	Med: Medical and surgical sciences
M:MolMed	Med: Molecular medical science
M:Nursg	Med: Nursing
M: Vets	Med: Veterinary medicine & science
S:Biols	Sci: Biology
S: BioSci	Sci: Biosciences
S:Chem	Sci: Chemistry
S:CompSci	Sci: Computer science and information technology
S:Maths	Sci: Mathematical sciences
S:Pharma	Sci: Pharmacy
S:Physics	Sci: Physics & astronomy
S:Psych	Sci: Psychology
SS:BuiltEnv	SS: Built environment
SS:Bus	SS: Business
SS:Econ	SS: Economics
SS: Edu	SS: Education
SS: Law	SS: Law
SS: Pols	SS: Politics and international relations
SS: Socs	SS: Sociology & social policy

To get a more quantifiable view of how coherent the faculty groupings are, I will use two different measures. First, I will look at the mean and minimum overlap between the keyword lists of each school within each faculty. Scores on this measure for the original faculty-based subject groupings are shown in Table 2.

Table 2: keyword overlaps within faculty groups

	mean overlap %	**minimum overlap %**
Arts & Humanities ('Arts')	34.51	21.87
Engineering ('Eng')	44.76	36.12
Medical and Health Sciences ('Med')	40.06	23.95
Science ('Sci')	25.62	11.73
Social science, law and education ('SS')	27.18	16.54
Overall	**34.43**	**11.73**

Some faculties appear to be very homogenous; Eng and Med in particular achieve very high overlaps. However, as our above review of the overlap matrix suggested, overlaps within the Sci and SS faculties are rather lower, and the minimum overlap within the Sci faculty (the 11.73% overlap between Mathematical Sciences and Pharmacy) is particularly low.

The second means of assessing the coherence of groups is intended to produce a more directly pedagogically meaningful measure. It aims to determine how well served students from each school within a group would be by a vocabulary list based on that group as a whole. The method here is first to create a list of the top keywords (as defined above) for each school. These listings are taken to show the words which students in each school most need to know. Since the shortest keyword list found was 466 words long, we will consider only the top 466 words of each list. Second, a similar list of 466 keywords is generated for each group (in this case, each faculty) as a whole. The overlap between this list and the listing of each school within the group shows what percentage of the words students most need they would get from a generic group-based list. The more coherent the group, the better this coverage would be. Table 3 shows the mean and minimum 'top keyword coverage' achieved within each faculty group.

Table 3: coverage by top 466 keywords

	mean coverage %	minimum coverage %
Arts & Humanities (‘Arts’)	61.07	46.57
Engineering (‘Eng’)	71.24	63.73
Medical and Health Sciences (‘Med’)	65.63	50.86
Science (‘Sci’)	50.83	40.34
Social science, law and education (‘SS’)	49.72	38.84
Overall	**59.36**	**38.84**

Again, the Eng and Med faculties do reasonably well. However, it appears that many students in the Sci and SS faculties would not be well served by vocabulary listings based on these groupings. On average, students within these faculties would only get about half of the words they most need, with students in the school of Law getting little more than a third (38.84%).

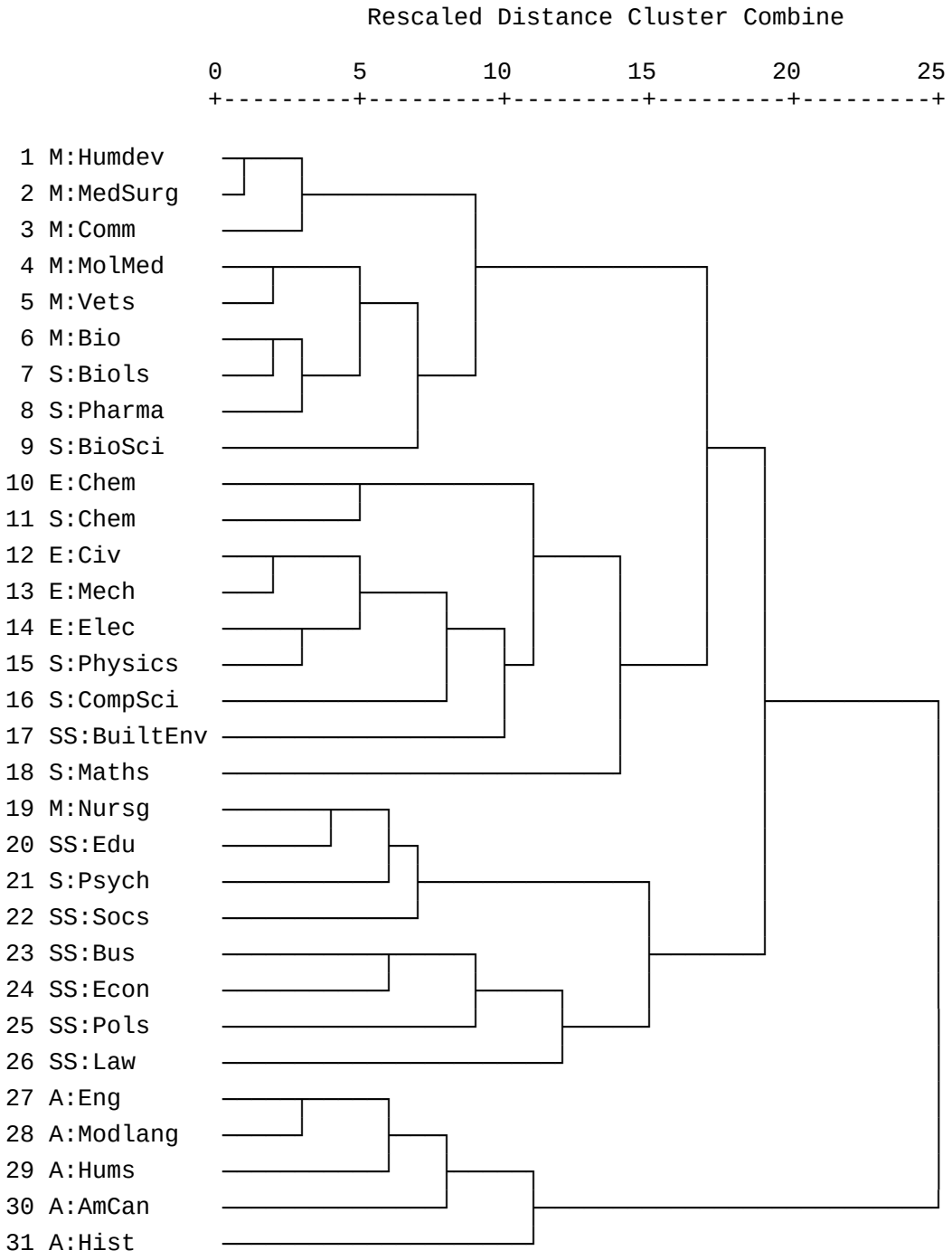
Using these measures of coherence, we can now ask whether the faculty groupings can be improved upon. One method which corpus linguists have used to group texts together in terms of quantitatively-defined similarities is that of hierarchical cluster analysis (McEnery & Wilson, 2001, pp. 92-95). This is a technique for grouping objects into classes of similar objects on the basis of their scores on a number of different variables. The technique starts by determining the degree of similarity of each object with each other object in terms of all the variables measured, a similarity which is expressed as a multi-dimensional Euclidean distance. Once a matrix of Euclidean distances between all objects has been produced, objects are then grouped according to one of a number of techniques. These techniques differ, firstly, in their starting point: *agglomeration* techniques start from individual objects then gradually combine these to form larger groups; the less commonly-used *division* techniques start with a single large group which is divided into smaller groups until individual objects are reached. Within the agglomeration technique, there are also a number of different means of determining how objects should be grouped. For example, on the *nearest neighbour linkage* method, groups of objects are merged at a given point in the analysis if at least one object in the group is sufficiently close to at least one object in the other group; on the *furthest neighbour linkage* technique, two groups only merge if the most distant objects from each group are sufficiently close; on the *group average linkage* method, groups merge only if the average distance between groups is small enough. Different techniques may produce different results for the same data, and

there is no generally accepted best method. The results of cluster analysis are typically represented as a branching 'dendogram' (see figure 1) showing the clusters found at each level of the analysis (Manly, 2005, pp. 125-130).

Our data set is somewhat different from that typically used in cluster analysis in that we are not trying to determine groups on the basis of their scores a number of different independent variables, but rather on the basis of their percentage keyword overlaps with a number of different objects. The basic concept remains the same, however: if two schools attain similar overlaps across the range of other schools, we will want to group them together. To take a concrete example, the school of Biomedical Sciences (in the Med faculty) and the school of Biology (in the Sci faculty) appear to mirror each other very closely in the degree to which they overlap with all other schools (see Table 1). We would therefore want to say that these schools are similar. To generate a cluster analysis based on this idea, we can treat each column in the matrix as a variable, representing the degree to which each object (represented by the rows) uses vocabulary similar to that found in a certain school. Thus, the first column represents a variable describing how similar each school's vocabulary usage is to that found in the school of American and Canadian Studies. The school of American and Canadian studies itself, of course, scores very highly on this variable (100), whereas schools in the Eng, Med, and Sci faculties score quite low.

While this is a rather unorthodox way of carrying out cluster analysis, we will be able to judge its validity in the pragmatic terms of whether it produces a more coherent set of vocabulary groupings than that provided by the faculties. Since different clustering methods can produce different results, a number of different types offered by *SPSS 12* were piloted. The methods were also tried both with and without first standardizing the overlap values (standardization was achieved by dividing all 'variables' by their standard deviation). There were, in fact, only minor disagreements between methods in the groupings yielded, and all groupings constituted some improvement over the faculty-based groups. The best results and the most easily readable dendogram were found for an analysis based on average between groups linkage without standardization. The dendogram produced by this method is shown in Figure 1 (rows are numbered for ease of reference; these numbers do not refer to any part of the analysis; see the key to Table 1 for clarification of abbreviated school names).

Figure 1: Dendrogram using Average Linkage (Between Groups)



The dendogram reads from left to right and represents the clusters of schools formed at each level of Euclidean distance. For example, the schools of Human Development and Medical and Surgical Sciences (rows 1 and 2) combine at the first level of analysis. These are then joined by Community Health Sciences. At the next level, this group of three combines with a larger group (which contains the schools of Molecular and Medical Science, Veterinary Medicine and Science, Biomedical Science, Biology,

Pharmacy, and Biosciences) to form a group of nine. At the next level, this group in turn combines with another group of nine (containing the Eng schools plus Chemistry, Physics and Astronomy, Computer Science and Mathematical Science from the Sci faculty and the school of Built Environment from SS). This group of 18 schools then combines with another group of eight, comprising all of the SS schools (except Built Environment) plus Nursing and Psychology. Finally, this large group combines with a group of four schools corresponding to the Arts and Humanities faculty.

An intuitive reading of the dendogram appears to suggest five major groupings of schools:

- *rows 1-9:* Human Development; Medical and Surgical Sciences; Community Health Sciences; Molecular and Medical Science; Veterinary Medicine and Science; Biomedical Sciences; Biology; Pharmacy; Biosciences;
- *rows 10-18*: Mechanical, Materials and Manufacturing Engineering; Chemistry; Chemical, Environmental and Mining Engineering; Electrical and Electronic Engineering; Civic Engineering; Physics and Astronomy; Computer Science and Information Technology; Built Environment; Mathematical Sciences;
- *rows 19-22*: Education; Sociology and Social Policy; Nursing; Psychology;
- *rows 23-26*: Business; Economics; Politics and International Relations; Law;
- *rows 27-31*: English Studies; Modern Languages and Cultures; Humanities; American and Canadian Studies; History.

These appear to form quite intuitively satisfying groupings, which I shall label, respectively:

- Life Sciences
- Science and Engineering
- Social-Psychological
- Social-Administrative
- Arts and Humanities

We can confirm that these groupings provide some improvement over the original faculty-based groups by considering the overlaps and the top 466 keyword coverage of each group. These are shown in tables 4 and 5.

Table 4: keyword overlaps within 'between group clusters'

	mean overlap %	**minimum overlap %**
Arts & Humanities	34.51	21.87
Life Sciences	39.90	29.12
Science and engineering	33.83	15.24
Social-Administrative	32.36	24.38
Social-Psychological	41.31	37.65
Overall	**36.38**	**15.24**

Table 5: coverage by top 466 keywords

	mean coverage %	**minimum coverage %**
Arts & Humanities	61.07	46.57
Life Sciences	64.02	56.22
Science and engineering	60.99	48.07
Social-Administrative	58.26	43.35
Social-Psychological	65.77	62.88
Overall	**62.02**	**43.35**

While the overall increase in mean overlap is modest (from 34.43% for faculty groupings to 36.38% for the new groups), it is important to note that there are now no overlaps as low as those which were previously seen for the rather disparate Sci and SS faculties. The mean overlaps of those faculty groupings were 25.62% and 27.18% respectively. The lowest mean overlap in our new groupings is a rather better 32.36% (in the Social-Administrative group). Similarly, the modest overall increase in percentage keyword 'coverage' by the top 466 keywords for each group (from 59.36% to 62.02%) is less important than the fact that the low coverage achieved in the Sci and SS faculties (of 50.38% and 49.72% respectively) is greatly improved upon; the lowest mean level of coverage now being 58.26% (in Social-Psychological), a marked improvement.

In sum, while the original five faculty-based groupings do appear to have reflected reasonably well the vocabulary use of the various schools in the corpus, the true variation between schools is more accurately captured by the five groups arrived at through cluster analysis. These groupings will therefore constitute a rather better basis for investigating the spread of collocations across the different disciplines within the

corpus and will be used in the following anlayses. The contents of the restructured corpus are summarised in Table 6 and shown in detail in Appendix C.

Table 6: summary of the academic corpus

Group	Total Words	Total Articles
Arts & Humanities	5,049,627	436
Life Sciences	6,156,089	991
Science & Engineering	8,242,417	1,270
Social-Administrative	2,850,789	215
Social-Psychological	2,778,426	339
Total	**25,077,348**	**3,251**

Limitations of the corpus

Before moving on to identifying academic collocations, some limitations of my corpus need to be acknowledged. One weakness stems from the somewhat untidy nature of the texts collected. Few journal articles consist solely of uninterrupted English prose. The most prominent example of such arguably 'extraneous' material is bibliographic information, which constitutes a substantial portion of the word count of the corpus (a manual review of 70 articles – suggested about 13%). Given the diverse ways in which bibliographic information is integrated into articles, however, (e.g. in parentheses, as footnotes, as endnotes), it was not, with the limited resources of this project, possible to excise such material in all cases. For the sake of consistency, it was therefore decided to leave all such information in place.

It might be argued that the formulaic nature of bibliographical information, and its separation from the main flow of prose, should exclude it from a consideration of collocation in academic writing. On the other hand, it might be countered that the 'collocations' which frequently appear in bibliographies (*et al; Journal of;* and *University Press* being the most prominent examples) are indeed part of academic vocabulary. I would argue that the best way forward is a pragmatic one: to remain conscious of the fact that that a substantial proportion of our corpus will be of this nature, and that some of the collocations found are likely to be drawn from material of this sort. Specific judgements about how such material should be used can be left for individual cases.

Aside from bibliographies, other examples of possibly 'extraneous' material include: non-English quotations (common in texts throughout the Arts and Humanities faculty, and also seen in the 'programming language' quoted in computer science texts); figures and tables (especially common in social sciences); and equations (common in economics, engineering and science). Similar considerations apply here as for bibliographic information: such material is often highly integrated into the prose of articles, and so the limited resources of the project did not allow for it to be removed. Moreover, this tight integration into the prose, and the fact that such material arguably constitutes part of the language of the discipline, means that it is debatable whether it should be excluded, even if this were possible. Again then, we will proceed by keeping in mind during analysis that such material is part of our data and is likely to affect the results.

Potentially more problematic is the imperfect nature of the way in which text was captured. Most corpus analysis software requires that corpora be stored as 'plain text' files. Since journals typically publish their articles as 'pdf' or 'html' files, it was necessary to convert the original files into the required format. This involved, for pdf files, using the 'save as text' facility of Adobe Reader and, for html files, copy and pasting the text into a blank Microsoft Notepad file. Unfortunately, both of these procedures proved to be somewhat error prone, introducing some misreading of characters and – the most common problem- some conflation of multiple words into a single string (e.g. *theinevitablevariations*). Moreover, words which are printed hyphenated across two lines appear as two words in the plain text files (e.g. *per-vasive*). Material of this sort is clearly unsatisfactory. In order to discover how serious the problem is likely to be, I examined manually, and with the aid of the Microsoft Word spell-checker, one article from one journal from each division of the corpus (a total of 83 articles, i.e. 2.5% of the 3,357 articles in the corpus). Averaging across all articles, 0.3% of the words in the subcorpus examined were erroneous in this way. 25 articles (30%) did not contain any errors of this sort at all, while in 76 articles (92%), less that 0.5% of the total word count was erroneous. Only in four articles (0.5%) did errors account for more than 1% of the word count. While we need to be aware that this problem exists, I would contend that these figures show that it is unlikely to have a large impact on our results.

6.4 Creating the academic collocation lists

Introduction

This section will describe two ways in which academic collocations could be defined. The first – on direct analogy to the traditional definition of academic vocabulary - is as word pairs which co-occur with at least moderate frequency across a wide range of academic disciplines, but which are not often found in other types of language. The second is as the collocations in which individual academic words are commonly found. The first approach is autonomous from any pre-existing academic word lists. Word pairs are included or excluded solely on the basis of their frequency as a pair, regardless of the frequencies of their component words. This means that words which would not normally be regarded as items of academic vocabulary might, in combination, be academic collocations if their combination created a pattern that was distinctively academic. The second approach, in contrast, would be directly dependent upon a pre-defined academic word list. The thinking behind such a definition would be that, having identified the words which students of EAP need to learn, the next step is to identify what they need to learn about these words in terms of the collocational environment in which they typically appear in the target language.

It is not clear, without empirical investigation, how much overlap there would be between listings of collocations defined in these two ways or whether either approach would produce a pedagogically useful listing. The present section will describe the creation of listings of academic collocations based on each definition. Section 6.5 will determine the degree of overlap between the two and evaluate their likely pedagogical value.

Key collocation approach

Our first definition – which I will call the 'key collocation' approach – characterises academic collocations as word pairs which co-occur with at least moderate frequency across a wide range of academic disciplines, but which are not often found in other types of language. This definition has a number of parts, which need to be unpacked. First, I will follow Jones and Sinclair's (1974) widely used precedent of limiting 'co-occurrence' to occurrences within a four word span (more on the precise specification of this below). Second, we need to state more explicitly what is meant by the

condition that collocations have 'moderate frequency' in academic writing, but not in other forms of language. One possibility here would be to set some more-or-less arbitrary frequency threshold below which collocations are deemed not to be 'frequent' in a corpus. Academic collocations would then be those which come above the threshold in academic texts but below it in non-academic texts. This does not seem satisfactory, however. Not only do we not have any principled means of setting such a threshold, but small, possibly random, differences in frequencies could result in collocations falling above the threshold in one corpus and below it in the other. Instead of this then, I will adopt the log-likelihood-based 'keyness' technique used in the previous section. On this approach, academic collocations will be those pairs which appear significantly more frequently in academic than in non-academic texts. This will be calculated by comparing the overall frequency of collocations in the academic corpus with their frequency in an 85 million word subsection of the BNC, comprising only non-academic texts (These sub-sections were identified using the BNC class codes devised by Lee (2001), and incorporated into the *Text Converter* utility of *WordSmith Tools*). We have seen that Scott (2007) uses a significance threshold of $p < .1x10^{-7}$ as his criterion of 'keyness', a figure which was adopted above. In this case, however, rather than setting any (necessarily arbitrary) level of significance, I will use log-likelihood to produce a ranked list of the collocations which are 'most key' to academic writing. This list can then be used to select, for example, the 'top 1,000' collocates which students of EAP need to learn. To avoid wrongly attributing keyness to unimportant collocations simply because they fail to appear in the BNC, only collocations appearing at least once per million words in the academic corpus will be included in the analysis. To eliminate frequently co-occurring words which do not stand in any interesting collocational relationship, pairs with mutual information scores of less than four will be excluded. This cut-off point both accords with the findings in Section 4.4 that MI scores of greater than 3.7 appeared to be informative about psychological associations and proved, through extensive piloting, to make intuitively appropriate distinctions. Finally, since academic collocations need to be moderately frequent across a range of academic disciplines, only word pairs which meet the last two frequency criteria (minimum frequency of one/million words and minimum mutual information of four) in all five of the sub-corpora described in the previous section (i.e. *arts and humanities; medicine; science*

and engineering; social-administrative; social-psychological) will be considered for inclusion in the final academic collocation list.

The first step in putting this definition into practice was to generate, for each of the five academic sub-corpora, a listing of all word pairs which co-occur within a four word span more than once per million words and with a mutual information score of at least four. This was achieved using the *Word list* function in *WordSmith Tools*. This enables the user to generate a listing of all the collocations of all the word types in a corpus. An important methodological point, which is not acknowledged in the *WordSmith* documentation (but which quickly became clear on consulting concordance lines generated separately through *WordSmith's Concord* tool), is that *Word list* only provides information about collocates occurring to the right-hand side of node words. Thus, for example, in a listing generated for the arts and humanities sub-corpus, the node word *as* lists among its collocates the word *well*, with 2,250 occurrences in 5 million words; the word *well*, meanwhile, lists among its collocates the word *as* with 1,782 occurrences. Consulting a concordance list for *well* shows that these two figures correspond to appearances of *as* to its left- and right-hand sides respectively.

The listing produced then, is one of node words plus collocates which meet the frequency criteria within a span of + 4 words only. This, of course, means that our frequency-based criteria are somewhat stricter than would be the case were a +/- 4 word span used. However, it is unlikely that much useful information will have been lost in this way: the minimum criterion of one occurrence per million words is a necessarily arbitrary one, and adding the specification that the occurrence must be within a + 4 word span simply raises this arbitrary bar somewhat (collocations which meet the threshold only with a minus four word span will, of course, appear listed under the prior word; collocations which meet the criterion in both directions will be listed twice, once under each word). The more important frequency factor is the relative frequencies in the BNC and academic corpora, which are used to generate the 'key collocations' list. Since the right-hand-only rule goes equally for both corpora, it will not have had an important effect here.

A possible criticism of this approach is that it does not merely 'raise the bar' for initial inclusion in the listings, but rather systematically favours collocations which appear in a fixed order. That is, if there were two different word pairs which both co-occurred in a +/- 4 word span with a frequency of 1.5 per million, but one of the pairs showed no preference for order (the collocate appearing 0.75 times/million to the left and 0.75 times/million to the right of the node), whereas the other had a fixed order (the collocate appearing 1.5 times/million to the right of the node), only the latter would be included in our listing. While this bias must be acknowledged, it should be noted that it will apply only to those cases which are in any case at the borderline in meeting the frequency threshold; i.e. those with a frequency of between one and two occurrences per million words in a +/- 4 word span. What the bias in effect means is that, for collocations which are in this borderline area, a 'boost' will be given to word pairs which appear in a relatively fixed order. Since it could be argued that fixed order collocations may be more salient, and so more useful targets for learning, than more variable collocations, this may not be an unfair bias. It should be noted, moreover, that this method will not prevent us from finding more variable collocations, provided that they occur more than twice per million words (in fact, most variable collocations will be detected well below this level; two occurrences per million words is the threshold required for the extreme, and probably very rare, case in which co-occurrence is exactly evenly distributed between the two sides). Once a collocation has been noted as an important academic pairing, further concordance-based work will be undertaken to find the major patterns in which it appears.

Once separate listings of collocations had been generated for each sub-corpus, collocations which were common to all corpora were identified using the functionality of Microsoft Excel. For these shared collocations, an overall frequency figure for the academic corpus as a whole was then calculated by summing the frequencies in each sub-corpus. To determine which collocations are distinctively academic, frequency counts for these collocations were then generated (again using *WordSmith Tools*) for the non-academic sub-sections of the BNC. Microsoft Excel was then used to calculate log-likelihood ratios comparing the frequency of each collocation in the two corpora and to rank collocations according to the size of this ratio, so yielding an ordered listing of the most 'distinctively academic' collocations. Finally, a number of collocations were manually removed from the listing. Collocations were removed if:

- they included an acronym or abbreviation, a proper name, an article, or a number or ordinal other than *one* and *first*;
- the collocation corresponded to a single Latin word (e.g. *ad hoc*, *per cent*);
- the majority of their occurrences appeared to be in writing outside the main text of the articles, e.g. in bibliographies, copyright information, or acknowledgements;
- they appeared on the listing twice (because they are frequent in both 'directions' – see discussion above); the more highly-ranked of the two appearances was kept in each case.

Once this process had been completed, a final list of the most distinctively academic collocations was created comprising the 1,000 highest ranked items. All of these collocations have log-likelihood ratios of greater than 82, indicating that they are far more frequent in academic writing than in everyday English. New frequency counts were then created for each collocation, using the *Concord* feature of *Wordsmith tools*, such that both left and right-hand occurrences of each collocation would be included in the count. These top 1,000 items are not meant, of course, to represent an exhaustive inventory of academic phraseology. They are intended, rather, as a pedagogically-manageable body of learning targets to which learners should pay special attention and – more immediately - as a sample from which we can evaluate the success of this search strategy. These 1,000 collocations are listed in Appendix D, along with their frequencies.

Collocations of academic keywords

My second approach to defining academic collocations – which I will call the 'collocations of academic keywords' approach - defines academic collocations as those pairings in which academic words are typically found in academic writing. This approach requires, as a first step, that we identify a set of academic words. Because, as we saw in Section 4.2, the identification of collocations is usually thought to require an analysis of unlemmatised word forms, existing listings of academic vocabulary, which list broad 'word families', rather than forms, were not considered suitable for

this purpose. It was necessary, therefore, to compile a new listing of academic vocabulary.

This listing was created using the keyword technique described in Section 6.3. That is, keywords were defined as items which:

- contain four or more alphabetical characters;
- appear in the school at least 20 times per million words;
- appear in at least 20% of texts in the school;
- appear in the school significantly more frequently than in the non-academic parts of the BNC (as described above), with the threshold for significance set at $p < .1x10^{-7}$.

Separate analyses were again carried out for each of the five sub-sections of the corpus (i.e. *arts and humanities; life sciences; science and engineering; social-administrative; social-psychological*)' items which were found in be key in all sections were considered academic words. A small number of words which follow-up concordance analysis showed to appear mainly outside of the main body of the text of articles (e.g. *http; authors; journal; references*) were manually excluded. This technique yielded a listing of 112 academic words, shown in Table 7.

Table 7: academic keywords

ABSENCE	DERIVED	LOCATED	REPRESENT
ACCORDING	DETERMINE	LOCATION	REPRESENTS
ACTIVE	DIFFERENCE	MEASURE	RESPECTIVELY
ACTIVITY	DIFFERENCES	MODEL	RESPONSE
ADDITION	DIFFERENT	MODELS	RESULTING
ANALYSES	DIRECT	MOREOVER	ROLE
ANALYSIS	DIRECTLY	MULTIPLE	SHOWN
APPEARS	EFFECTS	NEGATIVE	SHOWS
ASSOCIATED	FACTORS	NOTED	SIGNIFICANT
BASED	FIGURE	OBSERVATION	SIGNIFICANTLY
BETWEEN	FUNCTION	OBSERVATIONS	SIMILAR
BOTH	FURTHERMORE	OBSERVED	SIMILARLY
CAPACITY	GROUPS	OCCUR	SOURCES
CASES	HIGHLY	OCCURS	SPECIFIC
CHARACTERISTICS	HOWEVER	PATTERN	STRONGLY
CHARACTERIZED	IDENTIFIED	PATTERNS	STRUCTURE
COLLECTED	IDENTIFY	POSITIVE	SUBSEQUENT
COMMONLY	IMPACT	POTENTIAL	SUGGESTS
COMPARE	INDICATE	PRESENCE	TERM
COMPARISON	INDICATES	PRESENTED	THEREFORE
CONDITION	INFLUENCE	PRIOR	THESE
CONSISTENT	INITIAL	PROCESSES	THUS
CONTENT	INITIALLY	RELATED	TYPES
CONTRAST	INTERACTION	RELATION	UNIQUE
CRITICAL	INTERNAL	RELATIONSHIP	VALUES
DEFINED	LARGER	RELATIVE	VARIOUS
DEGREE	LIMITED	RELATIVELY	WHEREAS
DEMONSTRATE	LITERATURE	RELEVANT	WITHIN

It should be noted that this listing is not intended to be an exhaustive inventory of academic vocabulary. The conservative $p < .1x10^{-7}$ significance threshold is a conventional, but arbitrary one, and a more liberal cut-off point would have provided a longer listing of items. However, as was noted in the previous section, the purpose of this analysis is not to generate a comprehensive listing of academic items, but rather to identify a manageable sub-set of these items on which learners would do well to concentrate their attention. Our conservative analysis is well-suited to these aims.

The second step in this analysis involved finding the typical collocates of these keywords. The *Concord* function of WordSmith Tools was used to generate listings of the common collocates of each word within each section of the academic corpus. A keyword-collocate combination was considered an interesting collocation if it

occurred with a frequency of at least once per million words (using a +/- 4-word span) and with a MI score of at least 4. Pairs were considered academic collocations only if they met these criteria within all five sections of the corpus. Pairs including articles, or numbers or ordinals other than *one* and *first* were again excluded, as were collocations usually found outside the main part of the text and duplicate cases (where two keywords are collocates of each other). This analysis yielded a listing of 656 academic collocations, shown in Appendix E.

6.5 Evaluating the academic collocation lists

Introduction

I noted at the outset of this chapter that a number of possible obstacles stand in the way of creating an academic collocation list. In particular, I pointed out that it was not clear how academic collocations should be defined and identified, whether a useful and generically academic set of items existed, or how collocations of more than two words should be handled. Section 6.4 has introduced two possible means of defining and identifying academic collocations. We now need to ask whether these have produced useful listings. The first part of this section will describe and compare the contents of the two listings. The second part will describe a study which aims to determine whether increased expertise in academic writing is associated with increased use of these items. Such an association could be taken to indicate that the collocations are indeed useful targets for learning. Section 6.5 will then address the issue of collocations of more than two words.

The contents of the lists

The most immediately obvious difference between the two collocation listings is their lengths. The key collocation approach yielded 1,924 word pairs meeting Scott's $p < 1 \times 10e^{-7}$ threshold; these were cut down to 1,000 collocations in order to provide a manageably-sized listing. The collocations of academic keywords approach, in contrast, returned only 656 collocations. Far longer listings could have been obtained from both methods, of course, by relaxing some of the criteria (e.g. by using lower significance thresholds in identifying either keywords or key collocations). The relative brevity of the keyword-based listing should draw our attention to an important point however: individual academic words often do not have a large number of

generically academic collocates. Of the 112 academic words listed in Table 7, only 44 had five or more collocates which met the minimum frequency requirements in all five sections of the corpus. 16 academic words had one collocate only, and eight (*condition, measure, moreover, observation, similarly, specific, structure, whereas*) had none. This is not to say that these keywords do not have strong collocates in academic writing – the word *structure*, for example, collocates very strongly with the word *crystal* in both science and engineering and life sciences sub-corpora. However, these collocates are not shared across all five disciplinary areas, and so are not generically 'academic'. If, as these data suggest, only a minority of academic words have a large number of academic collocates, it may not be suitable to base collocation listings entirely around listings of academic vocabulary, as this approach has attempted to do.

The second point to note is that the two approaches generate largely distinct sets of collocations. Of the 656 collocations generated on the collocations of academic keywords approach, only 240 are also found amongst the 1,000 collocations generated on the key collocations approach. The remainder of this section will evaluate the contents of the two listings and ask whether one is likely to be more useful to learners than the other.

We can compare the two lists, firstly, in terms of the overall frequency in academic writing of the collocations they identify. In general, word pairs found by the key collocations approach have somewhat higher frequencies of occurrence in the academic corpus as a whole than those found through the keywords method, though the difference is not dramatic. The top 20% most frequent collocations on the key collocations list all have frequencies of greater than 37 per million words, a rate of occurrence equivalent to that of items within the top 3,000 word forms in the BNC. In comparison, the top 20% of collocations on the keyword-based listing have frequencies of 25 per million words or more, which would put them among the 4,000 most frequent BNC word forms. The median frequency across all 1,000 items on the key collocation list is 15 per million words, compared to 14 per million words for the keyword-based list. The least frequent item on the key collocation list (*to-delineate*) occurs 1.76 times per million words, while all of the bottom 30% of items have a frequency of less than 10/million – a rate lower than that of words in the top 7,500

forms in the BNC. On the keyword-based listing, the least frequent item (*main-sources*) occurs 1.24 times per million words, and the bottom 30% of items all have frequencies of lower than 5/million (BNC figures in this paragraph are taken from Leech et al., 2001).

If we take the simple line that more frequent collocations are likely to be better targets for learners, the key collocation approach appears on these data to provide the better listing. The same principle would also suggest that the top few hundred collocations on the list will be of great value to learners, while the relatively low frequency of the bottom few hundred pairs may indicate that these are best reserved for the most advanced learners. These interpretations are open to question, however. Our finding in Section 5.4 that learners of English seem to pick up very high frequency collocations from input should make us ask whether such items really need to be taught, or whether they will simply be acquired naturally. It is not necessarily the case, therefore, that more frequent is always better. With regard to the lower frequency collocations, we can also observe that even relatively infrequent collocations may be worth teaching if they can be included within broader, more common patterns. These issues will be addressed in more detail below.

A second point of note regarding the contents of our listings is that the great majority of pairs are 'grammatical' collocations – i.e. they contain at least one non-lexical word (I take 'non-lexical' words to comprise prepositions, determiners, primary and modal verbs, conjunctions, subordinating adverbs, pronouns and numerals and ordinals other than *one* and *first*). Of the 1,000 collocations on the key collocation list, 763 are grammatical in this sense, and 237 are lexical (i.e. consist of two lexical words). Of the 656 collocations found by the collocates of keywords method, 421 (64%) are grammatical, 235 (36%) lexical. This may be a disappointment to some teachers. Gledhill (2000, pp. 73-79) has noted that many researchers systematically eliminate grammatical collocations from their analyses, considering collocation between lexical items to be the only sort worthy of examination; I also suspect that such items are not what many teachers have in mind when they think of collocation (see, for example, the definitions of collocation given by contributors to Lewis's edited collection of papers (2000)).

If lexical collocations are indeed the more interesting type, then the keyword-based listing, which has a slightly larger proportion of such items (36%, compared to 24% for the key collocation list) may be preferred. However (as Gledhill also argues), an exclusive focus on lexical collocations may be misguided. As Chapter 2 described, those linguistic frameworks which engage seriously with formulaic language have denied any absolute distinction between lexis and grammar, seeing the terms as referring to end-points on a spectrum, rather than clear and mutually-exclusive categories (e.g., Langacker, 1987, p. 3; Sinclair, 1991, p. 108). Pattern grammar asserts that supposedly 'abstract' grammatical patterns are often strongly associated with specific lexical instantiations (Hunston & Francis, 2000, p. 96), while Sinclair (2004b, pp. 30-35) and Hoey (2005, p. 40) have shown that lexical items often 'favour' particular grammatical forms. One benefit to learners of a listing of high frequency grammatical collocations is that the most typical versions of the patterns they need, and the most typical patterns of the words they need, can be brought to their attention.

To take a concrete example, the key collocations list includes 36 pairs instantiating the often-taught 'reporting' pattern 'verb + *that*'. Of these 36, many contain alternate forms of the same verb (e.g. *assume, assumed, assumes, assuming* all get separate entries). Collapsing these together leaves 16 distinct verbs: *argue, assume, conclude, confirm, demonstrate, emphasize, hypothesize, imply, indicate, note, predict, reveal, show, speculate, suggest, suppose*. Both lemmatised and non-lemmatised versions of this listing would, I suspect, be of great value to learners trying to get to grips with this pattern. Instead of simply learning the abstract form ('verb + *that*), learners could be introduced to the patterns through these instantiations. Learning the collocations as pairs may both provide learners with a good basis for getting to grips with the meaning and use of the pattern and bias them towards using it in the most lexically appropriate ways. It is also worth noting that such patterning underlines the benefits of including in our listing relatively infrequent collocations (e.g. *hypothesize that*, which occurs only 4.5 times per million words), which can be introduced to extend the range of particular higher-level patterns.

Another benefit of listing grammatical collocations is that they may draw attention to those productive patterns which are tied to specific lexis in a way that usually leads

them to be overlooked by traditional grammars (these would be examples of the more 'concrete' constructions dealt with in construction grammars). One example is the collocation *and-respectively*, as in:

> The survival rates after 12 and 24 months were 88 per cent **and** 83 per cent, **respectively**, for the dogs with single tumours.

> Two **and** one asterisks denote, **respectively**, that the estimates are statistically significant at the 5% and 10% levels…

This form is, I would suggest, likely to be of great use to students of academic English, playing as it does a vital text-organising function which would be difficult to manage by other means. A strong indicator of this usefulness is its high frequency of occurrence in the corpus – on average, 250 appearances per million words. It is also, I suspect, likely to be neglected by many EAP teachers. It seems likely that one reason that researchers have not been interested in grammatical collocations is that such pairs lack the striking salience of collocations like *significant difference* or *control group*; however, I would argue that this lack of salience makes it all the more important that researchers bring such items to teachers' and learners' attention, since they are otherwise likely to be passed over unnoticed.

Related to the prevalence of grammatical items on our collocation listings is the fact that many do not overlap with 'academic vocabulary' as it has been traditionally defined. Of the 1,000 collocations on the key collocations list, only 425 include an item from the AWL. Of the 656 collocations on the keyword-based list, 229 (35%) include an item from the AWL. I would argue that this lack of overlap indicates a shortcoming of traditional approaches to identifying academic vocabulary. The majority of the 509 individual words on the key collocations list which are not in Coxhead's list appear to have been excluded from the latter because they or one of their inflectionally or derivationally-related forms are found in West's General Service List. The figure cannot be determined precisely because it is not always obvious if a related form will have been excluded; however, I was able to identify 456 forms (90%) which seem likely to have been excluded for this reason. Examples of items in this category (together with the collocations in which they are listed) are:

address (*address-issue*)
control (*control-group*)
findings (findings-suggest; our-findings; similar-findings; these-findings)
means (by-means)
paper (*this-paper*)
resulting (*resulting-in*)

Such items highlight a serious disadvantage of Coxhead's strategy of eliminating from her word list any items which are related to words found on the GSL. While intermediate learners coming to EAP for the first time are likely to have met some form of the word families to which these words belong, many of the usages seen here are, I would suggest, likely not to be known. Few intermediate learners will be unfamiliar with the noun *address*, for example, but it seems unlikely that the verb form listed here – and still less its important academic collocation with *issue* – will be common knowledge to learners starting out in EAP. Similar remarks apply to *findings* and *resulting*, which learners are also more likely to known in other grammatical forms (the verb *find* and the noun *result*). In other cases, learners may be familiar with a word in the grammatical form listed, but not with its meaning. *Paper* and *control*, for example, take on special senses and enter into distinctive collocations in academic writing which learners are unlikely to have encountered elsewhere. *Means* combines these two factors – an unfamiliar form (learners are, I would suggest, most likely to be familiar with the verb *mean*) with a specialised meaning. Finally, high frequency vocabulary can be seen to take on meanings within a specific collocation which may be different from that with which learners are most familiar. This is seen in the example of *control-group*. These examples are not untypical of items which appear in my listings but are not on the AWL, and I would argue that they probably require specific pedagogical attention. Indeed, the fact that they are superficially familiar to learners may make them all the more problematic, since learners may not even notice when they have not understood them properly. The strategy of eliminating all high frequency words from academic word lists therefore seems a somewhat suspect one; many items which are excluded by this strategy may be of considerable importance for learners of EAP.

In sum, we have seen so far that the collocations on both of our listings range from the extremely frequent to the rather infrequent, that they contain a large proportion of 'grammatical' items, and that they are composed primarily of words which are not found on the AWL. The listing generated by the key collocations method produced a rather larger set of items than the keyword-based method, and this is related to the fact that many typically 'academic' words do not have generically academic collocations. We also saw that the majority of collocations on each listing were not found on the other. Word pairs on the key collocation listing were, on the whole, somewhat more frequent in academic writing than those on the keyword-based list, and included a rather higher proportion of 'grammatical' items, though these differences were not large. I have argued that grammatical collocations are worthy objects for teaching and that the fact that many words featured here are not on the AWL indicates a problem with traditional methods of identifying academic vocabulary. I have also noted the possibility that the very high frequency of some items may mean that they do not need explicitly to be taught, while the low frequency of others may not mean that they are too obscure to be worthy of learners' attention, if they form part of larger patterns. The value of higher and lower frequency items will be explored further in the next section.

The use of academic collocations by expert and novice writers

Introduction

This section will continue to evaluate the contents of our two sets of academic collocations. On the assumption that the best targets for learning are items which proficient writers use frequently but which learner writers do not, it will explore whether increased expertise in academic writing is associated with increased use of the items listed. As well as evaluating the lists as wholes, we will also examine further the issue of whether grammatical and very high frequency collocations need to be taught.

Materials

In order to evaluate whether expert writers make more use of our academic collocations than novices, several small corpora were created which would enable a comparison of the writing of academics at different levels of expertise. Beginner writers are represented by first year undergraduate essays; intermediate writers are

represented by third and fourth year undergraduate essays; expert writers are represented by research articles published in journals. Undergraduate essays were taken from the collection of writing compiled for the British Academic Written English (BAWE) corpus <http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/>. Research articles were taken from the same online journals used in compiling the main research article corpus (See Section 6.4). No articles found in the main corpus were used.

We saw in Section 6.4 that there is a high degree of heterogeneity in the vocabulary used in writing from different academic faculties. In particular, writing in the arts and humanities has a different profile from that of other disciplines. The present study will therefore include two separate analyses: one for writing in the arts and humanities, and one for writing in the sciences (at the time of writing, BAWE holdings in engineering, medicine, and the social sciences were too small for our present purposes). Separate comparative corpora were therefore compiled for these two areas. Their composition is shown in Tables 8 and 9. Since the principle comparison to be made will be between writing at these different levels of expertise, and since it is possible that writing from different subject areas will differ in the extent to which they use academic collocations, the principle criterion in balancing these corpora was that each level be similar in disciplinary make-up to the others. The corpora were therefore compiled such that the percentage contribution of each discipline represented was approximately equal across levels. In the first instance, an equal number of texts were taken from each discipline for each level (the number of texts used for each discipline was determined by what was available in BAWE); imbalances in terms of numbers of words were then adjusted by adding or removing texts as required. It was, in general, possible to retain roughly equal numbers of texts and equal proportions of words at each level. Since the inclusion of texts from different disciplines was constrained by what was available within BAWE, these corpora cannot be said to be a rigorously balanced representation of science or arts and humanities writing as a whole. Some major disciplines are not represented (e.g. chemistry, mathematics) and there is some imbalance between the disciplines which are included (e.g. the small number of texts from agriculture, psychology, archaeology and philosophy). It should also be noted that ‘science’ subjects in BAWE do not correspond exactly to those in our corpus.

Results should therefore be taken as provisional and in need of replication with other corpora.

Table 8: science comparison corpus

	1st year UG			3rd/4th year UG			journals		
Discipline	**texts**	**words**	**% of total words**	**texts**	**words**	**% of total words**	**texts**	**words**	**% of total words**
Agriculture	4	3,696	4.32	4	8,744	7.17	4	21,122	7.00
Biology	13	18,761	21.91	13	37,416	30.67	13	77,869	25.81
Computing	12	18,938	22.12	12	20,479	16.79	12	56,093	18.59
Food sciences	13	19,169	22.39	13	19,533	16.01	13	52,014	17.24
Physics	4	5,672	6.62	4	8,722	7.15	3	24,773	8.21
Psychology	12	19,380	22.64	12	27,109	22.22	6	69,789	23.13
Total	**58**	**85,616**	**100**	**58**	**122,003**	**100.00**	**51**	**301,660**	**100**

Table 9: arts and humanities comparison corpus

	1st year UG			3rd/4th year UG			journals		
Discipline	**texts**	**words**	**% of total words**	**texts**	**words**	**% of total words**	**texts**	**words**	**% of total words**
Archaeology	2	4,086	3.23	2	5,833	3.29	2	18,668	3.69
Classics	7	12,346	9.77	7	17,997	10.14	6	55,606	10.98
American studies	5	15,675	12.41	5	15,336	8.64	6	39,263	7.76
English	23	38,496	30.47	20	65,324	36.81	21	177,594	35.08
History	17	41,962	33.22	17	54,474	30.70	15	174,167	34.40
Linguistics	6	8,030	6.36	6	8,928	5.03	5	40,982	8.09
Philosophy	3	5,737	4.54	1	9,561	5.39	3	33,560	6.63
Total	**63**	**126,332**	**100**	**58**	**177,453**	**100**	**58**	**506,280**	**100**

The three levels of each corpus are intended to represent the language of writers with increasing levels of experience and expertise in academic writing. While it seems reasonable to suppose that the different levels indeed coincide with such increases, it needs to be noted that in comparing undergraduate essays with research articles, we are not only comparing different types of writer, but also – to a certain extent – different genres of text. It is therefore possible that any differences between the journal and undergraduate texts will be due to this difference in genre, rather than to a difference between writers. While this lack of parallelism may have some impact on results, I believe that the comparison can nevertheless be defended. While there are no doubt certain differences in the aims, methods, and intended audiences of the two genres, I would argue that in most cases a reasonably straight line can be drawn between undergraduate writing and published research articles. As Hyland (2008, p.

47) notes in defence of his own comparison of research articles with student writing, articles are the primary "model of good academic writing" which students are encouraged to emulate. There is therefore likely to be no closer parallel form of written output for expert and apprentice academics; undergraduate essays are likely to be far more similar to research articles than they are to textbooks, for example. Nevertheless, the difference in genre between the two text types will need to be borne in mind when interpreting results.

Analysis

Key collocations

Our analysis will first look at the occurrence at different levels of writing of those collocations identified on the key collocation method. The analysis will then be repeated for the list generated by the collocates of keywords method.

Our analysis will look first at the total number of academic collocations used in a given length of text. I argued in Section 5.4 that the validity of generalisations from whole-corpus frequency counts can be questioned because such counts provide no record of the variation between individual texts. My suggested solution was to make separate counts for each text in a corpus and to use these counts as the basis for standard inferential tests, in which each text is represented as an individual case. The same approach will be adopted here. First, the number of (tokens of) academic collocations found in each individual text is recorded. Since it would have been prohibitively labour-intensive to perform these counts using today's commercially available corpus interfaces, I created a specialised search tool using the Python programming language <www.python.org>, in conjunction with the Natural Language Tool Kit <www.sourceforge.net> to extract the data automatically. A subset of the derived data was checked against results provided by *WordSmith Tools* to confirm that they provided the same figures.

As the three levels of each corpus differed in size (getting larger from the lowest to the highest level – see Tables 8 and 9), collocation frequency counts were then normalised to occurrences per 500 words with the formula:

500 * total number of occurrences of keywords/total word tokens in text

Table 10 shows the median number of (tokens of) academic collocations found in texts at each level of science writing, normalised to occurrences per 500 words (medians are used because counts were not normally distributed across levels). As a point of comparison, it also shows the number of times the 770 individual words which make up these collocations (labelled 'collocation components'). Kruksall-Wallis tests showed significant differences between levels on both counts. Follow-up Mann-Whitney tests reveal significant differences between both sets of undergraduate texts and the journal articles in the number of collocations used (1st year vs. journals: $U = 874.0$, $p < .001$, $r = -.34$; 3rd/4th year vs. journals: $U = 829.0$, $p < .001$, $r = -.38$). However, there was no significant difference between the two sets of undergraduate writing ($U = 1640.0$, $p > .05$, $r = -.02$). A similar, but reversed, pattern was found for the number of individual collocating words used: both sets of undergraduates used more of these words than the journal articles (1st year vs. journals: $U = 671.0$, $p < .001$, $r = -.47$; 3rd/4th year vs. journals: $U = 950.0$, $p < .001$, $r = -.31$), while there was no significant difference between the two sets of undergraduates ($U = 1497.0$, $p > .05$, $r = -.09$).

Table 10: key academic collocation use (tokens) in different levels of science writing

	1st year	**3rd/4th year**	**Journal**	**Kruksall-Wallis**
median collocations/500 words	21.20	20.14	25.78	$H(2) = 19.06$ $p < .001$
median collocation components /500 words	216.86	216.74	201.10	$H(2) = 23.14$ $p < .001$

In short, although undergraduate writers are familiar with the individual words found on the collocation listing – indeed they make more use of them than journal writers – they are rather less likely to combine them in conventional collocations.

Repeating this analysis for the arts and humanities corpus gives a strikingly different pattern of results (Table 11). A Kruksall-Wallis test again indicates significant differences between levels in the number of collocations used. However, the trend runs in the opposite direction to that seen for science writers: Mann-Whitney tests

show that both sets of undergraduate writing make greater use of academic collocations than do journal articles (1st year vs. journals: $U = 874.0$, $p < .001$, $r = -.35$; 3rd/4th year vs. journals: $U = 829.5$, $p < .001$, $r = -.38$). As before, there was no significant difference between the two sets of undergraduates ($U = 1640.0$, $p > .05$, $r = -.02$). The use of individual words from the collocations followed the same pattern (1st year vs. journals: $U = 671.0$, $p < .001$, $r = -.47$; 3rd/4th year vs. journals: $U = 950.0$, $p < .001$, $r = -.31$; 1st year vs. 3rd/4th year: $U = 1497.0$, $p > .05$, $r = -.09$).

Table 11: key academic collocation use (tokens) at different levels of arts and humanities writing

	1st year	3rd/4th year	Journal	**Kruksall-Wallis**
median collocations/500 words	14.49	14.91	11.08	$H(2) = 19.06$ $p < .001$
median collocation components/500 words	198.71	197.33	186.98	$H(2) = 23.14$ $p < .001$

The collocations on our list do not, therefore, seem to be positively associated with developing expertise in arts and humanities writing; indeed, expert writers are rather less likely to use such items than beginners. It is also worth noting that these academic collocations appear to play a much smaller role overall in arts and humanities than in science writing. The median number of collocations used per 500 words across all levels of arts and humanities writing is 14.27, compared to 22.43 for science writing, a statistically significant difference ($U = 5276$, $p < .001$, $r = -.56$).

We saw in Section 5.4 that learners of English tend to make much use of modifier-noun collocations which have a very high frequency in the BNC, but fail to use many collocations with high MI scores. It is worth asking, therefore, whether collocations with either very high frequency or very high MI scores have a different status from other items on our listing. In the light of our previous results, it seems possible that the highest frequency items will not need to be explicitly taught because learners will already know them, while items with very high MI scores may warrant special pedagogical attention. Since academic collocations in general do not seem to be

important learning targets for writers in the arts and humanities, this analysis will be limited to the science corpus.

Since these learners' input will be best estimated by frequencies of occurrence within the science section of the main corpus, frequencies and MI scores were re-calculated for all collocations using that section only (the present analysis uses 'raw-frequency', rather than – as in Section 5.4 - t-score. This is to ensure that 'grammatical collocations' are included amongst the items). On the basis of these figures, two sub-lists were then created - one containing the 20% of collocations with the highest frequencies and one containing the 20% with the highest MI scores. The high frequency list comprised items with frequencies of at least 34 occurrences per million words; the high MI list comprised items with MI scores of at least 7.25.

Table 12 shows the use made by science writers at each level of these two sets of collocations. Both sets can be seen to follow the same general pattern as the collocation listing as a whole: there is an overall difference in use between levels, with both sets of undergraduate texts using significantly fewer collocations than journal articles, but not differing significantly from each other (high frequency collocations: 1st year vs. journals $U = 1136.0$, $p < .05$, $r = -.20$; 3rd/4th year vs. journals $U = 815.0$, $p < .001$, $r = -.39$; 1st year vs. 3rd/4th year $U = 815.0$, $p < .001$, $r = -.39$; high MI collocations: 1st year vs. journals $U = 826.0$, $p < .001$, $r = -.38$; 3rd/4th year vs. journals $U = 768.0$, $p < .001$, $r = -.41$; 1st year vs. 3rd/4th year $U = 1626.5$, $p > .05$, $r = -.03$). Collocations with very high frequency or mutual information do not, therefore, seem to behave differently from the other collocations on our lists.

Table 12: key academic collocation use (tokens) at different levels of science writing – high frequency/MI only

	1st year	3rd/4th year	Journal	**Kruksall-Wallis**
median collocations/500 words: most frequent 20% of collocations only (N = 200, min freq. = 34/million)	14.41	13.70	16.60	$H(2) = 14.33$ $p < .001$
median collocations/500 words: 20% highest MI collocations only (N = 200, min MI = 7.25)	2.50	2.54	3.67	$H(2) = 22.60$ $p < .001$

I noted above that a large proportion of the academic collocation list comprises 'grammatical collocations'. I argued at the time that such items are legitimate targets for learning. It might be thought, however, that grammatical collocations would already be familiar to learners, following simply from their knowledge of grammar. We can now test whether this is indeed the case. Again, our analysis will be limited to the science corpus. Table 13 shows the use by science writers at the three levels of the 749 grammatical and 251 lexical collocations on our listing. Both sets of collocations can be seen to follow the same pattern as the collocation list in general: there is an overall difference in use between levels, with both sets of undergraduate texts using significantly fewer collocations than journal articles, but not differing significantly from each other (grammatical collocations: 1st year vs. journals $U = 888.0$, $p < .001$, $r = -.34$; 3rd/4th year vs. journals $U = 820.0$, $p < .001$, $r = -.38$; 1st year vs. 3rd/4th year $U = 1623.0$, $p > .05$, $r = -.03$; lexical collocations: 1st year vs. journals $U = 920.0$, $p < .001$, $r = -.33$; 3rd/4th year vs. journals: $U = 820.0$, $p < .05$, $r = -.19$; 1st year vs. 3rd/4th year 1-3: $U = 1435.0$, $p > .05$, $r = -.13$). It does not seem to be the case, therefore, that grammatical collocations are picked up by learners more readily than lexical collocations.

Table 13: key academic collocation use (tokens) at different levels of science writing – grammatical vs. lexical collocations

	1st year	3rd/4th year	Journal	**Kruksall-Wallis**
median collocations/500 words: grammatical collocations only (N = 749)	18.74	18.32	22.48	$H(2) = 18.99$ $p < .001$
median collocations/500 words: lexical collocations only (N = 251)	1.63	1.80	2.48	$H(2) = 11.38$ $p < .005$

We saw in Section 5.4 that non-native learners of English are far more likely to repeatedly use the same collocations than are natives. Because of this, comparisons of collocation types in learner and native writing gave somewhat different results from comparisons of collocation types. It is therefore also worth asking how the number of academic collocation tokens used varies across levels in our comparative corpora.

Comparing collocation types across texts of different lengths (such as we have here) is rather more difficult than comparing tokens. Researchers have traditionally used a 'type-token ratio' (TTR) to standardise type counts. This is calculated by dividing the number of word types in a sample by the number of word tokens. This measure has proved problematic however. As Richards (1987) points out, TTR is strongly negatively correlated with sample length: as the number of tokens increases, TTR tends to decrease, other things being equal. In child language research, this has led to the paradoxical finding that older and more proficient children, who tend to produce longer samples of language, have lower TTRs than younger and less proficient children. In an attempt to overcome this problem, Malvern and Richards (1997) suggest a measure which they call the 'Mean Segmental Type-Token Ratio' (MSTTR). This calculates individual TTRs for successive text segments of a standard length and takes an average of these. The same approach is used by *WordSmith Tools* to generate a 'standardised type:token ratio' (Scott, 1999).

MSTTR will need to be adapted slightly for application to the present study. The diversity we are aiming to measure here differs from the usual case in that we are

looking not at the total range of vocabulary used, but rather at a particular set of items. Given this focus, there are two distinct ways in which diversity could be measured. First, we could measure the total number of collocation types used in a given length of text (e.g. collocation types per 500 words). This would give us an indication of the repertoire of collocations which a writer commonly employs. Second, we could measure the total number of collocation types used per collocation token. This would give us an indication of how much a writer repeats individual collocations (a score of 1 indicating no repetition; a score of 0.5 indicating an average of two appearances of each type). When researchers are interested in all of the vocabulary in a sample (rather than a specific list of items), these two measures are identical and repertoire is inseparable from repetition; in our case, however, the two are distinct. Since our interest is in the repertoire of collocations demonstrated, rather than in degree of repetition, we will focus exclusively on the former type.

Table 14 shows the median diversity in key collocation use, as measured by mean collocation types per 500 words, for each level of science writing. Following up this analysis with Mann-Whitney tests, we find that journal writing uses a significantly wider range of collocations than undergraduate writing, while there is no significant difference between the two sets of undergraduate writing (1st year vs. journals: U = 813.0, $p < .001$, r = -.42; 3rd/4th year vs. journals: U = 933.5, $p < .001$, r = -.32), but not between the two groups of undergraduates (U = 1581.5, $p > .05$, r = -.05)

Table 14: diversity in key academic collocation use at different levels of science writing

	1st years	**3rd/4th years**	**journal articles**	**Kruksall-Wallis**
median collocation types per node type/500 words	15.15	15.71	19.45	$H(2) = 18.11$ $p < .001$

Table 15 repeats this analysis for writing in the arts and humanities. Again, results repeat the patterns seen in the count of collocation tokens: Research articles use a significantly narrower range of academic collocations than undergraduates (1st year vs. journal U = 1229.50, $p < .001$, r = -.28; 3rd/4th year vs. journal U = 823.00, $p < .001$, r

= -.44), while the two sets of undergraduate writing do not differ significantly from each other (U = 1543.00, p > .05, r = -.13).

Table 15: diversity in key academic collocation use at different levels of arts and humanities writing

	1st years	**3rd/4th years**	**journal articles**	**Kruksall-Wallis**
median collocation types per node type/500 words	11.00	12.25	8.96	$H(2) = 22.60$ $p < .001$

Collocations of academic keywords

Table 16 replicates the above analyses of collocation tokens for science writing for the word pairs identified by the collocations of academic keywords method.

Table 16: collocations of academic keywords use (tokens) at different levels of science writing

		1st year	3rd/4th year	Journal	Kruksall-Wallis
all collocations	median collocations/500 words	7.71	8.64	11.68	$H(2) = 20.36$ $p < .001$
	median collocation components/500 words	200.27	195.40	176.90	$H(2) = 36.42$ $p < .001$
top 20% most frequent collocations (N = 132, freq. > 16/million in science)	median collocations/500 words	5.07	5.97	7.85	$H(2) = 24.47$ $p < .001$
	median collocation components /500 words	141.92	135.27	130.69	$H(2) = 16.29$ $p < .001$
top 20% highest MI collocations (N = 132, MI. > 6.7 in science)	median collocations/500 words	1.00	1.35	2.22	$H(2) = 18.21$ $p < .001$
	median collocation components /500 words	43.63	42.69	45.14	$H(2) = 16.29$ $p > .05$
grammatical collocations only (N = 421)	median collocations/500 words	6.73	7.16	9.22	$H(2) = 17.31$ $p < .001$
	median collocation components /500 words	175.43	171.68	155.50	$H(2) = 39.12$ $p < .001$
lexical collocations only (N = 235)	median collocations/500 words	1.00	0.92	0.71	$H(2) = 0.51$ $p > .05$
	median collocation components /500 words	39.76	33.34	42.32	$H(2) = 0.55$ $p > .05$

The patterns seen in Table 16 are similar to those found for the key collocation list: research articles use significantly more academic collocations than undergraduate writing. This pattern applies to the listing as a whole, to the most frequent 20% of collocations, to collocations with very high mutual information scores, and to grammatical collocations. The keyword-based list differs from the key collocation list, however, in that lexical collocations do not differ across levels. This is probably a result of the very low frequencies of occurrence of these items (between 0.7 and 1.0 occurrences per 500 words). Our previous finding that the two levels of undergraduate writing do not differ from each other in the amount of academic collocations used is replicated for the collocation listing as a whole and for grammatical collocations. However, this time 3rd/4th years use significantly more very high frequency collocations than do 1st years (Table 17). This suggests that some learning of these high frequency collocations may take place over time for undergraduates, though it should be noted that the absolute gain in small (from 5.07 to 5.97 collocations per 500 words) and the effect size modest (r = -.19).

Table 17: Mann-Whitney tests comparing collocation use at different levels of science writing

	1st year vs. journals	**3rd/4th year vs. journals**	**1st year vs. 3rd/4th year**
all collocations	U = 769.5, p < .001, r = -.41	U = 953.0, p < .001, r = -.31	U = 1412.0, p > .05, r = -.14
most frequent collocations	U = 827.0, p < .001, r = -.38	U = 989.0, p < .005, r = -.29	U = 1307.0, p < .05, r = -.19
high MI collocations	U = 795.0, p < .001, r = -.40	U = 1002.0, p < .005, r = -.28	U = 1424.0, p >.05, r = -.13
grammatical collocations	U = 974.5, p < .005, r = -.38	U = 989.0, p < .005, r = -.29	U = 1466.0, p > .05, r = -.11

As before, research articles in the arts and humanities use fewer academic collocations than undergraduate writing (Table 18). Also replicating our previous result, writing in the arts and humanities overall (averaged across the three levels) used significantly fewer academic collocations than writing in the sciences (arts and humanities Mdn = 5.79, science Mdn = 9.80, U = 7746.5, p < .001, r = -.42) Again, therefore, it seems

that the collocations we have identified do not constitute an important target for students in these areas.

Table 18: collocations of academic keywords use (tokens) at different levels of arts and humanities writing

		1st year	3rd/4th year	Journal	Kruksall-Wallis
all collocations	median collocations/500 words	6.51	5.83	4.51	$H(2) = 9.10$ $p < .01$
	median collocation components/500 words	185.99	186.41	173.11	$H(2) = 10.27$ $p < .01$

Tables 19 and 20 show the standardised number of collocation types used by writers at each level of the two comparative corpora. Again, these data replicate the familiar pattern: science research articles use a significantly wider range of collocations than undergraduates, while the two groups of undergraduates do not differ significantly from each other (1st year vs. journals $U = 844.50$, $p < .001$, $r = -.36$; 3rd/4th year vs. journals $U = 1040.00$, $p < .05$, $r = -.23$; 1st year vs. 3rd/4th year $U = 1323.50$, $p > .05$, $r = -.15$); arts and humanities research articles use a significantly narrower range of collocations than undergraduates, which do not differ from each other(1st year vs. journals $U = 1295.50$, $p < .005$, $r = -.25$; 3rd/4th year vs. journals $U = 1155.50$, $p < .005$, $r = -.27$; 1st year vs. 3rd/4th year $U = 1781.5$, $p > .05$, $r = -.02$)

Table 19: diversity in collocations of academic keywords use at different levels of science writing

	1st years	3rd/4th years	journal articles	Kruksall-Wallis
median collocation types per node type/500 words	6.40	7.25	9.25	$H(2) = 14.80$ $p < .001$

Table 20: diversity in collocations of academic keywords use at different levels of arts and humanities writing

	1st years	3rd/4th years	journal articles	Kruksall-Wallis
median collocation types per node type/500 words	5.50	5.00	3.87	$H(2) = 10.69$ $p < .005$

Summary and conclusions: the value and limitations academic collocations

I noted in Section 6.2 that the apparent topic- and genre-specificity of collocation might imply that a substantial listing of generic academic collocations cannot be identified. The present research shows that this idea is at least partly true. Though it proved possible to identify a reasonably large number of collocations that were common across academic disciplines, these items do not appear to be good learning targets for students in arts and humanities disciplines. Writers in these areas make little use of such collocations overall, and progression from undergraduate to journal writing is marked by an actual decrease in use. Section 6.3 showed that the vocabulary used by writers in the arts and humanities is rather different from that used in other academic disciplines. It now also seems that arts and humanities students are not likely to benefit from studying generically academic collocations. Together, these findings make a strong case for treating the needs of students in these areas separately from those other of EAP learners. This would both enable more relevant listings to be created for arts and humanities students and allow us to include in our inventory of generic academic vocabulary the many items which are currently excluded because they are not common in arts and humanities writing, in spite of their high frequencies in all other areas.

Further support for the idea that generic academic collocations are rare comes from the fact that neither of our identification procedures found a large number of lexical collocations. It may be, therefore, that the majority of such collocations are indeed specific to disciplines or small groups of disciplines. If it is only lexical collocations that we wish to teach, an academic collocation list may not be a viable project. I have argued, however, that collocation teaching should incorporate both lexical and grammatical pairings, since the latter, as much as the former, form an important part of the formulaic patterning that makes up proficient language. The finding that grammatical items show a similar profile of occurrence across levels of expertise as lexical pairs provides support for their inclusion.

If the need to teach grammatical collocations is accepted, and if the genre of EAP can be limited to disciplines outside of the arts and humanities, then the present research

suggests that an academic collocation list may be a viable project. Taken together, our two methods of identification were able to locate a substantial number of items. Since the use of these items was found to significantly distinguish expert science writers from novices, it seems likely that pairs from both lists will be useful targets for learning. This conclusion is not clear-cut, however. In particular, the finding that third and fourth year undergraduate science writers do not make any more use of academic collocations than first years (the one exception being very high frequency collocations on the keyword-based listing, where a small increase in use was found) should give us pause for thought.

One interpretation of this finding is that learners of academic English do not, in two to three years of undergraduate education, pick up the collocations they need. This reading would emphasise the importance of making such a listing a focus of explicit learning for students in these areas. However, in view of the evidence put forward in Chapter 5 that adult second language learners are capable of learning collocations from input, it would be rather surprising if undergraduates were not able to make any such progress over the length of their education. A second interpretation of the data is that the differences in aims, methods, and audience of undergraduate writing and research articles call for a different linguistic response. On this view, the lower number of academic collocations in undergraduate writing is not a sign of linguistic deficit, but rather of a competent response to a different language task (Hyland (2008) notes a similar possibility with regard to lexical bundles). All of the writing included in the corpus which provided our undergraduate data had received a minimum grade of 65% (a IIi in the British degree classification system) in university assessments (Nesi et al., 2005), so these writers have clearly been successful in meeting the demands put upon them. Were the appropriate data to become available, it would be interesting to see how the use of academic collocations differs between more and less successful student writers. Relative underuse of these collocations by weaker students, in comparison to stronger students, would give a much better indication that these collocations are important targets for undergraduate learning. If, on the other hand, our collocations simply mark a typological distinction between types of writing, they may not represent a particularly useful learning inventory for undergraduates. However, since they so clearly distinguish undergraduate from research writing, they

are likely to be an important learning focus for students making the transition from taught courses to research work – e.g. for first year doctoral students.

In sum, both methods of collocation identification described here were able to uncover substantial numbers of academic collocations. Since the two methods identified largely different sets of items, future research may wish to use both in tandem. Since many academic keywords do not appear to have generic academic collocations, and since many important academic collocations are made up of words which aren't specifically academic, it certainly does not seem wise to base academic collocation listings entirely around existing academic word lists. Future work may also wish to exclude arts and humanities disciplines from analyses of 'mainstream' academic vocabulary use in view of the distinctive profile of word and collocation use in this area. Finally, while increased use of academic collocations distinguishes writing in academic journals from student writing, it is not yet clear whether this is because undergraduates lack these linguistic resources or because of the different requirements of the two types of writing. In either case, however, it seems likely that students making the transition from undergraduate to research-based courses would benefit from studying these collocations.

6.6. Future directions: identifying longer collocations

The prime motivation for creating a listing of academic collocation was that, seen in the light of what we know about formulaic language, listings of individual words appear inadequate. Word lists fail to inform learners about the phrasal items which make up so much of the lexicon of competent language users, and even individual words may be misrepresented when analysed outside of their typical collocational contexts. We have made some progress towards overcoming these problems as they apply to listings of academic vocabulary by identifying two-word combinations which are likely to be important for students of EAP. However, this is only the first step towards identifying a truly phrasal academic vocabulary. As the overview in Chapter 2 made clear, phrasal relations stretch well beyond recurrent word pairs. Even leaving aside more abstract forms of phrasal patterning, such as semantic preference and prosody (Sinclair, 2004a, pp. 32-33), collocation (in the sense of the frequent of co-occurrence of words) is not limited to two-word combinations, but also refers to larger

mutually-predicting sets of items, which any thorough phrasal listing certainly ought to include but which will have been missed by our two-word listings.

One type of example is found in longer fixed chunks which have been artificially divided by our two-word search strategies. This can be illustrated by the three-word collocation *with respect to*, which our analysis has divided into the separate items *with-respect* and *respect-to*. That these collocations are not genuinely distinct from each other is demonstrated by the facts that 97% of occurrences of the *with-respect* collocation appear adjacent to the word *to*, while 94% of occurrences of the *respect-to* collocation appear adjacent to the word *with*.

More common – and more problematic – than such fixed forms are sets of collocations which are syntagmatically related to each other with more moderate degrees of regularity. This can be illustrated by the related collocations *there-significant, no-significant, statistically-significant, significant-differences,* and *significant-between*, which clearly stand in a syntagmatic relationship to each other (the exact phrase *there were no significant differences between* appears in its entirety 54 times in the 25-million word corpus), though the longer 'chunk' they form in combination is far from invariable, such that each of these collocations also frequently appears without the others. The associations between these collocations are further emphasised by the fact that some collocates of *significant* are also recorded as collocates of each other. Thus, the following word pairs, both parts of which are collocates of *significant* are also recorded as collocates of each other: *no-difference; there-differences; interaction-between; difference-between; differences-between; interaction-between*. In other cases, collocates of *significant* themselves have collocates such that the pair together frequently co-occurs with *significant.* This is seen in the cases of *main-effects* and *showed-no*. Neither *main* nor *showed* are recorded as collocates of *significant*. However, 27% of all occurrences of the collocation *main-effects* and 19% of occurrences of *showed-no* are found within a four-word span of *significant.*

In short, the collocations identified on our listings themselves have complex arrays of collocational relations. With this in mind, it is clear that a listing of two-word combinations cannot be said to have finally solved the problems inherent in single-

word lists; rather it has shifted them along one. Just as with single-word lists, by ignoring the syntagmatic context beyond our two-word pairings, we will miss much of what learners need to know, and run the risk of misrepresenting the items we do present. This is not to say that no progress has been made: we have learned much about two-word combinations and how they can be identified, and this should inform any further investigations. The question now, however, must be how we can move beyond our listings of two-word pairs to identify larger collocational patterns.

Counting multi-word lexical bundles - of the sort listed by Biber et al (2004) and Ellis et al (2007) – does not seem to offer a good solution. As I argued above, the lexical bundle approach is severely weakened by its failure to detect collocations which are discontinuous or which exhibit positional variability. One recent approach which does seem to offer a way forward however, is that of Cheng et al (2006), who have developed the concept of the c*oncgram* with the aim of identifying exactly the sorts of multi-word collocations we are pursuing. A concgram is defined as "all of the permutations of constituency variation and positional variation generated by the association of two or more words" (2006, p. 414). Concgrams are identified by an iterative building up of associations. The first step in this process is to find and list all of the unique words in a corpus. Second, associated word pairs are identified by listing all of the words which co-occur with each word within a given span. Three-word concgrams are then found by listing the words which occur within a span of each two-word pair. Four-word concgrams are found by identifying words which co-occur with the three-word concgrams, and so on. Cheng et al illustrate the sorts of item which can be identified by such a search with the 3-word concgram *Asia-world-city*, which is instantiated in their spoken corpus as *world city of of of Asia*; w*orld city of Asia; Asia world city*; and *Asia's world city* (2006, pp. 415-416). A facility for undertaking this type of analysis was originally instantiated in a program called ConcGram, developed by Greaves, and is now also available within *WordSmith Tools 5.0* (Scott, 2008)

This approach to finding collocations beyond the two-word level appears to be a promising one, overcoming the problems with discontinuous and positionally variable collocations suffered by lexical-bundle searches. However, much more research will be required into the behaviour of concgrams before any useful pedagogical listing can be produced. One key issue is that of how linguistically interesting concgrams might

be picked out from comprehensive listings. Cheng et al (2006) argue that applying minimum t-score or MI thresholds at the two-word level leads to many items of linguistic interest being ignored, and warn against their application. However, some automated (and so, presumably, frequency-based) filtering methods clearly need to be developed, given the huge number of items produced by concgram analysis: a search of a 20 million-word subset of our research article corpus (with arts and humanities texts excluded) using the ConcGram utility in WordSmith Tools with a 5-word search span, identifies 27,737 distinct two- to five-word concgrams based around the word *significant* alone.

A further issue is that concgrams – like collocations – are often not entirely distinct items, but overlap with each other. Putting concgram listings into a format usable for teachers and learners would require us to find ways of summarising such overlapping items in ways which highlight their commonalities while not disguising their variation. As an illustration, we can consider the 44 concgrams which include the collocation *statistically-significant* and which appear with a frequency of at least once per million words in the 20 million-word academic sub-corpus. These concgrams and their frequencies are shown in Table 21.

Table 21: concgrams including *statistically significant* which appear more than once per million words in academic writing

concgram	frequency per million words
ARE STATISTICALLY SIGNIFICANT	3.15
BE STATISTICALLY SIGNIFICANT	1.65
CONSIDERED STATISTICALLY SIGNIFICANT	3.95
CONSIDERED STATISTICALLY SIGNIFICANT RESULTS	1.85
DIFFERENCE WAS NOT STATISTICALLY SIGNIFICANT	1.35
NEGATIVE AND STATISTICALLY SIGNIFICANT	1.1
NO STATISTICALLY SIGNIFICANT DIFFERENCES	2.65
NO STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN	1.05
NOT STATISTICALLY SIGNIFICANT	11.85
NOT STATISTICALLY SIGNIFICANT IN	1.25
NOT STATISTICALLY SIGNIFICANT P	1.05
POSITIVE AND STATISTICALLY SIGNIFICANT	1.4
STATISTICALLY SIGNIFICANT AND	2.55
STATISTICALLY SIGNIFICANT AT	4.15
STATISTICALLY SIGNIFICANT AT THE	2.25
STATISTICALLY SIGNIFICANT DIFFERENCE	5.35
STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN	1.65
STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN THE	1
STATISTICALLY SIGNIFICANT DIFFERENCE IN	1.9
STATISTICALLY SIGNIFICANT DIFFERENCES	7.4
STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN	2.15
STATISTICALLY SIGNIFICANT DIFFERENCES IN	1.75
STATISTICALLY SIGNIFICANT EFFECT	1.15
STATISTICALLY SIGNIFICANT FOR	2.55
STATISTICALLY SIGNIFICANT IN	4.2
STATISTICALLY SIGNIFICANT IN THE	1.3
STATISTICALLY SIGNIFICANT INCREASE	1
STATISTICALLY SIGNIFICANT OF	1
STATISTICALLY SIGNIFICANT P	3.85
STATISTICALLY SIGNIFICANT RESULTS	3.2
STATISTICALLY SIGNIFICANT THE	2.95
THERE WAS A STATISTICALLY SIGNIFICANT	1.2
THERE WAS NO STATISTICALLY SIGNIFICANT	1.65
THERE WERE NO STATISTICALLY SIGNIFICANT	1.55
WAS A STATISTICALLY SIGNIFICANT	1.4
WAS CONSIDERED STATISTICALLY SIGNIFICANT	2.85
WAS CONSIDERED STATISTICALLY SIGNIFICANT RESULTS	1.4
WAS NO STATISTICALLY SIGNIFICANT DIFFERENCE	1.15
WAS NOT STATISTICALLY SIGNIFICANT	4.35
WAS STATISTICALLY SIGNIFICANT	2.85
WERE NO STATISTICALLY SIGNIFICANT	1.55
WERE NO STATISTICALLY SIGNIFICANT DIFFERENCES	1.35
WERE NOT STATISTICALLY SIGNIFICANT	3.05
WERE STATISTICALLY SIGNIFICANT	5.55

A number of consistent patterns can be observed here. Particularly prominent is the group of concgrams based around the form:

there	*was* *were*	*no*	*statistically significant difference(s)*	*between* *in*	*the*

This seems to be a good candidate for a multi-word collocation which could be offered to learners. Such a pattern is satisfying in that it can subsume several of the concgrams on our list. However, care must be taken not to overemphasise the regularity here, since many of these concgrams also take part in other phraseologically interesting patterns.

Consider, for example, the relationship between the four four-word concgrams:

statistically significant difference between
statistically significant difference in
statistically significant differences between
statistically significant differences in

and what we might call their three-word 'root':

statistically significant difference(s)

The latter concgram appears as a continguous bundle 248 times in our 20 million word corpus (the frequency of this fixed bundle is slightly lower than the frequency of the concgram, which includes seven non-continguous occurrences). Of these 248 occurrences, 144 are immediately followed by either *between* or *in*. So the four-word pattern accounts for 58% of occurrences of the three-word bundle. While this is clearly an important pattern, however, it will also be important to bring learners' attention to the large minority of occurrences of the three-word bundle which do not conform to it. One type of extension is found in the lower frequency prepositions which could take the place of *between* or *in*, e.g.:

There are statistically significant differences across *portfolios in their covariation with consumption growth*

There were no statistically significant differences by *group*

More importantly, we need to note the alternative patterns in which the shorter bundle appears, and whose existence might be obscured by a focus on the more prominent pattern described above. For example, inspection of a concordance listing for *statistically significant difference(s)* also reveals the less frequent, but potentially pedagogically valuable, pattern:

statistically significant difference(s)	*was/were*	*found*
		observed
		identified
		seen
		discovered
		determined
		revealed
		detected
		noted

Similarly, even when the longer, 4-word version of the concgram is found, it does not invariably fall into the full pattern given above. Learners' attention could also be profitably drawn, for example, to the patterns exemplified in:

Our results revealed statistically significant differences between…
We did not obtain statistically significant differences between…
We found no statistically significant differences between…

In short, care needs to be taken to ensure that 'summary patterns' (such as *there was/were no significant difference(s) in/between the*) do not obscure other less frequent but pedagogically-valuable forms. Uncovering the various patterns which can be found within concgrams in this manner and presenting them in ways which will be

accessible to learners seems likely to be a very worthwhile – though very challenging – task.

A method for identifying longer collocations which could be rather more easily implemented would be to use our existing collocation listings as the basis for a vocabulary highlighting program, of the sort developed for individual words by Cobb (http://www.lextutor.ca/vp/). Such a program would allow users to input a text of their choice and would return the same text with all academic collocations highlighted. We can see how this would work by looking at the following two paragraphs, taken from one research article from the comparative corpus described in Section 5.5. The article comes from the Journal of Dairy Science and reports research into changes in the pH of cheese during ripening (Upreti & Metzger, 2007):

> The different forms of P (i.e., water-soluble, organic, and bound inorganic P) **were determined** at d 1, and wk 1, 2, 3, 4, 8, 16, 32, and 48 during ripening. Total P in the cheeses **was determined by** ashing (550°C for 24 h) a 1-g cheese sample and olorimetrically determining the P content of the ash (AOAC, 1995; method number 991.25). Water-soluble P **was measured** on the filtrate obtained for soluble Ca analysis using a colorimetric method (AOAC, 1995; method number 991.25). The concentration of bound organic P **was measured** using a method we had previously developed (Upreti and Metzger, 2006b). This method utilizes 12% TCA to precipitate and isolate the casein present in cheese. Subsequently, this precipitate was ashed in the presence of calcium chloride and the P content of the ash **was determined**. The concentration of bound-inorganic P in cheese **was determined by** subtracting the **measured** values of water-soluble and organic P from the total P content of the cheese.
>
> **To characterize** the shifts in the ratio of water-soluble P to total P, water-soluble P as a **percentage of total** P (WSPTP) **was calculated**. Calculations indicated that WSPTP was in the range of 30 to 45% at d 1, and diverged to give a range of 30 to 60% after 48 wk of ripening (data not shown). **To evaluate** the **relationship between** WSPTP **with respect to** pH, values of

> WSPTP were plotted as a function of pH for all the cheeses studied (Figure 3b). A regression analysis indicated a linear **relationship between** WSPTP **and** cheese pH (r = 0.70). The regression line had a lower slope for WSPTP (= 54.45) **compared with** WSCTC (= 75.55), which indicated that **there** were **differences** in the relative solubilization of Ca **compared with** P. The ratio of the slopes of the regression line of WSCTC and WSPTP is 1.39 (75.55 ÷ 54.45 = 1.39). **This indicates that**, on average, for 1 mol of P, 1.39 mol of Ca is solubilized per unit change in pH. A **higher rate** of solubilization of Ca **compared with** P **with** a **decrease in** pH has been reported by others (Dolby et al., 1937; Czulak et al., 1969; Lucey and Fox, 1993).

A number of points can be made about this text. Firstly, a number of collocations of more than two words have been successfully highlighted. These include both continuous bundles - *was determined by; percentage of total; with respect to; this indicates that* – and discontinuous collocations - *relationship between…and*. Secondly, the highlighting emphasises the way in which a collocation may be repeated through the course of a text, sometimes with slight variations. This is seen, for example, in the first paragraph with the repetition of *was measured* and with the chain: *were determined-was determined by-was determined-was determined by*. Such repetitions are not only interesting only from the discourse perspective – demonstrating how repeated collocations can create cohesion in a text (Hoey, 1991), but also indicate how collocations can vary internally – cf. *was determined-were determined* – and how they can exist as both longer (*was determined by*) and shorter (*was determined*) chunks. Finally, potential 'collocations between collocations' which are not noted on our collocation listings also emerge through highlighting. This is seen here in such extended co-occurrences as:

> *percentage of…was calculated*
> *To evaluate the relationship between…with respect to…*
> *A higher rate of…compared with…*

While our present listings do not indicate these collocations of collocations as significant partners, it seems likely that they may represent useful patterns for any learners who would be reading this particular text.

This way of using the collocation lists accords well with current thinking on the teaching of collocations. Many teachers emphasise the importance of fostering learner independence in learning collocations, and recommend encouraging students to identify collocations in the texts they read (Conzett, 2000; Hill et al., 2000; Woolard, 2000). Such an approach is advantageous both in that it emphasises the collocations learners are most likely to need (i.e. those encountered in their own reading) and in that it demonstrates their use in authentic contexts, rather than merely providing abstracted lists of items which need to be reconextualized. However, a problem of this approach is that it is not clear, if learners are not already familiar with the collocations, how they are to know which word pairs are the high frequency ones. A tool such as the one suggested here would enable learners to identify the collocations in texts they read without the help of a teacher and without assuming prior knowledge of the thing they are trying to learn.

While this approach appears highly promising, there is a potential problem with the automatic highlighting of texts in that some word pairs may be highlighted when they are not involved in a genuine collocational relationship. This is seen, for example, in the highlighting of *by-measured* in the first paragraph above:

> **was determined by** subtracting the **measured** values

Here, *by* is highlighted both because it collocates with *determined* and because it collocates with *measured.* The former pairing is unproblematic. The latter word pair can also form an academic collocation, exemplified in examples such as:

> the front wing loads were **measured by** an additional balance
> energy of the ionizing laser was **measured by** the R-752 universal radiometer

However, the co-occurrence highlighted here is clearly not an instantiation of this pattern. Learners would need to be warned to be on the look-out against such potentially misleading cases.

6.7 Summary and conclusions: academic collocations

This chapter set out to examine whether it is feasible to construct a pedagogically-valuable listing of academic collocations, of the sort which is already widely used for the teaching of individual words. Two methods were described for deriving collocation listings and their outputs were evaluated. We found that the two approaches produced largely different set of collocations, and that use of both sets significantly distinguishes expert from learner writers, suggesting that they will both be good targets for learning at some point in students' academic careers. It seems, therefore, that future work in this area would benefit from using both approaches to collocation identification. Both listings yielded a large number of collocations of words which do not feature on the Academic Word List, suggesting that it would not be wise simply to use an existing listing of academic words as a starting point for identifying academic collocations.

While our results suggest that an academic collocation list is feasible, a number of caveats were noted. First, academic collocations are usually grammatical, rather than lexical. This goes against the archetypal notion of 'interesting collocations' held by many researchers and teachers. However, I have argued that grammatical collocations are in fact pedagogically important items. Second, writing in the arts and humanities differs from writing in other academic disciplines in its use of both individual words and of collocations to such an extent that future research may be best advised to treat arts and humanities separately from 'mainstream' academic writing. Third, and perhaps most important, listings of two-word collocations must be seen as merely a first step in producing a listing of academic phraseology. Two-word collocations themselves demonstrate complex arrays of collocational relations which also need to be accounted for. The concgram approach reported by Cheng et al (2006) appears to offer a promising route forward here, though much work remains to be done on identifying and organising the presentation of important concgram patterns. We also saw that a collocation highlighting program, modelled after Cobb's vocabulary highlighting program (http://www.lextutor.ca/vp/), may be a good method of bringing academic collocations to learners' attention in a way which demonstrates both the regularity and the variation of longer collocational patterns. Both of these approaches appear to offer good prospects for future work.

Chapter 7
Summary and conclusions: High frequency collocations and second language learning

This thesis has attempted to address three main questions regarding the implications of high frequency collocation for second language learning. The first was whether high frequency of occurrence in a general language corpus indicates that a word combination is likely to be part of most native speakers' mental systems, and hence something which second language learners ought to learn. Chapter 4 showed that frequency data do reliably predict psychological associations between words, as evidenced by word association norms. Various frequency measures were ranked according to their ability to predict associations, with the directional 'conditional probability' score doing best, and chi-squared and z-score being the best of the traditional corpus-linguistic methods. On all measures, the relationship between frequency and word association was very reliable but relatively modest in size. I have argued, however, that the methods used probably underestimate the true strength of the relationship because only the most salient mental collocations are likely to appear on word association norm lists.

In an attempt to tap mentally-represented collocations beyond the very salient pairs which are attested in such norms, a series of lexical decision tasks was used to measure the degree of 'priming' between collocating words. Results from these techniques were disappointing, however. While robust strategic and automatic priming was demonstrated between associated word pairs, priming was not found between collocating pairs which were not associates. These techniques therefore seem to add little to traditional word association tests in their ability to detect mental representations of collocations. These results question the validity of Hoey's (2005) claim that collocating words commonly co-occur because they prime each other. It is possible that more sophisticated techniques – especially techniques which do not rely on single-word presentation and which elicit a more natural form of response from participants – will find a processing advantage for a wider range of collocations.

However, it is not clear that such an advantage should be described in terms of 'priming', since this term has usually been used to refer to precisely the recognition advantage provided by single-word prompts which was tested in the studies reported here.

In sum, Chapter 4 provides evidence that there is a reliable link between corpus-based frequency data and the psychological representation of collocations. Frequency is *not*, therefore, a purely textual phenomenon. However, the precise strength of this relationship remains unclear and further research will be required to clarify exactly what various types of frequency data can tell us about likely mental representations. Until such research is completed, it is probably safe to assume that many high frequency collocations will be psychologically-valid targets for learners; however, teachers should remain aware that there may be mismatches between what is frequent and what learners need to learn. Frequency information needs, therefore, to be supplemented with other forms of evaluation.

The second question addressed by this thesis was that of whether adult second language learners tend to acquire the collocations they meet in input. Our results suggest that, contrary to the claims of some researchers, learners may indeed acquire many of the collocations they meet on a regular basis; though it seems that the relatively low levels of input to which learners are typically exposed tends to leave them with a distinctively 'non-nativelike' profile of collocational knowledge. In particular, the lower-frequency but highly salient collocations identified by high mutual information scores tend not to be well learned. We also found that enhanced input – provided here in the form of repeated exposure – can increase collocation learning dramatically.

Taken together, these results suggest that, though learners can be expected to pick up some collocations implicitly, teachers should also provide an explicit focus on collocation learning – and especially on the learning of those very salient pairs which learners appear to have difficulty in acquiring. Many more questions remain to be answered in this area, however. First, it is not clear how many exposures the average learner usually requires to achieve a stable representation of a collocation, or how much time can elapse between exposures and effective learning still take place.

Larger-scale longitudinal studies exercising tight control over learner input will be required to determine this typical course of collocation learning. Second, it seems likely that some form of conscious attention to a collocation will improve learning, but it is not clear what form this attention should best take. The study reported in Section 5.3 achieved good short-term results through phonological repetition; it would be interesting to see how durable this effect is over time and how it compares with more meaning-focused activities, such as translation or comprehension tasks. Third, it seems likely that – regardless of frequency - some types of collocation will be easier to learn than others, perhaps because they are more salient to learners for semantic reasons or because they are reminiscent of L1 collocations. Unpacking these factors may give us an additional, non-frequency-based, criterion for selecting which collocations ought to be explicitly taught. Such information would provide a valuable addition to attempts to construct listings of important collocations for learners. Fourth, L1 acquisition research has suggested that wide individual differences exist between children in the degree to which their learning approach is formula- or word-based. It is important to determine to what extent such variation also exists for second language learning adults and what factors affect this variation. Finally, the question – not addressed by any of the studies reported here – of whether and how the learning of formulaic language relates to the development of a creative language system remains very much open and an important priority for future work.

The third question addressed by this thesis was that of whether, and how, a pedagogically-useful listing of academic collocations can be constructed. Our results suggest that it can, but that its contents will differ from what many teachers and researchers have typically considered 'interesting collocations'. In particular, most collocations which are used across academic disciplines are 'grammatical' collocations, and most do not involve vocabulary which is found on Coxhead's (2000) Academic Word List. I have suggested that the traditional focus on 'lexical collocations' is a misguided one and that Coxhead's strategy of excluding from 'academic vocabulary' any words related to items in general vocabulary is also mistaken. Moreover, I have argued that for the purposes of future vocabulary research and teaching, 'academic English' should be partitioned into two separate types – mainstream academic English and English for arts and humanities. These two streams

show distinct, but internally homogenous, patterns of use which mean that they are best represented as separate genres.

The principle challenge for future research in this area is to develop ways of identifying important collocations of more than two words and of presenting these collocations in ways which are both sufficiently concise to be pedagogically useable and sufficiently rich to do justice to the huge range of important patterns from which learners might benefit. Two possible methods for this have been explored, with promising early results. However, much work remains to be done in this area.

A final, somewhat broader conclusion, relating to all of the three research strands described here concerns the nature of collocation itself. This thesis has followed 'neo-Firthians' such as Sinclair and Hoey in defining collocations as word pairs which co-occur in texts with greater than random frequency. This view of collocation has some powerful advantages. One is that it allows collocations to be identified rapidly and with a high degree of reliability from samples of text far too large for human analysts to reliably handle. Another is that it helps us to identify those collocations which are semantically and syntactically regular but which many psychologically-oriented views of language maintain are likely to have a special 'holistic' status in the language system because of their high frequencies of occurrence. While the frequency-based approach is, thanks to these advantages, a powerful paradigm for studying language, it also has at least one fundamental shortcoming, which is reflected in the limitations of the studies reported here: i.e., it is based entirely on the formal co-occurrence of items, without consideration of the functional aspects of those forms.

The shortcomings of this approach are highlighted in a number of ways by the studies in this thesis. First, the priming studies described in Section 4.5 appear to indicate that there is a difference between 'mere' frequent collocations and collocations which are also psychological associates, in that the latter was shown to demonstrate priming while the former did not. Since both types were equally 'high-frequency collocations', this suggests that there are important aspects of collocation which cannot be explained in terms of frequency alone. Future research needs to unpick what additional factors are at work here, but to do so will necessarily involve moving beyond a paradigm in which collocations are simply defined in terms of their frequencies of occurrence.

Second, it was noted as a limitation of the acquisition studies reported in Sections 5.3 and 5.4 that they dealt only with the establishment of formal links between the constituent words of collocations, without considering how learners learn to use these collocations appropriately. This limitation is a direct concomitant of a purely formal, frequency-based, approach to the subject. Again, to progress in this area we will ultimately need to move beyond form to consider collocation use in the round. Thirdly, we saw in Section 6.6 that a formal frequency-based approach to collocation struggles to unravel multi-word patterns of co-occurring collocations (i.e. collocations between collocations) into distinct target items for learners. It seems likely that such patterns could be unpicked more successfully if function were integrated into our analyses. Moreover, listings of formal collocations such as those presented in Appendices D and E are, in their raw form, inadequate from a pedagogical point of view in that they offer the learner no obvious point of access. If such listings could be indexed by function – so that, for example, a student of EAP could look up 'collocations used to report data' or 'collocations used to cite other research', etc. – they would be of far greater pedagogical value. Such an approach would be more in tune with a communicative approach to language teaching, offering a 'syllabus' of collocations based on communicative functions, rather than linguistic forms.

In short, formal frequency of co-occurrence is a helpful, but ultimately limited paradigm for studying collocation. Further progress may require something closer to a construction grammar approach (as described in Section 2.3), in which form and function are seen as equally important aspects of linguistic items. The predominant focus within corpus linguistics on purely formal analyses of collocation has perhaps been inevitable, given the level to which most corpora have so far been developed. In the absence of sophisticated systems for integrating detailed and large-scale functional/semantic annotations, corpus linguists who want to identify collocations are still largely restricted to basing their searches on formal features – especially on orthographic words. While rich forms of corpus annotation are starting to be developed (McEnery et al., 2006, pp. 29-45), these are neither of the right sort nor (since identifying collocations requires such large corpora) on a large enough scale to enable an adequate extension of the current paradigm. Developing such systems should be a major priority for future research in the area.

References

ACTFL. (1985). *ACTFL proficiency guidelines*. New York: ACTFL Materials Center.

Adolphs, S., & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 107-126). Amsterdam: John Benjamins Publishing Company.

Aitchison, J. (1987). *Words in the mind: An introduction to the mental lexicon*. Oxford: Blackwell Publishing.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Arnold, I. V. (1986). *The English word*. Moscow: Vuisshaya Shkola.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101-114.

Balota, D. A. (1994). Visual word recognition: The journey from features to meaning. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 340-357). San Diego: Academic Press.

Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage-based models of language*. Stanford, California: CSLI Publications.

Barsalou, L., W. (1992). *Cognitive psychology: an overview for cognitive scientists*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tubingen: Gunter Narr Verlag Tubingen.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar : individual differences and dissociable mechanisms*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D. (2006). *University language*. Amsterdam: John Benjamins.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bley-Vroman, R. (2002). Frequency in production, comprehension, and acquisition. *Studies in Second Language Acquisition, 24*(02), 209-213.

Bobrow, S. A., & Bell, S. M. (1973). On catching on to idiomatic expressions. *Memory and Cognition, 1*, 343-346.

Bolander, M. (1989). Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. In K. Hyltenstam & L. Obler (Eds.), *Bilingualism across the lifespan: aspects of acquisition, maturity and loss* (pp. 73-86). Cambridge: Cambridge University Press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11-53). London: John Wiley and Sons, Inc.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language, 27*, 668-683.

Campion, M., & Elley, W. (1971). *An academic vocabulary list*. Wellington: New Zealand Council for Educational Research.
Carter, R. (2004). *Language and creativity: the art of common talk*. London: Routledge.
Charles, W. G., & Miller, G. A. (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics, 10*, 357-375.
Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International journal of corpus linguistics, 11*(4), 411-433.
Chomsky, N. (1965). *Aspects of a theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1991). Linguistics and adjacent fields: a personal view. In A. Kasher (Ed.), *The Chomskyan turn* (pp. 26-53). Oxford: Blackwell.
Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics, 16*(1), 22-29.
Clark, R. (1974). Performing without competence. *Journal of child language, 1*(1), 1-10.
Clear, J. (1993). Tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp. 271-292). Amsterdam: Benjamins.
Conzett, J. (2000). Integrating collocation into a reading & writing course. In M. Lewis (Ed.), *Teaching collocation: further developments in the lexical approach* (pp. 70-87). Boston: Thomson.
Cook, G. (2000). *Language play, language learning*. Oxford: Oxford University Press.
Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes, 23*, 397-423.
Coulmas, F. (1981). Introduction: conversational routine. In F. Coulmas (Ed.), *Conversational routine: explorations in standardized communication situations and prepatterned speech* (pp. 1-17). The Hague: Mouton Publishers.
Cowie, A. P. (1981a). Lexicography and its Pedagogic Applications: An Introduction. *Applied Linguistics, 2*(3), 203-206.
Cowie, A. P. (1981b). The Treatment of Collocations and Idioms in Learners' Dictionaries. *Applied Linguistics, 2*(3), 223-235.
Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 1-12). Houndsmills: Macmillan.
Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 3168-3171). Oxford: Pergamon.
Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 1-20). Oxford: Oxford University Press.
Cowie, A. P. (1998b). Phraseological dictionaries: some east-west comparisons. In A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 209-228). Oxford: Oxford University Press.
Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly, 34*(2), 213-238.
Coxhead, A. (2008). Phraseology and English for academic purposes. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149-161). Amsterdam: John Benjamins Publishing Company.
Croft, W. (2001). *Radical construction grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.
Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cummings, J., Benson, D. F., Walsh, M. J., & Levine, H. L. (1979). Left to right transfer of language dominance: a case study. *Neurology, 29*, 1547-1550.
De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon on EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). London: Addison Wesley Longman.
de Groot, A. M. B., & Nas, G. L. J. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of memory and language, 30*, 90.
De Keyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125-151). Cambridge: Cambridge University Press.
Dörnyei, Z., Durow, V., & Zahran, K. (2004). Individual differences and their effects on formulaic sequence acquisition. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 87-106). Amsterdam: John Benjamins Publishing Company.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics, 19*(1), 61-74.
Durrant, P. (2007). Review of Nadja Nesselhauf. *Collocations in a learner corpus*. *Functions of language, 14*(2), 251-261.
Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.
Ellis, N. C. (2002a). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(02), 143-188.
Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24*(02), 297-339.
Ellis, N. C. (2003). Constructions, chunking, and connectionism: the emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Oxford: Blackwell.
Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*(2), 305-352.
Ellis, N. C., & Larsen-Freeman, D. (2006). Language Emergence: Implications for Applied Linguistics--Introduction to the Special Issue (Vol. 27, pp. 558-589).
Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2007). *The processing of formulas in native and second-language speakers: psycholinguistic and corpus determinants*. Paper presented at the The 25th UWM Linguistics Symposium.
Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1998). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: The MIT Press.
Evert, S. (2004). Computational approaches to collocations. Retrieved 14 December, 2007, from www.collocations.de
Evert, S., & Krenn, B. (2001). *Methods for the qualitative evaluations of lexical association measures.* Paper presented at the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France.
Farghal, M., & Obeidat, H. (1995). Collocations: a neglected variable in EFL. *International review of applied linguistics in language teaching, 33*(4), 315-331.
Field, A. (2005). *Discovering statistics using SPSS* (Second ed.). London: Sage.
Fillmore, C., J. (1979). On fluency. In C. Fillmore, J., D. Kempler & S.-Y. W. Wang (Eds.), *Individual differences in language ability and language behaviour* (pp. 85-101). New York: Academic Press.

Fillmore, C., J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language, 64*(3), 500-538.
Firth, J. R. (1957). Modes of meaning. In *Papers in linguistics 1934-1951* (pp. 190-215). Oxford: Oxford University Press.
Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. R. Palmer (Ed.), *Selected papers of J.R. Firth 1952-1959* (pp. 168-205). Harlow: Longman.
Fitzpatrick, T. (2007). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319-331.
Foster, P. (2001). Rules and routines: a consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 75-94). London: Longman.
Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of language, 6*(1), 22-42.
Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition, 8*, 149-156.
Gibbs, R. W., & Nayak, N. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*, 100-138.
Gibbs, R. W., Nayak, N., & Cutting, C. (1989). How to kick the bucket and not decompose: analyzability and idiom processing. *Journal of memory and language, 28*, 576-593.
Ginzberg, R. S., Khidekel, S. S., Knyazeva, G. Y., & Sankin, A. A. (1979). *A course in modern English lexicology* (2nd ed.). Moscow: Vuisshaya Shkola.
Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. London: Chicago University Press.
Goldberg, A. E. (2006). *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.
Goldberg, A. E. (2007). *Learning the general from the specific*. Paper presented at the UWM linguistics symposium on formulaic language.
Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.
Granger, S., Dagneaux, E., & Meunier, F. (2002). *International Corpus of Learner English*. Louvain: UCL Presses Universitaires de Louvain.
Graves, R., & Landis, T. (1985). Hemispheric control of speech expression in aphasia. *Archives of Neurology, 42*, 249-251.
Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J.R.Firth* (pp. 148-162). London: Longmans, Green and Co. Ltd.
Handl, S. (2008). Essential collocations for learners of English: The role of collocational direction and weight. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (Vol. 43-65). Amsterdam: John Benjamins Publishing Company.
Hansen, L., Umeda, Y., & McKinney, M. (2002). Savings in the relearning of second language vocabulary: The effects of time and proficiency. *Language Learning, 52*(4), 653-678.
Haswell, R. (1991). *Gaining ground in college writing: tales of development and interpretation*. Dallas: Southern Methodist University Press.

Herbst, T. (1996). What are collocations: sandy beaches or false teeth? *English Studies, 4*, 379-393.
Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching collocations: Further developments in the lexical approach* (pp. 88-117). Boston: Thomson.
Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and cognitive processes, 6*(3), 169-205.
Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
Hoffman, S., & Lehmann, H.-M. (2000). Collocational evidence from the British National Corpus. In J. M. Kirk (Ed.), *Corpora Galore: Analyses and techniques in describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998)* (pp. 17-32). Amsterdam: Rodopi.
Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161-186). Oxford: Oxford University Press.
Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
Hunston, S., & Francis, G. (2000). *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10*(4), 785-813.
Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1).
Hyland, K., & Tse, P. (2007). Is there an 'Academic Vocabulary'? *TESOL Quarterly, 41*(2), 235-253.
Hymes, D. H. (1972). On communicative competence. In J.B.Pride & J.Holmes (Eds.), *Sociolinguistics*. Harmondsworth: Penguin.
Ingvar, D. H. (1983). Serial aspects of language and speech relatedd to prefrontal cortical activity. A selective review. *Human neurobiology, 2*, 177-189.
Irujo, S. (1986). A piece of cake: learning and teaching idioms. *ELT Journal, 40*(3), 236-242.
James, W. (1890). *The Principles of Psychology*. New York: Holt.
Jespersen, O. (1924/1976). Living grammar. In D. D. Bornstein (Ed.), *Readings in the theory of grammar* (pp. 82-93). Cambridge, MA: Winthrop Publishers.
Jiang, N., & Nekrasova, T., M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*(3), 433-445.
Jones, S., & Sinclair, J. M. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de lexicologie, 24*, 15-61.
Kaszubski, P. (2000). *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: a contrastive, corpus-based approach.* Adam Mickiewicz Univeristy, Poznán.
Katz, J. J. (1973). Compositionality, idiomaticity, and lexical substitution. In S. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 357-376). New York: Holt, Reinhart, and Winston.
Kay, P., & Fillmore, C., J. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language, 75*(1), 1-33.

Kemmer, S., & Barlow, M. (2000). Introduction: a usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage based models of language* (pp. vii-xxviii). Stanford: CSLI Publications.
Kempler, D., Van Lancker, D., Marchman, V., & Bates, E. (1999). Idiom comprehension in children and adults with unilateral brain damage. *Developmental Neuropsychology, 15*, 327-349.
Kent, G. H., & Rosanoff, A. J. (1910). A study of association in insanity. *American Journal of Insanity, 67*, 37-96, 317-390.
Kinsbourne, M. (1971). The minor cerebral hemisphere as a source of aphasic speech. *Transactions of the American neurological association, 96*, 141-145.
Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: University Press.
Kjellmer, G. (1990). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111-127). London: Longman.
Klarskov Mortensen, H. J. (2003). Significance tests calculator. Retrieved 22 January 2008, from http://www.hjkm.dk/TZ/default.htm
Krashen, S., & Scarcella, R. (1978). On routines and patterns in language acquisition and performance. *Language Learning, 28*(2), 283-300.
Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 37-54). Amsterdam: John Benjamins Publishing Company.
Lamb, S. (2000). Bidirectional processing in language and related cognitive systems. In M. Barlow & S. Kemmer (Eds.), *Usage based models of language* (pp. 87-119). Stanford, CA: CSLI Publications.
Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume 1 Theoretical prerequisites*. Stanford: Stanford University Press.
Langacker, R. W. (1991). *Foundations of cognitive grammar: Volume 2: Descriptive application*. Stanford: Stanford University Press.
Larsen, B., Skinhoj, E., & Lassen, H. A. (1978). Variations in regional cortical blood flow in the right and left hemispheres during automatic speech. *Brain, 10*, 193-200.
Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language learning & technology, 5*(3), 37-72.
Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Longman.
Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. London: Thomson Heinle.
Lewis, M. (Ed.). (2000). *Teaching collocations: further developments in the lexical approach*. Boston: Thomson.
Lieven, E., V.M., Pine, J. M., & Dresner Barnes, H. (1992). Individual differences in early vocabulary development: redefining the referential-expressive distinction. *Journal of child language, 19*, 287-310.
Lieven, E., V.M., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 168-196). London: Routledge.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492-527.
Lorenz, G. (1999). *Adjective intensification - learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.
Lucas, M. (2000). Semantic priming without association: a meta-analytic review. *Psychonomic Bulletin & Review, 7*(4), 618-630.
MacWhinney, B. (Ed.). (1999). *The emergence of language*. London: Lawrence Erlbaum.
Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon: Multilingual matters.
Manly, B. F. J. (2005). *Multivariate statistical methods: A primer*. London: Chapman & Hall/CRC.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for specific purposes, 19*, 63-86.
McEnery, T., & Wilson, A. (2001). *Corpus linguistics: an introduction* (Second ed.). Edinburgh: Edinburgh University Press.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Abingdon: Routledge.
McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1155-1172.
McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: semantic mediated priming occurs in lexical decisions. *Journal of memory and language, 27*, 545-559.
Mel'cuk, I. (1998). Collocations and lexical foundations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 23-53). Oxford: Oxford University Press.
Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental psychology, 90*, 227-234.
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.
Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning, 48*(3), 323-364.
Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2. *Studies in second language acquisition, 21*(01), 49-80.
Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: General, 106*(3), 226-254.
Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G. W. Humphreys

(Eds.), *Basic processes in reading: visual word recognition* (pp. 264-336). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Nelson, K. (1981). Individual differences in language development: implications for development and language. *Developmental Psychology, 17*(2), 170-187.
Nesi, H., Gardner, S., Forsyth, R., Hindle, D., Wickens, P., Ebeling, S., et al. (2005). *Towards the compilation of a corpus of assessed student writing: An account of work in progress.* Paper presented at the Corpus Linguistics 2005, University of Birmingham.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.
Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.
Newmeyer, F. (2003). Grammar is grammar and usage is usage. *Language, 79*, 682-707.
Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language, 70*(3), 491-538.
Oppenheim, N. (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 220-240). Ann Arbor: University of Michigan Press.
Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.
Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.
Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological Research, 66*, 180-194.
Peters, A. M. (1977). Language-learning strategies: does the whole equal the sum of the parts? *Language, 53*(3), 560-573.
Peters, A. M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.
Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1223-1237.
Pine, J. M., & Lieven, E., V.M. (1993). Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech. *Journal of child language, 20*, 551-571.
Pinker, S. (1999). *Words and rules: the ingredients of language*. London: Phoenix.
Praninskas, J. (1972). *American university word list*. London: Longman.
Rastle, K., Harrington, J., & Coltheat, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly journal of experimental psychology, 55*(A), 1339-1362.
Raupach, M. (1984). Formulae in second language speech production. In H. W. Dechert, D. Mohle & M. Raupach (Eds.), *Second language productions* (pp. 114-137). Tubingen: Gunter Narr.
Richards, B. (1987). Type/token ratios: what do they really tell us? *Journal of child language, 14*, 201-209.
Robinson, P., & Ellis, N. C. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.

Ryding, E., Bradvik, B., & Ingvar, D. H. C. (1987). Changes of regional cerebral blood flow measured simultaneously in the right and left hemisphere during automatic speech and humming. *Brain 110*, 1345-1358.
Saussure, F. D. (1916/1965). *Course in general linguistics*. New York: McGraw-Hill.
Scarcella, R. (1979). Watch up!: a study of verbal routines in adults second language performance. *Working papers on Bilingualism, 19*, 79-88.
Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition, 14*, 357-385.
Schmitt, N. (forthcoming). Instructed second language vocabulary learning. *Language teaching research*.
Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: a longitudinal study. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 55-86). Amsterdam: John Benjamins Publishing Company.
Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 127-151). Amsterdam: John Benjamins Publishing Company.
Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 173-189). Amsterdam: John Benjamins Publishing Company.
Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32*, 219-250.
Scott, M. (1996). WordSmith Tools. Oxford: Oxford University Press.
Scott, M. (1999). WordSmith Tools users help file. Oxford: Oxford University Press.
Scott, M. (2007). *Homing in on the text-initial cluster*. Paper presented at the Aston corpus symposium.
Scott, M. (2008). WordSmith Tools 5.0. Oxford: Oxford University Press.
Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 382-408). Oxford: Blackwell.
Segalowitz, N., & Hulstijn, J. H. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 371-388). Oxford: Oxford University Press.
Sereno, J. A. (1991). Graphemic, associative, and syntactic priming effects as a brief stimulus onset asynchrony in lexical decision and naming *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 459-477.
Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 272-281.
Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In memory of J.R. Firth* (pp. 410-430). London: Longman.
Sinclair, J. M. (1987). Collocation: a progress report. In R. Steele & T. Threadgold (Eds.), *Language topics: Essays in honour of Michael Halliday* (Vol. 2, pp. 319-331). Amsterdam: John Benjamins.
Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (2004a). The search for units of meaning. In *Trust the text: language, corpus and discourse* (pp. 24-48). London: Routledge.
Sinclair, J. M. (2004b). Trust the text. In *Trust the text* (pp. 9-23). London: Routledge.
Sinclair, J. M. (Ed.). (1990). *Collins COBUILD English grammar*. London: Harper-Collins.
Sinclair, J. M. (Ed.). (1995). *Collins COBUILD English dictionary*. London: Harper-Collins.
Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, 64*(3), 429-258.
Speedie, L. J., Wertman, E., T'air, J., & Hellman, K. M. (1993). Disruption of automatic speech following a right basal-ganglia lesion. *Neurology, 43*, 1768-1774.
Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of psycholinguistic research, 19*, 317-330.
Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative methods. *Functions of language, 2*(1), 1-33.
Stubbs, M. (1996). *Text and corpus analysis*. Oxford: Blackwell.
Stubbs, M. (2001). Texts, corpora, and problems of interpretation: a response to Widdowson. *Applied Linguistics, 22*(2), 149-172.
Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior, 18*, 523-534.
Tannen, D. (1989). *Talking voices: repetition, dialogue and imagery in conversational discourse*. Cambridge: Cambridge University Press.
Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teacher's college, Columbia University.
Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics, 31*, 1655-1674.
Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge MA, London: Harvard University Press.
Tomasello, M., & Brooks, P. J. (1999). Early syntactic development: a construction grammar approach. In M. Barret (Ed.), *The development of language* (pp. 161-190). Cambridge: Cambridge University Press.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The Development of Fluency in Advanced Learners of French. *Applied Linguistics, 17*(1), 84-119.
Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (in preparation). Processing advantages of lexical bundles.
Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: an eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 153-172). Amsterdam: John Benjamins Publishing Company.
Upreti, P., & Metzger, L. E. (2007). Influence of calcium and phosphorus, lactose, and salt-to-moisture ratio on cheddar cheese quality: pH changes during ripening. *Journal of dairy science, 90*(1), 1-12.
Van Ek, J. A., & Alexander, L. G. (1980). *Threshold Level English*. Oxford: Pergamon Press.
Van Lancker-Sidtis, D. (2004). When novel sentences spoken or heard for the first time in the history of the universe are not enough: toward a dual-process model of language. *International Journal of Language and Communication Disorders, 39*(1), 1-44.

Van Lancker, D., & Cummings, J. (1999). Expletives: neurolinguistic and neurobehavioral perspectives on swearing. *Brain research reviews, 31*, 83-104.
Van Lancker, D., & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain and Language, 32*, 265-277.
Van Lancker, D., McIntosh, R., & Grafton, S. (2003). PET activation studies comparing two speech tasks widely used in surgical mapping: localization of Broca's area. *Brain and Language, 85*, 245-261.
Weinreich, U. (1963). Lexicology. In T. Sebeok (Ed.), *Current trends in linguistics* (Vol. 1, pp. 60-93). The Hague: Mouton.
West, M. (1953). *A general service list of English words*. London: Longman.
Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied Linguistics, 10*(2), 128-137.
Wilkins, D. A. (1976). *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development*. London: Oxford University Press.
Williams, G. C. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics, 3*(1), 151-171.
Williams, J. N. (1996). Is automatic priming semantic. *European journal of cognitive psychology, 8*(2), 139-151.
Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *The Canadian modern language review, 63*(1), 13-33.
Woolard, G. (2000). Collocation - encouraging learner independence. In M. Lewis (Ed.), *Teaching collocation: further developments in the lexical approach* (pp. 28-46). Boston: Thomson.
Wray, A. (1992). *The focusing hypothesis: the theory of left hemisphere lateralized language re-examined*. Amsterdam: John Benjamins.
Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied linguistics, 21*(4), 463-489.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), *Formulaic sequences: acquisition, processing and use* (pp. 249-268). Amsterdam: John Benjamins.
Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua, 117*, 543-578.
Xue, G., & Nation, P. (1984). A university word list. *Language learning and communication, 3*, 215-229.
Yalden, J. (1987). *The communicative syllabus: evolution, design and implementation*. London: Prentice-Hall International.
Yang, H. Z. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing, 1*(2), 93-103.
Yorio, C. A. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly, 14*(4), 433-442.
Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam & L. K. Obler (Eds.), *Bilingualism across the lifespan* (pp. 55-72). Cambridge: Cambridge University Press.

Appendix Ai: Items for Priming Study One

Set 1

condition	prime	target	BNC occurrences	t-score	MI
collocate	atomic	bomb	102	10.10	11.63
collocate	death	penalty	324	17.97	9.26
collocate	deep	sigh	49	6.98	8.80
collocate	elder	brother	246	15.68	11.13
collocate	foreign	affairs	1074	32.74	9.87
collocate	gold	rush	52	7.19	8.23
collocate	heart	attack	644	25.33	9.03
collocate	huge	amounts	73	8.52	8.32
collocate	intense	heat	48	6.91	8.50
collocate	middle	ages	660	25.67	10.52
collocate	opinion	polls	425	20.61	12.41
collocate	peace	talks	379	19.44	9.36
collocate	private	sector	1823	42.66	10.23
collocate	soft	drinks	135	11.61	10.01
collocate	violent	crime	85	9.20	8.83
collocate	waiting	list	346	18.54	8.20
collocate	warning	signs	80	8.92	8.58
collocate	wild	flowers	172	13.09	9.23
collocate	window	frames	79	8.88	9.58
collocate	wire	fence	55	7.41	10.57
unrelated	atomic	extent	0	n/a	n/a
unrelated	death	purpose	0	n/a	n/a
unrelated	deep	corps	0	n/a	n/a
unrelated	elder	jacket	0	n/a	n/a
unrelated	foreign	tiles	0	n/a	n/a
unrelated	gold	release	0	n/a	n/a
unrelated	heart	leader	0	n/a	n/a
unrelated	huge	mining	0	n/a	n/a
unrelated	intense	movies	0	n/a	n/a
unrelated	middle	boots	0	n/a	n/a
unrelated	opinion	school	0	n/a	n/a
unrelated	peace	coast	0	n/a	n/a
unrelated	private	shirt	0	n/a	n/a
unrelated	soft	assets	0	n/a	n/a
unrelated	violent	yacht	0	n/a	n/a
unrelated	waiting	beings	0	n/a	n/a
unrelated	warning	factor	0	n/a	n/a
unrelated	wild	error	0	n/a	n/a
unrelated	window	guns	0	n/a	n/a
unrelated	wire	trees	0	n/a	n/a

nonword	atomic	spenk	n/a	n/a	n/a
nonword	atomic	sprinz	n/a	n/a	n/a
nonword	death	mugned	n/a	n/a	n/a
nonword	death	spads	n/a	n/a	n/a
nonword	deep	galds	n/a	n/a	n/a
nonword	deep	ghworned	n/a	n/a	n/a
nonword	elder	glipce	n/a	n/a	n/a
nonword	elder	kafts	n/a	n/a	n/a
nonword	foreign	moarsts	n/a	n/a	n/a
nonword	foreign	squemb	n/a	n/a	n/a
nonword	gold	gnoidd	n/a	n/a	n/a
nonword	gold	leathe	n/a	n/a	n/a
nonword	heart	snuscks	n/a	n/a	n/a
nonword	heart	strecte	n/a	n/a	n/a
nonword	huge	gimped	n/a	n/a	n/a
nonword	huge	plenc	n/a	n/a	n/a
nonword	intense	skroate	n/a	n/a	n/a
nonword	intense	tefe	n/a	n/a	n/a
nonword	middle	phraull	n/a	n/a	n/a
nonword	middle	queams	n/a	n/a	n/a
nonword	opinion	cranths	n/a	n/a	n/a
nonword	opinion	jeefs	n/a	n/a	n/a
nonword	peace	cwiked	n/a	n/a	n/a
nonword	peace	skeuged	n/a	n/a	n/a
nonword	private	kourgs	n/a	n/a	n/a
nonword	private	shruite	n/a	n/a	n/a
nonword	soft	phrylc	n/a	n/a	n/a
nonword	soft	swazz	n/a	n/a	n/a
nonword	violent	clemped	n/a	n/a	n/a
nonword	violent	scwacte	n/a	n/a	n/a
nonword	waiting	creeves	n/a	n/a	n/a
nonword	waiting	phlorpe	n/a	n/a	n/a
nonword	warning	founnth	n/a	n/a	n/a
nonword	warning	thwyffs	n/a	n/a	n/a
nonword	wild	crupced	n/a	n/a	n/a
nonword	wild	danns	n/a	n/a	n/a
nonword	window	ghwaigg	n/a	n/a	n/a
nonword	window	spylgn	n/a	n/a	n/a
nonword	wire	snokes	n/a	n/a	n/a
nonword	wire	wouch	n/a	n/a	n/a

Set 2

condition	prime	target	BNC occurrences	t-score	MI
collocate	coal	mining	130	11.39	10.34
collocate	crucial	factor	104	10.17	8.58
collocate	deputy	leader	187	13.65	9.00
collocate	dual	purpose	47	6.84	8.81
collocate	east	coast	501	22.35	9.30
collocate	fatal	error	28	5.28	9.10
collocate	human	beings	1302	36.07	11.94
collocate	leather	jacket	140	11.83	11.02
collocate	lesser	extent	404	20.09	11.18
collocate	liquid	assets	102	10.09	9.92
collocate	luxury	yacht	12	3.46	9.40
collocate	machine	guns	101	10.03	9.26
collocate	officer	corps	32	5.64	8.23
collocate	palm	trees	148	12.16	10.55
collocate	press	release	367	19.11	8.78
collocate	primary	school	980	31.20	8.16
collocate	riding	boots	22	4.68	8.28
collocate	roof	tiles	40	6.32	9.76
collocate	silent	movies	13	3.60	8.50
collocate	silk	shirt	64	7.99	10.20
unrelated	coal	ages	0	n/a	n/a
unrelated	crucial	heat	0	n/a	n/a
unrelated	deputy	attack	0	n/a	n/a
unrelated	dual	crime	0	n/a	n/a
unrelated	east	penalty	0	n/a	n/a
unrelated	fatal	sigh	0	n/a	n/a
unrelated	human	polls	0	n/a	n/a
unrelated	leather	brother	0	n/a	n/a
unrelated	lesser	bomb	0	n/a	n/a
unrelated	liquid	sector	0	n/a	n/a
unrelated	luxury	talks	0	n/a	n/a
unrelated	machine	talks	0	n/a	n/a
unrelated	officer	rush	0	n/a	n/a
unrelated	palm	fence	0	n/a	n/a
unrelated	press	amounts	0	n/a	n/a
unrelated	primary	list	0	n/a	n/a
unrelated	riding	signs	0	n/a	n/a
unrelated	roof	affairs	0	n/a	n/a
unrelated	silent	flowers	0	n/a	n/a
unrelated	silk	drinks	0	n/a	n/a
nonword	coal	steubs	n/a	n/a	n/a
nonword	coal	culfed	n/a	n/a	n/a
nonword	crucial	dreaze	n/a	n/a	n/a

nonword	crucial	goarlds	n/a	n/a	n/a
nonword	deputy	falk	n/a	n/a	n/a
nonword	deputy	cwanged	n/a	n/a	n/a
nonword	dual	sleinte	n/a	n/a	n/a
nonword	dual	scwink	n/a	n/a	n/a
nonword	east	pawks	n/a	n/a	n/a
nonword	east	trenths	n/a	n/a	n/a
nonword	fatal	sckorlt	n/a	n/a	n/a
nonword	fatal	serle	n/a	n/a	n/a
nonword	human	strul	n/a	n/a	n/a
nonword	human	joargn	n/a	n/a	n/a
nonword	leather	gwinse	n/a	n/a	n/a
nonword	leather	scwyfe	n/a	n/a	n/a
nonword	lesser	murms	n/a	n/a	n/a
nonword	lesser	slaub	n/a	n/a	n/a
nonword	liquid	flygued	n/a	n/a	n/a
nonword	liquid	stursh	n/a	n/a	n/a
nonword	luxury	vadds	n/a	n/a	n/a
nonword	luxury	ghlofte	n/a	n/a	n/a
nonword	machine	fromfed	n/a	n/a	n/a
nonword	machine	creubbe	n/a	n/a	n/a
nonword	officer	froomth	n/a	n/a	n/a
nonword	officer	phirbed	n/a	n/a	n/a
nonword	palm	phansed	n/a	n/a	n/a
nonword	palm	blymph	n/a	n/a	n/a
nonword	press	folphed	n/a	n/a	n/a
nonword	press	gruiffs	n/a	n/a	n/a
nonword	primary	skwabb	n/a	n/a	n/a
nonword	primary	splirm	n/a	n/a	n/a
nonword	riding	ghwunse	n/a	n/a	n/a
nonword	riding	kwett	n/a	n/a	n/a
nonword	roof	ghraked	n/a	n/a	n/a
nonword	roof	skwave	n/a	n/a	n/a
nonword	silent	phlizz	n/a	n/a	n/a
nonword	silent	kwalned	n/a	n/a	n/a
nonword	silk	chunths	n/a	n/a	n/a
nonword	silk	scresc	n/a	n/a	n/a

Appendix Aii: Items for Priming Study Two

Set 1

condition	prime	target	BNC occurrences	t-score	MI
Level 1	subject	content	32	5.41	4.54
Level 1	former	student	26	4.85	4.35
Level 1	human	culture	47	6.62	4.85
Level 1	greater	concern	40	6.08	4.68
Level 1	likely	effects	40	5.94	4.02
Level 1	special	unit	51	6.81	4.43
Level 1	recent	figures	31	5.25	4.14
Level 1	complex	series	28	5.04	4.40
Level 2	past	decade	357	18.85	8.58
Level 2	armed	struggle	117	10.80	9.16
Level 2	double	doors	115	10.69	8.35
Level 2	foreign	debt	226	14.98	8.04
Level 2	stone	floor	67	8.08	6.29
Level 2	music	hall	162	12.59	6.55
Level 2	colour	scheme	126	11.10	6.53
Level 2	rapid	growth	243	15.56	9.07
Level 3	estate	agent	328	18.10	10.49
Level 3	current	affairs	188	13.64	7.53
Level 3	cutting	edge	173	13.13	9.27
Level 3	feature	film	67	8.11	6.75
Level 3	village	green	105	10.10	6.08
Level 3	card	game	38	6.03	5.54
Level 3	pretty	girl	87	9.21	6.31
Level 3	parish	church	411	20.23	9.03
control	future	owner	3	1.12	1.51
control	strange	freedom	2	1.15	2.43
control	chief	impact	2	0.83	1.28
control	fixed	text	2	0.27	0.30
control	marked	loss	2	0.95	1.62
control	direct	risk	2	0.44	0.54
control	entire	stay	2	1.02	1.84
control	pure	kind	2	0.86	1.36
control	simple	links	2	1.00	1.76
control	real	motion	3	1.14	1.54
control	central	index	3	1.22	1.74
control	proper	homes	2	1.14	2.39
control	fine	welcome	2	0.83	1.29
control	funny	picture	2	1.09	2.13
control	milk	bill	3	1.37	2.26
control	tiny	range	3	1.13	1.52
control	famous	train	2	1.06	2.00
control	unique	match	2	1.13	2.30

control	normal	list	2	0.34	0.39
control	silent	march	2	1.02	1.84
control	chosen	paper	3	1.17	1.63
control	stupid	word	2	1.01	1.79
control	dark	office	3	-0.07	-0.06
control	massive	room	3	1.02	1.28

Set 2

condition	prime	target	BNC occurrences	t-score	MI
Level 1	true	owner	24	4.72	4.78
Level 1	total	freedom	21	4.36	4.35
Level 1	real	impact	38	5.89	4.51
Level 1	full	text	63	7.67	4.89
Level 1	complete	loss	26	4.82	4.17
Level 1	lower	risk	25	4.71	4.13
Level 1	short	stay	65	7.77	4.80
Level 1	worst	kind	34	5.63	4.88
Level 2	close	links	181	13.39	7.66
Level 2	slow	motion	107	10.32	8.68
Level 2	price	index	168	12.90	7.63
Level 2	private	homes	77	8.65	6.19
Level 2	warm	welcome	180	13.38	8.72
Level 2	mental	picture	60	7.67	6.63
Level 2	finance	bill	78	8.72	6.27
Level 2	narrow	range	93	9.54	6.57
Level 3	express	train	38	6.10	6.64
Level 3	football	match	147	12.07	7.89
Level 3	shopping	list	144	11.96	8.28
Level 3	protest	march	42	6.40	6.25
Level 3	daily	paper	59	7.52	5.54
Level 3	spoken	word	101	9.98	7.26
Level 3	post	office	1324	36.32	9.17
Level 3	waiting	room	130	11.16	5.58
control	fish	content	2	0.96	1.65
control	modern	student	4	1.54	2.11
control	prison	culture	2	1.05	1.94
control	active	concern	2	0.89	1.44
control	useful	effects	4	1.47	1.90
control	fresh	unit	2	0.91	1.50
control	minor	figures	3	1.41	2.44
control	older	series	2	0.53	0.68
control	lost	decade	3	1.34	2.13
control	current	struggle	2	0.98	1.70
control	brown	doors	2	1.15	2.42
control	farm	debt	2	1.16	2.46
control	vast	floor	2	1.06	2.01

control	wooden	hall	2	1.14	2.34
control	normal	scheme	2	0.38	0.46
control	extra	growth	2	0.57	0.74
control	major	agent	4	1.42	1.79
control	complex	affairs	2	0.90	1.46
control	south	edge	2	0.23	0.25
control	grey	film	2	1.03	1.89
control	winter	green	4	1.51	2.03
control	lovely	game	3	1.23	1.80
control	tired	girl	2	1.04	1.93
control	biggest	church	3	1.21	1.72

Word-non-word pairs

prime	**target**
adult	jeefs
awful	wouch
blood	phrylc
brain	danns
channel	clemped
classic	spylgn
clean	creeves
clear	crupced
client	thwyffs
code	queams
constant	plenc
crucial	mugned
damage	kourgs
date	phlorpe
drive	ghwaigg
error	skroate
extreme	galds
food	moarsts
friendly	scwacte
late	cranths
lead	skeuged
leader	strecte
live	shruite
love	spenk
male	sprinz
market	founnth
near	snuscks
options	squemb
port	snokes
practice	joargn
present	spads
prime	glipce
quiet	kafts
ready	leathe
record	swazz
return	tefe

royal	steubs
severe	gnoidd
show	culfed
square	goarlds
starting	falk
strong	sleinte
target	scwink
trading	pawks
white	trenths
working	sckorlt
wrong	serle
youth	strul

Appendix Aiii: Items for Priming Study Three

Set 1

condition	prime	target	BNC occurrences	t-score	MI
Level 0	famous	saying	2	0.61	0.81
Level 0	weak	ground	2	1.03	1.87
Level 0	market	experts	3	1.18	1.66
Level 0	royal	lunch	2	0.89	1.42
Level 0	fixed	levels	2	0.88	1.40
Level 0	huge	powers	2	1.05	1.95
Level 0	direct	danger	2	0.93	1.55
Level 0	main	concept	2	0.30	0.34
Level 1	subject	content	32	5.41	4.54
Level 1	former	student	26	4.85	4.35
Level 1	human	culture	47	6.62	4.85
Level 1	greater	concern	40	6.08	4.68
Level 1	likely	effects	40	5.94	4.02
Level 1	special	unit	51	6.81	4.43
Level 1	recent	figures	31	5.25	4.14
Level 1	complex	series	28	5.04	4.40
Level 2	past	decade	357	18.85	8.58
Level 2	armed	struggle	117	10.80	9.16
Level 2	double	doors	115	10.69	8.35
Level 2	foreign	debt	226	14.98	8.04
Level 2	stone	floor	67	8.08	6.29
Level 2	music	hall	162	12.59	6.55
Level 2	colour	scheme	126	11.10	6.53
Level 2	rapid	growth	243	15.56	9.07
Level 3	estate	agent	328	18.10	10.49
Level 3	current	affairs	188	13.64	7.53
Level 3	cutting	edge	173	13.13	9.27
Level 3	feature	film	67	8.11	6.75
Level 3	village	green	105	10.10	6.08
Level 3	card	game	38	6.03	5.54
Level 3	pretty	girl	87	9.21	6.31
Level 3	parish	church	411	20.23	9.03
Control	useful	office	0	n/a	n/a
Control	front	links	0	n/a	n/a
Control	single	mixture	0	n/a	n/a
Control	central	list	0	n/a	n/a
Control	final	range	0	n/a	n/a
Control	simple	match	0	n/a	n/a
Control	strong	loss	0	n/a	n/a
Control	easy	train	0	n/a	n/a
Control	true	motion	0	n/a	n/a
Control	real	stay	0	n/a	n/a
Control	complete	status	0	n/a	n/a
Control	short	room	0	n/a	n/a
Control	total	owner	0	n/a	n/a
Control	full	welcome	0	n/a	n/a
Control	lower	word	0	n/a	n/a

Control	worst	access	0	n/a	n/a
Control	close	bill	0	n/a	n/a
Control	price	engine	0	n/a	n/a
Control	warm	freedom	0	n/a	n/a
Control	finance	impact	0	n/a	n/a
Control	slow	balance	0	n/a	n/a
Control	private	march	0	n/a	n/a
Control	mental	paper	0	n/a	n/a
Control	narrow	risk	0	n/a	n/a
Control	express	text	0	n/a	n/a
Control	shopping	picture	0	n/a	n/a
Control	daily	tower	0	n/a	n/a
Control	post	measure	0	n/a	n/a
Control	football	homes	0	n/a	n/a
Control	protest	index	0	n/a	n/a
Control	spoken	journey	0	n/a	n/a

Set 2

condition	prime	target	BNC occurrences	t-score	MI
Level 0	useful	balance	2	0.80	1.20
Level 0	final	status	4	1.29	1.50
Level 0	front	engine	2	0.70	0.99
Level 0	simple	access	3	0.88	1.02
Level 0	single	tower	2	0.99	1.75
Level 0	strong	mixture	2	1.06	1.99
Level 0	central	measure	2	0.55	0.71
Level 0	easy	journey	2	0.94	1.57
Level 1	true	owner	24	4.72	4.78
Level 1	total	freedom	21	4.36	4.35
Level 1	real	impact	38	5.89	4.51
Level 1	full	text	63	7.67	4.89
Level 1	complete	loss	26	4.82	4.17
Level 1	lower	risk	25	4.71	4.13
Level 1	short	stay	65	7.77	4.80
Level 1	worst	kind	34	5.63	4.88
Level 2	close	links	181	13.39	7.66
Level 2	slow	motion	107	10.32	8.68
Level 2	price	index	168	12.90	7.63
Level 2	private	homes	77	8.65	6.19
Level 2	warm	welcome	180	13.38	8.72
Level 2	mental	picture	60	7.67	6.63
Level 2	finance	bill	78	8.72	6.27
Level 2	narrow	range	93	9.54	6.57
Level 3	express	train	38	6.10	6.64
Level 3	football	match	147	12.07	7.89
Level 3	shopping	list	144	11.96	8.28
Level 3	protest	march	42	6.40	6.25
Level 3	daily	paper	59	7.52	5.54
Level 3	spoken	word	101	9.98	7.26
Level 3	post	office	1324	36.32	9.17
Level 3	waiting	room	130	11.16	5.58
Control	famous	powers	0	n/a	n/a

Control	fixed	concern	0	n/a	n/a
Control	weak	scheme	0	n/a	n/a
Control	huge	agent	0	n/a	n/a
Control	market	danger	0	n/a	n/a
Control	direct	experts	0	n/a	n/a
Control	royal	growth	0	n/a	n/a
Control	main	saying	0	n/a	n/a
Control	subject	film	0	n/a	n/a
Control	former	doors	0	n/a	n/a
Control	human	series	0	n/a	n/a
Control	greater	floor	0	n/a	n/a
Control	likely	content	0	n/a	n/a
Control	special	hall	0	n/a	n/a
Control	recent	culture	0	n/a	n/a
Control	complex	decade	0	n/a	n/a
Control	past	ground	0	n/a	n/a
Control	armed	concept	0	n/a	n/a
Control	double	girl	0	n/a	n/a
Control	foreign	green	0	n/a	n/a
Control	stone	affair	0	n/a	n/a
Control	music	struggle	0	n/a	n/a
Control	colour	debt	0	n/a	n/a
Control	rapid	student	0	n/a	n/a
Control	estate	unit	0	n/a	n/a
Control	current	lunch	0	n/a	n/a
Control	cutting	figures	0	n/a	n/a
Control	feature	church	0	n/a	n/a
Control	village	game	0	n/a	n/a
Control	card	effects	0	n/a	n/a
Control	pretty	levels	0	n/a	n/a

Word-non-word pairs

prime	target
actual	blymph
afraid	clemped
alone	cranths
ancient	creeves
annual	creubbe
blue	crupced
brief	culfed
broad	danns
brown	falk
certain	flygued
clean	folphed
clear	founnth
common	fromfed
crucial	froomth
dead	galds
eastern	ghlofte
empty	ghwaigg

equal	glipce
extreme	gnoidd
fast	goarlds
female	gruiffs
fine	gwinse
flat	jeefs
formal	joargn
glad	kafts
golden	kourgs
grand	leathe
grey	moarsts
hard	mugned
healthy	pawks
inner	phansed
larger	phirbed
latest	phlorpe
legal	phrylc
light	plenc
living	queams
longer	rurdes
lovely	sckorlt
male	scwacte
middle	scwink
minor	scwyfe
modern	serle
nice	shruite
perfect	skeuged
poor	skroate
prime	slaub
proper	sleinte
proud	snokes
quiet	snuscks
rich	spads
rural	spenk
smaller	sprinz
solid	spylgn
stupid	squemb
sweet	steubs
thick	strecte
thin	strul
tiny	stursh
tired	swazz
unique	tefe
vital	thwyffs
western	trenths
white	vadds
wide	wouch

Appendix Aiv: Items for Priming Study Four

Set One

condition	prime	target
associate	fruit	APPLE
pure collocate	dear	BOY
unrelated	homes	BUTTER
neutral	*	CARPET
associate	hat	COAT
pure collocate	poured	CUP
unrelated	liked	DOCTOR
neutral	*	FLOWER
associate	hand	FOOT
pure collocate	worked	HARD
unrelated	afraid	HAMMER
neutral	*	HIGH
associate	cold	HOT
pure collocate	anger	HURT
unrelated	mental	JUMP
neutral	*	LIGHT
associate	find	LOSE
pure collocate	middle	NIGHT
unrelated	goal	PEPPER
neutral	*	QUEEN
associate	stream	RIVER
pure collocate	expects	SELL
unrelated	sites	SHEEP
neutral	*	SLEEP
associate	fast	SLOW
pure collocate	faint	SMILE
unrelated	p1	SOUR
neutral	*	STREET
associate	chair	TABLE
pure collocate	common	THREAD
unrelated	user	TOWN
neutral	*	WALK

Set Two

condition	prime	target
pure collocate	big	APPLE
unrelated	output	BOY
neutral	*	BUTTER
associate	rug	CARPET
pure collocate	black	COAT
unrelated	listen	CUP
neutral	*	DOCTOR
associate	stem	FLOWER
pure collocate	six	FOOT
unrelated	noted	HARD
neutral	*	HAMMER
associate	low	HIGH
pure collocate	bowl	HOT
unrelated	detail	HURT
neutral	*	JUMP
associate	dark	LIGHT
pure collocate	stand	LOSE
unrelated	neck	NIGHT
neutral	*	PEPPER
associate	king	QUEEN
pure collocate	stretch	RIVER
unrelated	equal	SELL
neutral	*	SHEEP
associate	bed	SLEEP
pure collocate	deep	SLOW
unrelated	file	SMILE
neutral	*	SOUR
associate	road	STREET
pure collocate	oak	TABLE
unrelated	raise	THREAD
neutral	*	TOWN
associate	run	WALK

Set Three

type	prime	target
unrelated	draw	APPLE
neutral	*	BOY
associate	bread	BUTTER
pure collocate	thick	CARPET
unrelated	double	COAT
neutral	*	CUP
associate	nurse	DOCTOR
pure collocate	wild	FLOWER
unrelated	occur	FOOT
neutral	*	HARD
associate	nail	HAMMER
pure collocate	metres	HIGH
unrelated	talks	HOT
neutral	*	HURT
associate	skip	JUMP
pure collocate	throw	LIGHT
unrelated	imagine	LOSE
neutral	*	NIGHT
associate	salt	PEPPER
pure collocate	beauty	QUEEN
unrelated	crisis	RIVER
neutral	*	SELL
associate	wool	SHEEP
pure collocate	go	SLEEP
unrelated	prince	SLOW
neutral	*	SMILE
associate	sweet	SOUR
pure collocate	parked	STREET
unrelated	star	TABLE
neutral	*	THREAD
associate	city	TOWN
pure collocate	short	WALK

Set Four

type	prime	target
neutral	*	APPLE
associate	girl	BOY
pure collocate	sugar	BUTTER
unrelated	bridge	CARPET
neutral	*	COAT
associate	coffee	CUP
pure collocate	phoned	DOCTOR
unrelated	card	FLOWER
neutral	*	FOOT
associate	soft	HARD
pure collocate	club	HAMMER
unrelated	laid	HIGH
neutral	*	HOT
associate	pain	HURT
pure collocate	triple	JUMP
unrelated	pair	LIGHT
neutral	*	LOSE
associate	day	NIGHT
pure collocate	add	PEPPER
unrelated	pound	QUEEN
neutral	*	RIVER
associate	buy	SELL
pure collocate	beef	SHEEP
unrelated	quick	SLEEP
neutral	*	SLOW
associate	frown	SMILE
pure collocate	turned	SOUR
unrelated	client	STREET
neutral	*	TABLE
associate	needle	THREAD
pure collocate	ancient	TOWN
unrelated	classociatees	WALK

Appendix B: Training materials for recall study

(NB. target nouns and their paired adjectives are highlighted here in bold; they were not highlighted in the training materials seen by participants)

Set One

List One

sentence	**type**
You can buy a **cheap ball** from that shop on the corner.	target collocation
How big is the **average brain**?	target collocation
Hot chocolate is an **excellent drink** on a cold evening.	target collocation
They live on a **quiet farm** in Kent.	target collocation
We couldn't open the **huge gate.**	target collocation
This is the only **clean lake** in the area.	target collocation
He had a thick, **powerful neck**.	target collocation
The bells made a lovely **soft noise** in the distance.	target collocation
Extra buses were introduced on the **busy route** into the city.	target collocation
He wrote a **detailed text**, explaining everything fully.	target collocation
She worked on an emergency supply **boat**.	control
She was looking forward to getting home to her **flat**.	control
Wherever I go, I find that these maps are a very useful **guide**.	control
The afternoon **heat** made us sweat.	control
She found an old **leaf**.	control
She thought the **radio** was a waste of money.	control
There had been a **rise** in the cost of living.	control
He skied down the mountain at high **speed**.	control
He wanted to go on a **tour** of the Middle East.	control
He always found it difficult to choose a **wine**.	control

List Two

sentence	**type**
She worked on an emergency **medical boat**.	target collocation
She was looking forward to getting home to her **warm flat**	target collocation
Wherever I go, I find that these maps are a very **effective guide**.	target collocation
The **southern heat** made us sweat.	target collocation
She found a **beautiful leaf**.	target collocation
She thought the **expensive radio** was a waste of money.	target collocation
There had been an **obvious rise** in the cost of living.	target collocation
He skied down the mountain at a **dangerous speed**.	target collocation
He wanted to go on a **religious tour** of the Middle East.	target collocation
He always found it difficult to choose a **suitable wine**.	target collocation
You can buy a **ball** from that shop on the corner.	control
How big is the human **brain**?	control
Hot chocolate is a wonderful **drink** on a cold evening.	control
They live on a **farm** in Kent.	control
We couldn't open the **gate**.	control

This is the only **lake** in the area.	control
He had a thick, short **neck**.	control
The bells made a lovely **noise** in the distance.	control
Extra buses were introduced on the **route** into the city.	control
He wrote a long **text**, explaining everything fully.	control

Filler sentences

There are lots of different types of birds in the park.
The Earth is around four point five billion years old.
The police were keeping a file on his activities
There was a car in front of the house.
She is a good judge of character.
We had a lovely meal yesterday.
He used to have a model horse made of wood.
Every time he opens his mouth he says something stupid.
The dog escaped through a hole in the fence.
The shops were crowded during the January sale.
We went to America by ship.
I've got an old tape of her singing.
Crime in the city had increased sharply.
Have you seen this video?
That club has a five pound entry charge.
She lost her phone when she was on holiday.
There is a one hundred pound overdraft limit.
He always wears a gold watch.
I damaged a plate when I was washing up.
I'm going for a walk after dinner.

Set Two

List One

sentence	**type**
It was only a **cheap ball**, but he was sorry to lose it.	target collocation
The **average brain** must get about five to six hours sleep a night to work well.	target collocation
Lemon and honey is an **excellent drink** if you have a cold.	target collocation
He owns a house by a **quiet farm**.	target collocation
He pushed open the **huge gate** at the end of the path.	target collocation
There was a **clean lake** where the children could swim.	target collocation
They were impressed by the athlete's **powerful neck** and shoulders.	target collocation
The **soft noise** of the door opening downstairs woke her up.	target collocation
The **busy route** between Birmingham and London was closed for a week.	target collocation
She read the **detailed text** carefully	target collocation
They went to the island on a passenger **boat**.	control
He loved to sit in his **flat** and watch the snow outside.	control
This is a simple and practical **guide** to computer programming.	control
He loved the summer **heat**.	control
There was one **leaf** still on the tree.	control
I got her a **radio** for her birthday.	control
They had seen a huge **rise** in their standards of living.	control
The police stopped her for going at an illegal **speed**.	control
She went on a long **tour**, visiting all the churches in the region.	control
Can you recommend a **wine** to have with fish?	control

List Two

sentence	**type**
They went to the island on a **medical boat**	target collocation
He loved to sit in his **warm flat** and watch the snow outside.	target collocation
This is a simple and **effective guide** to computer programming.	target collocation
He loved the **southern heat**.	target collocation
There was one **beautiful leaf** still on the tree.	target collocation
I got her an **expensive radio** for her birthday.	target collocation
They had seen an **obvious rise** in their standards of living.	target collocation
The police stopped her for going at a **dangerous speed**.	target collocation
She went on a **religious tour**, visiting all the churches in the region.	target collocation
Can you recommend a **suitable wine** to have with fish?	target collocation
It was only an old **ball**, but he was sorry to lose it.	control
The **brain** must get about five to six hours sleep a night to work well.	control
Lemon and honey is a very soothing **drink** if you have a cold.	control
He owns a house by a **farm**.	control

He pushed open the **gate** at the end of the path.	control
There was a big **lake** where the children could swim.	control
They were impressed by the athlete's strong **neck** and shoulders.	control
The **noise** of the door opening downstairs woke her up.	control
The **route** between Birmingham and London was closed for a week.	control
She read the **text** carefully.	control

Filler sentences

She was very upset when her pet bird escaped from its cage.
Have you ever been a victim of crime?
Some people think that the sun orbits the earth.
Do you know the entry code for this door?
She kept a nail file in her handbag.
I don't like sitting at the front in the cinema.
He won a gold medal for running.
He had holes in his trouser pockets.
The judge sentenced him to two years in prison.
They put a limit on the number of people who could come.
They had a quick meal in an Italian restaurant.
After eating the sweets he had a red mouth.
Can you give me your phone number?
He ate an enormous plate of pasta.
The shops made a big profit in the sale.
They watched the ships coming into the harbour.
Can you send me that tape?
The video is about two hours long.
She always went for a long walk at the weekends.
We had a weekend’s camping in a wood.

Appendix C – detailed contents of the academic corpus

Table 1: Arts and Humanities Group

School	School word count	Division	Division word count	Journal	Journal word count	Number of articles
American and Canadian Studies	1,011,085	American and Canadian Studies	505,636	American Quarterly	101,349	12
				Journal of American Studies	105,417	11
				Reviews in American History	99,995	28
				American Literature	104,281	9
				Canadian Literature	94,594	16
		Film and Television Studies	505,449	Cinema Journal	101,136	14
				Film and History	101,183	13
				Film Quarterly	102,578	15
				Velvet Light Trap	100,251	10
				Wide Angle	100,301	22
English Studies	1,021,702	Medieval Studies	256,701	Chaucer Review	48,557	5
				Early Medieval Europe	46,570	3
				Essays in Medieval Studies	50,022	8
				Medieval and Early Modern Studies	55,749	4
				Journal of Medieval History	55,803	4
		Modern English Language	258,355	Applied Linguistics	56,721	5
				Language	51,489	3
				Pragmatics, Journal of	51,722	4
				Studies in Second Language Acquisition	50,625	3
				Text	47,798	5
		Modern English Literature	255,936	Children's Literature	49,396	4
				English Studies	53,793	8
				Language and Literature	50,201	6
				Journal of Modern Literature	51,561	8
				Studies in English Literature 1500-1900	50,985	7

		Drama	250,710	Drama Review	49,030	5
				Modern Drama	47,244	5
				New Theatre Quarterly	48,477	6
				Theatre Journal	54,004	5
				Theatre Research International	51,955	6
History	1,007,941	History	1,007,941	Past and Present	198,701	13
				Comparative Studies of Society and History	202,737	13
				History Workshop Journal	206,569	18
				Journal of Modern History	203,709	12
				Journal of Social History	196,225	18
Humanities	997,503	Archaeology	155,253	Journal of Anthropological Archaeology	28,398	4
				Antiquity	28,801	4
				International Journal of Historical Archaeology	31,788	4
				Oxford Journal of Archaeology	34,911	4
				World Archaeology	31,355	4
		Art History	168,087	Aesthetics and Art Criticism, Journal of	31,618	3
				Art Book	32,076	14
				Art History	33,958	2
				Oxford Art Journal	39,599	3
				Third Text	30,836	4
		Classics	170,783	Classical Antiquity	40,254	2
				Classical Quarterly	28,155	4
				Greece and Rome	34,882	4
				Hesperia	30,494	2
				Mnemosyne	36,998	4
		Music	164,294	19th Century Music	33,112	2
				Early Music	32,627	3
				Music Analysis	30,622	2

				Music Theory Spectrum	35,493	2
				Popular Music	32,440	3
		Philosophy	172,308	Continental Philosophy Review	37,197	3
				Erkenntnis	26,576	3
				Journal of The History of Philosophy	39,425	3
				Philosophy	33,963	3
				Synthese	35,147	4
		Theology	166,778	International Journal of Systematic Theology	30,111	3
				Literature and Theology	37,039	5
				Modern Theology	30,404	3
				Religion	35,980	3
				Scottish Journal of Theology	33,244	4
Modern Languages and Cultures	1,011,396	French and Francophone Studies	201,289	French Forum	42,225	6
				French Historical Studies	40,633	4
				Journal of French Language Studies	43,756	4
				French Studies	37,168	5
				Nineteenth-Century French Studies	37,507	6
		German Studies	201,258	Journal of Comparative German Linguistics	59,213	3
				German Historical institute Bulletin	70,196	10
				German Life and Letters	71,849	8
		Hispanic and Latin American Studies	200,221	Bulletin of Hispanic Studies	40,772	4
				Bulletin of Spanish Studies	40,910	4
				Hispanic American Historic Review	39,317	3
				Hispanic Review	36,428	4
				Journal of Latin American Studies	42,794	3
		Russian and Slavonic Studies	197,889	Russian Linguistics	62,583	8

				Russian Review	69,594	6
				Slavonic and East European Review	65,712	6
		Critical Theory and Cultural Studies	210,739	Critical inquiry	48,712	3
				Critical Social Policy	42,549	5
				Postmodern Culture	42,817	4
				Representations	41,979	3
				Theory, Culture and Society	34,682	3

Table 2: Life Sciences Group

School	School word count	Division	Division word count	Journal	Journal word count	Number of articles
Biomedical Sciences	719,052	Biochemistry and Molecular Biology	178,007	Biochemical and Biophysical Research Communications	36,325	7
				Cell	36,837	4
				Embo Journal	37,125	4
				Molecular and Cellular Biology	33,135	3
				Nucleic Acids Research	34,585	4
		Medicine (General and Internal)	179,033	Annals of Internal Medicine	34,715	4
				British Medical Journal	34,574	6
				Canadian Medical Association Journal	36,371	8
				Lancet	34,738	5
				Medical Journal of Australia	38,635	9
		Neuroscience	182,112	Brain Research	32,097	4
				Journal of Comparative Neurology	40,353	3
				Neuron	34,004	3
				Journal of Neurophysiology	38,117	5
				Journal of Neuroscience	37,541	4
		Nutrition and Dietetics	179,900	American Journal of Clinical Nutrition	39,431	6
				British Journal of Nutrition	35,038	4

				Food Chemistry	35,060	6
				International Journal of Obesity	35,062	6
				Journal of Nutrition	35,309	7
Biology	626,168	Biochemistry and Molecular Biology	155,837	Journal of Biological Chemistry	28,299	4
				Cell	36,948	4
				Embo J	30,958	3
				Nucleic Acids Research	28,047	6
				Molecular and Cellular Biology	31,585	3
		Biology	154,566	Biometrics	26,998	4
				Journal of Experimental Biology	32,215	4
				Faseb J	33,840	5
				Proceedings of the Royal Society B: Biological Sciences	27,067	5
				Journal of Theoretical Biology	34,446	5
		Genetics and Heredity	157,983	Gene	24,469	4
				Genes and Development	33,525	3
				Genetics	34,875	4
				Human Molecular Genetics	35,225	4
				Molecular Biology and Evolution	29,889	4
		Zoology	157,782	Animal behaviour	32,302	4
				Animal Ecology	34,091	4
				Behavioral Ecology and Sociobiology	31,745	4
				Comparative Biochemistry and Physiology, Part A	31,478	4
				Journal of Zoology	28,166	
Biosciences	622,913	Agriculture, Dairy and Animal Science	153,642	Animal Reproduction Science	28,516	3
				Journal of Animal Science	28,525	4
				Applied Animal Behaviour Science	31,134	6
				Journal of Dairy Science	32,698	6

				Poultry Science	32,769	6
		Environmental Sciences	155,461	Environmental Science and Technology	32,004	5
				Water Research	32,070	6
				Atmospheric Environment	29,844	4
				Chemosphere	29,567	4
				Environmental Health Perspectives	31,976	4
		Food Science and Technology	163,510	Journal of Agricultural and Food Chemistry	32,900	5
				Food Chemistry	33,501	5
				Journal of Food Science	31,428	4
				International Journal of Food Microbiology	33,348	5
				Journal of the Science of Food and Agriculture	32,333	5
		Plant Sciences	150,300	Plant Physiology	31,081	3
				Plant Cell	30,904	3
				Phytochemistry	29,076	4
				Plant Journal	27,787	2
				New Phytologist	31,452	3
Community Health Sciences	719,703	Psychiatry	357,397	American Journal of Psychiatry	64,582	12
				Archives of General Psychiatry	73,917	10
				Biological Psychiatry	74,366	10
				British Journal of Psychiatry	74,261	12
				Journal of Neurology, Neurosurgery and Psychiatry	70,271	13
		Public, Environmental and Occupational Health	362,306	American Journal of Epidemiology	74,573	9
				Cancer Epidemiology, Biomarkers and Prevention	74,142	10
				Journal of Clinical Epidemiology	68,921	11

				Environmental Health Perspectives	74,188	10
				Journal of Tropical Medicine and Hygiene, American	70,482	13
Human Development	668,070	Obstetrics and Gynaecology	668,070	American Journal of Obstetrics and Gynecology	119,908	25
				BJOG	132,190	29
				Fertility and Sterility	125,069	25
				Gynecologic Oncology	132,616	25
				Human Reproduction	158,287	25
Medical and Surgical Sciences	728,911	Allergy	181,954	Allergy	38,498	7
				Journal of Allergy and Clinical Immunology	34,401	5
				Clinical and Experimental Allergy	38,153	7
				Contact Dermatitis	34,194	9
				Pediatric Allergy Immunology	36,708	8
		Cardiac and Cardiovascular Systems	177,959	Ajp Heart and Circulatory	34,459	4
				Journal of the American College of Cardiology	33,109	6
				American Journal of Cardiology	37,558	9
				Circulation	32,783	5
				Circulation Research	40,050	6
		Dermatology	178,618	Journal of American Academy of Dermatology	36,971	9
				Archives of Dermatology	35,339	8
				British Journal of Dermatology	35,196	8
				Dermatologic Surgery	36,217	11
				Journal of investigative Dermatology	34,895	5
		Surgery	190,380	Annals of Surgery	34,665	7
				British Journal of Surgery	36,927	9

				Clinical Orthopaedics	36,435	9
				Journal of Neurology, Neurosurgery and Psychiatry	35,537	7
				Journal of Neurosurgery	46,816	6
Molecular and Medical Science	719,857	Immunology	363,772	European Journal of Immunology	72,849	11
				Journal of Experimental Medicine	75,157	8
				Immunity	74,071	8
				Journal of Immunology	74,139	9
				Infection and Immunity	67,556	9
		Infectious Diseases	356,085	Emerging Infectious Diseases	68,178	15
				Epidemiology and Infection	70,581	14
				Journal of Infectious Diseases	71,037	12
				Pediatric Infectious Disease Journal	71,617	11
				Sexually Transmitted Diseases	74,672	14
Pharmacy	627,366	Pharmacology and Pharmacy	627,366	Antimicrobial Agents and Chemotherapy	124,589	20
				Biochemical Pharmacology	125,298	17
				European Journal of Pharmacology	126,448	20
				Pharmacology and Experimental Therapeutics	126,313	17
				Psychopharmacology	124,718	18
Veterinary Medicine and Science	724,049	Veterinary Sciences	724,049	Theriogenology	142533	32
				Vaccine	148084	21
				Veterinary Microbiology	146188	27
				Veterinary Parasitology	142391	30
				Veterinary Record	144853	27

Table 3:Science & Engineering Group

School	School word count	Division	Division word count	Journal	Journal word count	Number of articles
Built Environment	718,993	Architecture	718,993	Journal of Architectural Engineering	143,788	21
				Architectural Research Quarterly	144,831	25
				Design Studies	141,191	16
				Structural Design of Tall and Special Buildings	141,887	22
				Journal of Urban Design	147,296	16
Chemical, Environmental and Mining Engineering	1,253,980	Chemical Engineering	625,907	Catalysis Today	127,620	28
				Journal of Chemical and Engineering Data	126,537	27
				Chemical Engineering Science	122,306	13
				industrial Engineering and Chemistry Research	123,395	21
				Journal of Catalysis	126,049	18
		Environmental Engineering	628,073	Ecological Engineering	125,141	18
				Environmental Science and Technology	125,825	19
				Journal of Hazardous Materials	124,749	22
				Waste Management	128,155	20
				Water Research	124,203	24
Chemistry	629,553	Analytical Chemistry	124,366	Analytical Chemistry	25,795	4
				Journal of Chromatography A	23,094	4
				Analytical Biochemistry	25,667	3
				Analytica Chimica Acta	24,654	5
				Electrophoresis	25,156	4
		Applied Chemistry	127,480	Advanced Synthesis and Catalysis	26,782	3
				Carbohydrate Polymers	24,814	5

				Carbohydrate Research	24,726	3
				Microporous and Mesoporous Materials	24,232	5
				Journal of Natural Products	26,926	5
		Inorganic and Nuclear Chemistry	124,368	Dalton Transactions	24,542	2
				inorganic Chemistry	27,087	4
				Inorganica Chimica Acta	23,697	4
				Journal of Organometallic Chemistry	24,677	3
				Organometallics	24,365	3
		Organic Chemistry	127290	Tetrahedron	26,522	3
				Journal of Organic Chemistry	24,228	3
				Tetrahedron: Asymmetry	25,848	4
				European Journal of Organic Chemistry	23,464	2
				Bioorganic and Medicinal Chemistry	27,228	3
		Physical Chemistry	126,049	Journal of Chemical Materials	25,732	4
				Langmuir	24,080	4
				Journal of Physical Chemistry A	23,745	3
				Journal of Physical Chemistry B	25,819	4
				Surface Science	26,673	4
Civil Engineering	1,255,027	Civil Engineering	1,255,027	Computers and Structures	252,891	36
				Earthquake Engineering and Structural Dynamics	250,201	30
				Journal of Hydrology	252,058	29
				Transportation Research Part B	250,817	27
				Wind Engineering and industrial Aerodynamics, Journal of	249,060	33
Computer Science and	620,713	Artificial Intelligence	90,613	Artificial Intelligence	18,490	1

Information Technology				IEEE Transactions On Image Processing	19,243	2
				IEEE Transactions On Neural Networks	17,550	2
				IEEE Transactions On Pattern Recognition and Machine Intelligence	16,698	2
				Pattern Recognition	18,632	2
		Cybernetics	87,832	International Journal of Human-Computer Studies	18,010	2
				Presence: Teleoperators and Virtual Environments	16,075	2
				Behaviour and Information Technology	23,076	1
				Interacting With Computers	16,223	2
				Kybernetes	14,448	2
		Hardware and Architecture	90,332	Journal of Computer and System Sciences	14,691	1
				Computer Networks	21,347	2
				Computer Standards and Interfaces	18,797	3
				IEEE Transactions On Computers	18,003	2
				IEEE/ACM Transactions On Networking	17,494	2
		Information Systems	89,595	IEEE Transactions On Information Theory	17,345	3
				American Society For Information Science and Technology	17,992	2
				IEEE Transactions On Knowledge and Data Engineering	20,511	2
				Information Sciences	16,568	2
				Information and Management	17,179	2

		Interdisciplinary Applications	87,042	Journal of Computational Physics	20,588	2
				Molecular Graphics and Modelling	16,433	2
				International Journal For Numerical Methods in Fluids	16,121	2
				Bioinformatics	15,684	3
				IEEE Transactions On Medical Imaging	18,216	2
		Software Engineering	89,246	Mathematical Programming	16,525	1
				Computer-Aided Design	18,161	3
				ACM Transactions On Graphics	17,801	2
				Image and Vision Computing	17,708	3
				IEEE Transactions On Software Engineering	19,051	1
		Theory and Methods	86,053	Fuzzy Sets and Systems	19,333	3
				Information and Computing	20,697	1
				Journal of Algorithms	16,846	2
				IEEE Transactions On Evolutionary Computation	17,130	1
				IEEE Transactions On Parallel and Distributed Systems	12,047	1
Electrical and Electronic Engineering	1,257,103	Electrical and Electronic Engineering	1,257,103	Automatica	252,929	28
				Image and Vision Computing	251,230	35
				Microelectronic Engineering	250,781	78
				Semiconductor Science and Technology	251,483	60
				Solid-State Electronics	250,680	62
Mathematical Sciences	626,655	Applied Mathematics	208,802	Physica D - Nonlinear Phenomena	40,505	7
				Mathematical Analysis and Applications	41,960	8

				Communications On Pure and Applied Mathematics	42,535	3
				Linear Algebra and Its Applications	42,017	9
				Mathematics of Computation	41,785	4
		Mathematics	208,887	Journal of Algebra	40,792	4
				Journal of Differential Equations	41,774	4
				Journal of Functional Analysis	39,019	3
				Discrete Mathematics	43,966	5
				Duke Mathematical Journal	43,336	3
		Statistics	208,966	Statistics in Medicine	45,578	5
				Annals of Statistics	42,234	7
				Journal of The Royal Statistical Society B: Statistical Methodology	40,386	4
				Chemometrics and Intelligent Laboratory Systems	38,536	7
				Annals of Probability	42,232	5
Mechanical, Materials and Manufacturing Engineering	1,254,449	Biomedical Engineering	418,693	Artificial Organs	86,948	19
				Biomaterials	83,642	13
				Journal of Biomechanics	81,653	13
				Pacing and Clinical Electrophysiology	86,172	20
				Physics in Medicine and Biology	80,278	10
		Mechanical Engineering	419,661	International Journal of Heat and Mass Transfer	82,701	10
				Journal of Aerosol Science	83,079	9
				Journal of Sound and Vibration	82,006	11
				Probabilistic Engineering Mechanics	82,103	14
				Proceedings of The Combustion institute	89,772	13

		Manufacturing Engineering	416,095	Assembly Automation	81,074	21
				Composites Part A: Applied Science and Manufacturing	84,110	16
				International Journal of Advanced Manufacturing Technology	84,727	17
				International Journal of Production Economics	84,881	11
				Journal of Manufacturing Science and Engineering	81,303	12
Physics and Astronomy	625,944	Astronomy and Astrophysics	102,524	Astrophysical Journal	21,358	2
				Monthly Notices of The Royal Astrophysical Society	24,581	2
				Astronomical Journal	14,147	2
				Icarus	23,746	2
				Solar	18,692	4
		Atomic, Molecular and Chemical	105,419	Journal of Chemical Physics	23,674	3
				Physical Review A - Atomic, Molecular and Optical Physics	18,244	3
				Physical Chemistry A - Spectroscopy, Kinetics, Environment and General Theory	21,191	2
				Physics B - Atomic, Molecular and Optical Physics	19,877	3
				Physical Chemistry Chemical Physics	22,433	5
		Condensed Matter	101,187	Physical Review B - Condensed Matter	17,439	2
				Thin Solid Films	22,934	6
				Journal of Physics - Condensed Matter	19,537	4
				Journal of Magnetism and Magnetic Materials	21,352	6

				Solid State Communications	19,925	7
		Fluids and Plasmas	107,932	Physical Review E - Statistical Physics, Plasmas, Fluids and Related Interdisciplinary Topics	22,210	4
				Journal of Fluid Mechanics	21,629	2
				Physics of Fluids	21,391	4
				Plasma Physics and Controlled Fusion	19,706	3
				Annual Review of Fluid Mechanics	22,996	2
		Nuclear Physics	109,962	Physical Review C - Nuclear	21,510	3
				Nuclear Physics A	20,720	3
				Journal of Physics G - Nuclear and Particle Physics	22,663	3
				Energy Conversion and Management	23,645	3
				Hyperfine Interactions	21,424	5
		Particles and Fields	98,920	Physical Review D - Particles and Fields	18,451	2
				Nuclear Physics B	21,377	2
				Journal of High Energy Physics	17,204	2
				Astroparticle Physics	20,526	2
				Cosmology and Astroparticle Physics	21,362	3

Table 4: Social-Administrative Group

School	School word count	Division	Division word count	Journal	Journal word count	Number of articles
Business	712,835	Business	362,529	Academy of Management Journal	70,606	7
				Administrative Science Quarterly	72,949	4
				Journal of Marketing	74,855	6
				Journal of Marketing Research	72,441	6

				Strategic Management Journal	71,678	6
		Business Finance	350,306	Journal of Finance	71,394	5
				Journal of Monetary Economics	68,338	6
				Accounting and Economics	68,412	5
				Journal of Accounting Research	70,684	4
				Journal of Banking and Finance	71,478	6
Economics	716,849	Economics	716,849	Econometrica	147,201	11
				Journal of Econometrics	139,916	12
				Journal of Financial Economics	143,085	8
				Journal of Political Economy	144,560	10
				Quarterly Journal of Economics	142,087	8
Law	711,649	Law	711,649	Columbia Law Review	149,021	6
				Harvard Law Review	143,541	10
				Journal of Law and Economics	137,203	12
				Michigan Law Review	139,187	6
				University of Chicago Law Review	142,697	12
Politics and International Relations	709,456	Political Science	353,357	American Journal of Political Science	70,039	6
				American Review of Political Science	66,461	5
				Journal of Politics	69,878	7
				Public Choice	72,723	8
				Public Opinion Quarterly	74,256	8
		International Relations	356,099	International Organization	71,906	5
				International Security	67,505	4
				Journal of Common Market Studies	74,742	7
				World Economy	69,300	10
				World Politics	72,646	5

Table 5: Social-Psychological Group

School	School word count	Division	Division word count	Journal	Journal word count	Number of articles
Education	712,352	Education and Educational Research	712,352	Journal of College Student Development	143,893	15
				Health Education Research	139,229	20
				Research in Science Teaching	140,612	11
				Journal of School Health	142,157	26
				Science Education	146,461	13
Nursing	722,363	Nursing	722,363	Journal of Advanced Nursing	145,914	21
				Cancer Nursing	141,443	21
				Heart and Lung	141,714	27
				Nursing Research	145,978	25
				Research in Nursing and Health	147,314	18
Psychology	630,188	Developmental Psychology	209,818	Child Development	40,758	5
				Developmental Psychology	40,197	3
				Journal of Autism and Developmental Disorders	42,222	5
				Abnormal Child Psychology	46,430	4
				Journal of Adolescent Health	40,211	8
		Experimental Psychology	208,478	Neuropsychologia	43,074	5
				Experimental Psychology Journal of Human Perception and Performance	41,142	3
				Journal of Cognitive Neuroscience	41,949	5
				Experimental Psychology Learning, Memory and Cognition	41,596	3
				Perception and Psychophysics	40,717	4
		Social Psychology	211,892	Personality and individual Differences	38,420	7
				Organizational Behavior and Human Decision Processes	41,265	3

				Journal of Personality	41,013	4
				Journal of Experimental Social Psychology	44,682	4
				Sex Roles	46,512	5
Sociology and Social Policy	713,523	Sociology	356,840	American Journal of Sociology	78,342	4
				Social Forces	73,879	8
				Annual Review of Sociology	67,136	6
				Social Problems	69,483	5
				Sociology of Health and Illness	68,000	6
		Social Work	356,683	Child Abuse and Neglect	70,918	11
				American Journal Community Psychology	68,209	9
				Journal of Community Psychology	71,750	8
				Children and Youth Services Review	72,439	8
				British Journal of Social Work	73,367	9

NB. all word counts were calculated using WordSmith Tools 3. Other software packages may return slightly different figures.

Appendix D: Key academic collocations

word1	word2	frequency/ million words
ABILITY	TO	97.64
ACCORDING	TO	267.36
ACCOUNT	FOR	56
ACCOUNTED	FOR	22.48
ACROSS	DIFFERENT	8.76
ADAPTED	FROM	6.36
ADDITIONAL	INFORMATION	7.56
ADDRESS	ISSUE	3.44
ADDRESS	QUESTION	3.72
ADHERENCE	TO	9.88
AFFECTED	BY	42.48
AFTER	INITIAL	9.96
AGE	GROUP	14.64
AGREES	WITH	7.92
ALIGNED	WITH	5.76
ALL	CASES	31.44
ALL	PARTICIPANTS	14.68
ALL	VARIABLES	18.24
ALLOW	US	8.08
ALLOWED	US	6.16
ALLOWS	TO	42.92
ALLOWS	US	13.32
ALSO	EVIDENT	3.64
ALSO	FOUND	23.88
ALSO	INDICATES	3.4
ALSO	SHOWED	8.64
ALSO	DEMONSTRATES	2.2
ALSO	INVESTIGATED	4.8
AMONG	GROUPS	13.4
AMONG	INDIVIDUALS	3.76
AMONG	VARIOUS	3.44
AMOUNT	OF	126.2
AMOUNT	INFORMATION	6.36
ANALOGOUS	TO	13.12
ANALYSIS	REVEALED	7.52
AND	RESPECTIVELY	249.68
ANY	GIVEN	15.04
APPEARS	BE	28.72
APPEARS	TO	56.96
APPLICABLE	TO	11.24
APPROPRIATE	FOR	25.52
ARE	COMMONLY	7.36
ARE	COMPARABLE	8.96

ARE	CONSISTENT	30.2
ARE	CORRELATED	11.28
ARE	DEPICTED	4.04
ARE	GENERALLY	19.44
ARE	HIGHLY	16.4
ARE	IDENTICAL	12.12
ARE	LIKELY	73.92
ARE	LISTED	17.2
ARE	LOCATED	11.36
ARE	MUTUALLY	3.48
ARE	PARENTHESES	8.28
ARE	PRESENTED	54.32
ARE	REPRESENTED	13.12
ARE	SENSITIVE	11.84
ARE	SHOWN	100.36
ARE	SUMMARIZED	12.32
ARE	TYPICALLY	11.96
ARE	VALID	6.04
ARGUE	THAT	38.16
ARGUED	THAT	36.28
ARGUING	THAT	15.56
ARISE	FROM	8.92
AS	CONSEQUENCE	20.48
AS	EVIDENCED	5.92
AS	FOLLOWS	81.2
AS	ILLUSTRATED	12.32
AS	MENTIONED	19.16
AS	NOTED	24.04
AS	SHOWN	126.16
ASCRIBED	TO	6.32
ASPECTS	OF	74.36
ASSIGNED	TO	33.2
ASSOCIATED	WITH	315.52
ASSOCIATION	BETWEEN	36.32
ASSUME	THAT	72.32
ASSUMED	BE	33.36
ASSUMED	THAT	31.64
ASSUMES	THAT	13.2
ASSUMING	THAT	20.76
ASSUMPTION	THAT	35.76
ASSUMPTIONS	ABOUT	7.28
AT	LEVEL	116.08
AT	POINTS	25.64
AT	SITE	29.96
AT	STAGES	10.36
ATTRIBUTABLE	TO	12.72
ATTRIBUTED	TO	44.12
BASED	ON	404.64

BASED	APPROACH	15.52
BASIS	FOR	37.24
BE	ADDRESSED	15.08
BE	APPLIED	23.28
BE	ATTRIBUTED	19.08
BE	CLASSIFIED	6
BE	CONSIDERED	74.56
BE	DETERMINED	27.68
BE	DISTINGUISHED	5.88
BE	EASILY	27.44
BE	EVALUATED	9.6
BE	EXPLAINED	31.28
BE	INFERRED	4
BE	INTERPRETED	17.88
BE	NOTED	27.12
BE	UNDERSTOOD	17.52
BE	VIEWED	13.88
BEEN	APPLIED	9.72
BEEN	ATTRIBUTED	3.48
BEEN	CHARACTERIZED	3.88
BEEN	CONDUCTED	7.2
BEEN	CONSIDERED	9.76
BEEN	DEMONSTRATED	14.88
BEEN	DESCRIBED	21.56
BEEN	DOCUMENTED	6.64
BEEN	EXTENSIVELY	8.64
BEEN	IDENTIFIED	13.64
BEEN	OBSERVED	16.92
BEEN	PREVIOUSLY	26.76
BEEN	PROPOSED	19.32
BEEN	SHOWN	51.56
BEEN	STUDIED	19.48
BEEN	SUGGESTED	13.36
BEEN	USED	48.2
BEEN	WIDELY	8.56
BEFORE	AFTER	37.68
BETTER	UNDERSTAND	6.48
BETTER	UNDERSTANDING	6.48
BETWEEN	AND	935.56
BETWEEN	GROUPS	45.84
BEYOND	SCOPE	5.72
BOTH	CASES	16.92
BOTH	GROUPS	23.36
BOTH	TYPES	9.28
BOTTOM	UP	5.08
BROAD	RANGE	6.64
BUT	ALSO	149.44
BUT	RATHER	32.76

BY	ADDING	13.08
BY	ANALYZING	4.08
BY	COMBINING	5.76
BY	COMPARING	15.92
BY	DIVIDING	7.76
BY	EXAMINING	8.88
BY	FOCUSING	5.52
BY	MEANS	56.44
CAN	APPLIED	11.92
CAN	ATTRIBUTED	8.68
CAN	BE	857.4
CAN	CONSIDERED	13.68
CAN	DIRECTLY	7.2
CAN	EASILY	23.28
CAN	EXPLAINED	15.08
CAN	EXTENDED	5.92
CAN	INTERPRETED	7.28
CAN	OBTAINED	24.36
CAN	OCCUR	9.96
CAN	POTENTIALLY	3.36
CAN	SEEN	51.92
CAN	USED	57.56
CANNOT	BE	87.04
CAPTURED	BY	7.6
CARRIED	OUT	94.68
CASE	STUDIES	15.6
CASE	STUDY	30.28
CASES	WHERE	12.96
CATEGORIZED	AS	6.2
CAUSED	BY	61.88
CHANGES	IN	156.96
CHARACTERIZED	BY	41.8
CLASSIFIED	AS	22.4
CLOSELY	RELATED	12.12
COGNITIVE	PROCESSES	5.28
COINCIDES	WITH	8.8
COLLECTED	FROM	25.84
COMBINED	WITH	29.8
COMMENTS	ON	15.08
COMMONLY	USED	13.92
COMPARABLE	TO	25.88
COMPARATIVE	ANALYSIS	6.4
COMPARATIVE	STUDIES	4.16
COMPARATIVE	STUDY	8.44
COMPARED	OTHER	13.96
COMPARED	THOSE	18.76
COMPARED	TO	152.04
COMPARED	WITH	165.88

COMPARISON	BETWEEN	21.44
COMPARISONS	BETWEEN	8.88
CONCERNS	ABOUT	9.12
CONCLUDE	THAT	31.8
CONCLUDED	THAT	26.48
CONCLUSIONS	DRAWN	4.44
CONFIRMED	BY	25.28
CONFIRMS	THAT	6.48
CONNECTED	TO	22.84
CONNECTION	BETWEEN	12.84
CONNECTIONS	BETWEEN	7.2
CONSISTED	OF	33
CONSISTENT	WITH	121.88
CONSISTING	OF	28.32
CONSISTS	OF	49.4
CONSTRAINTS	ON	10.12
CONSTRUCTED	BY	8.28
CONTRIBUTE	TO	48.44
CONTRIBUTED	TO	25.8
CONTRIBUTES	TO	16.44
CONTRIBUTING	TO	11.36
CONTROL	GROUP	44.2
CONTROL	GROUPS	13.28
CONVERTED	TO	13.08
CORRELATE	WITH	9.68
CORRELATED	WITH	45.88
CORRELATION	BETWEEN	48.48
CORRELATIONS	BETWEEN	16
CORRESPOND	TO	36.4
CORRESPONDS	TO	45.36
COULD	POTENTIALLY	4.6
CRITERIA	FOR	32.92
CRITICAL	ROLE	5.68
CROSS	SECTION	34.72
CURRENT	STUDY	27.8
DATA	AVAILABLE	24.56
DATA	COLLECTED	23
DATA	COLLECTION	28.72
DATA	SET	35.36
DATA	SUGGEST	13.88
DAYS	AFTER	44.24
DECISION	MAKING	37
DECREASE	IN	64.64
DEFINED	AS	100.76
DEFINED	BY	54.96
DEMONSTRATE	THAT	28.64
DEMONSTRATED	THAT	39.24
DEMONSTRATES	THAT	14.32

DEMONSTRATING	THAT	6.8
DENOTED	BY	16.36
DEPEND	ON	43
DEPENDENCE	ON	18.16
DEPENDING	ON	19.28
DEPENDS	ON	67.36
DEPICTION	OF	6.52
DERIVED	FROM	66.04
DESCRIBED	BY	46.04
DESCRIBED	HERE	5.48
DESCRIBED	ABOVE	25.28
DESCRIBED	DETAIL	5.08
DESCRIPTION	OF	55.84
DETAILED	ANALYSIS	6.64
DETERMINANTS	OF	21.2
DETERMINE	HOW	5.2
DETERMINE	IF	10.16
DETERMINE	WHETHER	23.8
DETERMINED	BY	84.96
DEVIATION	FROM	9.6
DEVIATIONS	FROM	8.84
DID	DIFFER	17.52
DIFFER	FROM	16.72
DIFFER	THEIR	4.6
DIFFERED	FROM	7.68
DIFFERENCE	BETWEEN	92.88
DIFFERENCES	AMONG	10.28
DIFFERENCES	BETWEEN	84.8
DIFFERENT	APPROACHES	5.28
DIFFERENT	GROUPS	18.44
DIFFERENT	LOCATIONS	3.76
DIFFERENT	METHODS	8.36
DIFFERENT	PATTERNS	6.56
DIFFERENT	SETS	4.52
DIFFERENT	STRATEGIES	4.4
DIFFERENT	TYPES	26.68
DIFFERS	FROM	12.68
DIRECT	EVIDENCE	4.52
DIRECTLY	FROM	13.56
DIRECTLY	RELATED	6
DISCREPANCY	BETWEEN	4.76
DISCUSSED	ABOVE	9.64
DISCUSSED	BELOW	6.88
DISTANCE	BETWEEN	25.48
DISTANCE	FROM	21.6
DISTINCT	FROM	13.16
DISTINCTION	BETWEEN	21.16
DISTINGUISH	BETWEEN	12.12

DIVIDED	INTO	25.36
DOES	DEPEND	7
DOES	NOT	274.16
DOES	REQUIRE	6.96
DRIVEN	BY	20.08
DUE	TO	374.12
DUE	FACT	9.56
DURING	FIRST	23.56
DURING	PERIOD	57.64
DURING	PERIODS	7.64
DURING	PHASE	15.48
EACH	CATEGORY	6.28
EACH	GROUP	29.84
EACH	INDIVIDUAL	14.76
EACH	PARTICIPANT	6.48
EARLY	PHASE	4
EFFECT	ON	184.36
EFFECTS	ON	112.92
EITHER	OR	128.92
EMBEDDED	IN	18.24
EMERGENCE	OF	26.44
EMPHASIZE	THAT	7.32
EMPIRICAL	EVIDENCE	8.56
ENGAGE	IN	24.72
ENROLLED	IN	9.28
ENTIRE	PERIOD	3.56
ESSENTIAL	ROLE	3.6
EVIDENCE	THAT	81.16
EVIDENCE	SUGGESTS	8.92
EVIDENCED	BY	7.64
EXAMINE	HOW	5.2
EXCLUDED	FROM	17.32
EXPLAINED	BY	32
EXPLAINED	FACT	3.2
EXPLANATION	FOR	22.52
EXPLANATIONS	FOR	9.2
EXPOSED	TO	44.4
EXPOSURE	TO	56.28
EXTENT	WHICH	24.88
EXTRACTED	FROM	19.28
FACILITATED	BY	3.96
FACTORS	SUCH	16.6
FACTORS	AFFECTING	6.24
FACTORS	INCLUDING	5.72
FIGURE	SHOWS	50.64
FINDINGS	SUGGEST	12.8
FOCUS	ON	73.84
FOCUSED	ON	48.6

FOCUSES	ON	23.32
FOCUSING	ON	21.56
FOLLOW	UP	90.12
FOLLOWED	BY	88.68
FOR	ASSISTANCE	19.76
FOR	DETERMINING	11.04
FOR	EXAMPLE	292.36
FOR	INSTANCE	78.04
FORMED	BY	16.48
FOUND	SIGNIFICANT	18.04
FRAMEWORK	FOR	19.6
FREQUENTLY	USED	7.24
FROM	PERSPECTIVE	28.16
FULL	SCALE	13.6
FURTHER	INVESTIGATION	6.84
FURTHER	RESEARCH	13.24
FUTURE	SHOULD	7.36
FUTURE	RESEARCH	19.48
GENERAL	POPULATION	14.16
GIVES	RISE	5.76
GREATER	THAN	68.16
GROUPED	INTO	4
HAD	EFFECT	24.64
HAS	ARGUED	10.24
HAS	DEMONSTRATED	17.68
HAS	DESCRIBED	13.8
HAS	DOCUMENTED	5.28
HAS	EXTENSIVELY	5.68
HAS	FOCUSED	7.36
HAS	IMPLICATIONS	7.04
HAS	INVESTIGATED	7.12
HAS	NOTED	7.32
HAS	OBSERVED	13.96
HAS	POTENTIAL	10.84
HAS	PROPOSED	12.04
HAS	PROVEN	4.44
HAS	RECOGNIZED	3.8
HAS	SHOWN	54.84
HAS	STUDIED	12.12
HAS	SUGGESTED	16.12
HAS	WIDELY	6.48
HAVE	ARGUED	10.88
HAVE	DEMONSTRATED	18.24
HAVE	DEVELOPED	20.44
HAVE	DOCUMENTED	4.68
HAVE	EXAMINED	10.96
HAVE	EXPLORED	4.16
HAVE	FOCUSED	8.12

HAVE	IDENTIFIED	15.2
HAVE	IMPACT	17.96
HAVE	INVESTIGATED	11.28
HAVE	PREVIOUSLY	17.64
HAVE	PROPOSED	16.6
HAVE	SHOWN	66.8
HAVE	STUDIED	13.84
HAVE	SUGGESTED	14.56
HEALTH	CARE	50.6
HIGH	DEGREE	8.96
HIGH	DENSITY	13.84
HIGH	FREQUENCY	18.2
HIGH	LEVELS	26.6
HIGH	LOW	8.8
HIGHER	THAN	98.56
HIGHER	DEGREE	3.92
HIGHER	LEVEL	13.04
HIGHER	LEVELS	27.68
HIGHER	ORDER	13.8
HIGHER	RATE	14.2
HIGHER	RATES	14.68
HIGHLY	SIGNIFICANT	5.8
HOWEVER	DOES	13.36
HOWEVER	THERE	33.72
HUMAN	DEVELOPMENT	8.96
HUMAN	NATURE	10.72
HYPOTHESIS	THAT	32.28
HYPOTHESIS	TESTING	3.68
HYPOTHESIZE	THAT	4.48
IDENTIFIED	AS	34.68
IDENTIFIED	BY	25.76
ILLUSTRATED	FIGURE	5.52
IMPACT	ON	75.2
IMPLICATED	IN	11.16
IMPLICATIONS	FOR	43.48
IMPLIES	THAT	46.36
IMPLY	THAT	15.16
IMPLYING	THAT	7.8
IMPORTANT	IMPLICATIONS	5.2
IMPORTANT	NOTE	11.12
IMPORTANT	ROLE	25.24
IN	ADDITION	204.72
IN	CONTRAST	126.76
IN	MANNER	52
IN	ORDER	262.88
IN	PARENTHESES	17.24
IN	TERMS	181.8
INCONSISTENT	WITH	6.8

INCORPORATED	INTO	12
INCREASE	IN	167.68
INDEBTED	TO	4.8
INDEXED	BY	3.24
INDICATE	THAT	71.8
INDICATED	BY	25.2
INDICATES	THAT	57.8
INDICATING	THAT	37.92
INDICATIVE	OF	11.36
INDIVIDUAL	DIFFERENCES	15.04
INDIVIDUALS	WHO	15.92
INFERRED	FROM	5.8
INFLUENCED	BY	31.48
INFLUENCES	ON	7.84
INFORMATION	PROCESSING	7.88
INFORMATION	REGARDING	6.4
INSENSITIVE	TO	4.88
INSIGHT	INTO	16.92
INSIGHTS	INTO	9.88
INSOFAR	AS	6.96
INTERACT	WITH	17
INTERACTED	WITH	3.12
INTERACTING	WITH	6.88
INTERACTION	BETWEEN	37.2
INTERACTIONS	AMONG	3.32
INTERACTIONS	BETWEEN	16.92
INTERACTS	WITH	7.2
INTERESTING	NOTE	6.84
INTERPLAY	BETWEEN	3.6
INTERPRETED	AS	20.08
INTERVIEWS	CONDUCTED	5.76
INTO	ACCOUNT	53.04
INTO	CATEGORIES	10
INTO	GROUPS	15.12
IS	APPARENT	13.92
IS	CHARACTERIZED	14.52
IS	CLEAR	59.2
IS	CLEARLY	30.2
IS	COMMONLY	10.84
IS	CONSISTENT	50.08
IS	CRUCIAL	15.04
IS	EVIDENT	22
IS	FEASIBLE	8.72
IS	ILLUSTRATED	15.32
IS	IMPORTANT	123.28
IS	INCONSISTENT	4.6
IS	INDICATIVE	3.64
IS	KNOWN	76.68

IS	LOCATED	16.52
IS	NECESSARY	46.36
IS	NOTEWORTHY	7.12
IS	POSSIBLE	104.4
IS	PROBLEMATIC	7.44
IS	PRONOUNCED	4.64
IS	REASONABLE	13.32
IS	STRAIGHTFORWARD	8.84
IS	SUFFICIENT	18.32
IS	SUFFICIENTLY	9.04
IS	UNCLEAR	12.8
ISOLATED	FROM	28.68
IT	APPEARS	22.68
IT	ASSUMED	19.64
IT	EVIDENT	10.48
IT	FOLLOWS	32.96
IT	NOTED	31.84
IT	NOTING	6.8
IT	POSSIBLE	101.28
IT	UNCLEAR	9.36
KNOWLEDGE	ABOUT	15.4
LACK	OF	100.52
LARGE	NUMBER	26.64
LARGER	THAN	44.56
LARGER	NUMBER	5.28
LAST	DECADES	6.32
LEAD	TO	94.12
LEADING	TO	53.48
LEADS	TO	77.44
LEAST	PARTIALLY	2.76
LESS	LIKELY	27.68
LESS	THAN	152.52
LINK	BETWEEN	17.52
LINKED	TO	37.96
LISTED	IN	26.96
LISTED	TABLE	15.84
LOCAL	GLOBAL	5.28
LOCATED	AT	13.68
LOCATED	WITHIN	4.16
LONG	TERM	84.36
LONGER	DURATION	3.48
LOW	DENSITY	10.44
LOW	FREQUENCY	13.08
LOW	LEVELS	18.24
LOWER	THAN	65.2
LOWER	LEVELS	13.56
MAIN	EFFECT	12.6
MAIN	EFFECTS	6.96

MALE	FEMALE	29.48
MALES	FEMALES	18.72
MAY	AFFECT	10.68
MAY	ALSO	49.88
MAY	DUE	17.88
MAY	EXPLAIN	10.52
MAY	OCCUR	9.68
MAY	PROVIDE	11.72
MAY	REFLECT	12.6
MAY	SERVE	4.76
MAY	USEFUL	7.6
MAY	LEAD	15.4
MEASURED	BY	53.92
MEDIATED	BY	22.8
MENTIONED	ABOVE	18.24
METHOD	USED	20.72
METHODS	USED	15.12
METROPOLITAN	AREAS	4.92
MIGHT	DUE	4.56
MIGHT	EXPLAIN	4.28
MODEL	FIT	12.12
MODIFIED	VERSION	3.04
MORE	ACCURATE	12.04
MORE	COMPLETE	5.8
MORE	COMPLEX	20.24
MORE	DETAILED	12.96
MORE	FREQUENT	8.28
MORE	FREQUENTLY	9.56
MORE	LIKELY	82.2
MORE	PRECISELY	8.84
MORE	PRONOUNCED	8.24
MORE	RECENTLY	13.76
MORE	SENSITIVE	9.44
MORE	SPECIFICALLY	11.88
MORE	STABLE	6.96
MOST	CASES	16.04
MOST	COMMON	22.56
MOST	COMMONLY	7.48
MOST	FREQUENT	5.44
MOST	FREQUENTLY	10.24
MOST	LIKELY	24.08
MOST	OFTEN	10
MOST	RELEVANT	4.36
MOTIVATED	BY	11.56
MUCH	HIGHER	13.88
MULTIPLIED	BY	6.48
MUTUALLY	EXCLUSIVE	3.88
NEGATIVE	EFFECT	11.68

NEGATIVE	EFFECTS	6.68
NEXT	SECTION	11.12
NO	DIFFERENCE	33.52
NO	SIGNIFICANT	61.84
NOT	AFFECT	22.44
NOT	ALTER	7.04
NOT	COMPLETELY	9.6
NOT	CORRESPOND	3.56
NOT	DEPEND	10.4
NOT	DIFFER	20.76
NOT	DIRECTLY	11.68
NOT	EXPLICITLY	5.96
NOT	OCCUR	9.28
NOT	REFLECT	8.32
NOT	REQUIRE	12.24
NOT	SPECIFIED	5.04
NOTE	THAT	134.56
NOTED	THAT	46.68
NOTED	ABOVE	5.96
NOTING	THAT	17.76
NOTION	THAT	18.48
NUMBER	OF	634.6
NUMEROUS	STUDIES	4.44
OBTAINED	FROM	109.08
OCCURRED	DURING	6.48
OCCURRENCE	OF	40.56
OCCURS	AFTER	3.2
OCCURS	AT	10.12
OCCURS	WHEN	9.44
ON	BASIS	77.08
ON	HAND	113.4
ON	SURFACE	56.84
ONE	DIMENSIONAL	14.52
ONE	EXPECT	11.68
OR	COMBINATION	11.04
ORDER	TO	244.08
ORDER	DETERMINE	6.64
ORDER	IDENTIFY	2.84
ORDER	INVESTIGATE	3.76
ORDER	UNDERSTAND	4.96
OTHER	FACTORS	25.88
OTHER	GROUPS	18.52
OTHER	HAND	84.88
OTHER	TYPES	11.64
OTHER	VARIABLES	14.84
OTHER	WORDS	42
OUR	ANALYSIS	27.52
OUR	DEFINITION	3.28

OUR	FINDINGS	27.8
OUR	KNOWLEDGE	15.56
OUR	RESULTS	72.6
OUR	STUDY	68.32
OUR	UNDERSTANDING	14.36
OVER	COURSE	9.4
OVER	ENTIRE	6.44
OVER	TIME	73.6
OVERLAP	BETWEEN	4.88
OVERVIEW	OF	17.88
OWING	TO	15.2
PARTIALLY	BY	8.4
PARTICIPANTS	WERE	46.24
PARTICIPATE	IN	28
PARTICIPATED	IN	19.8
PARTICIPATING	IN	13.12
PAST	DECADES	4.72
PAST	PRESENT	20.36
PAUCITY	OF	3.76
PERCEIVED	AS	15.72
PERCENTAGE	OF	88.08
PERCENTAGE	TOTAL	5.28
PERSPECTIVES	ON	11.44
PERTAINING	TO	4.36
PLAY	IMPORTANT	11.56
PLAY	ROLE	43.32
PLAYS	IMPORTANT	8.16
PLAYS	ROLE	28.4
POSITIVE	NEGATIVE	30.84
POSITIVE	VALUE	5.72
POSSIBILITY	THAT	30.96
POSSIBLE	EXPLANATION	6.12
PREDICTED	BY	16.4
PREDICTS	THAT	7.36
PREFERENCE	FOR	16.04
PRESENCE	ABSENCE	18.6
PRESENT	STUDY	76.28
PRESENTED	HERE	11.4
PRESENTED	ABOVE	2.68
PRESENTED	ARTICLE	2.88
PREVALENCE	OF	67.32
PREVIOUS	RESEARCH	12.48
PREVIOUS	SECTION	9.2
PREVIOUS	STUDIES	41.04
PREVIOUS	WORK	11.88
PRIMARILY	ON	7.08
PRIOR	TO	91.48
PROBABLY	DUE	6.52

PROBLEM	SOLVING	14.88
PRODUCED	BY	39.76
PROPORTION	OF	63.68
PROPOSED	BY	25.2
PROVIDE	EVIDENCE	13.8
PROVIDE	INFORMATION	16.32
PROVIDE	INSIGHT	4.52
PROVIDED	BY	49.04
PROVIDES	EVIDENCE	7.08
PROVIDES	INSIGHT	2.32
PUBLIC	POLICY	10.32
PUBLIC	PRIVATE	17.72
QUALITY	LIFE	38.76
QUANTITATIVE	ANALYSIS	9.24
QUESTIONS	REGARDING	2.76
QUITE	SIMILAR	4.68
RANGED	FROM	27.2
RANGES	FROM	8.24
RANGING	FROM	33.96
RATIONALE	FOR	5.52
RECEIVED	ATTENTION	8.04
RECENT	RESEARCH	8.28
RECENT	STUDIES	14.24
RECENT	STUDY	12.6
RECENT	WORK	7.36
REFER	TO	44.72
REFERRED	AS	27.64
REFERS	TO	37.16
REGARD	TO	37.32
REGARDLESS	OF	37.8
REGARDLESS	WHETHER	4.72
REGULATED	BY	10.12
RELATED	TO	190.72
RELATION	BETWEEN	30.04
RELATIONSHIP	BETWEEN	115.32
RELATIONSHIPS	AMONG	6.56
RELATIONSHIPS	BETWEEN	24.84
RELATIVELY	HIGH	12.56
RELATIVELY	LOW	12.6
RELATIVELY	SMALL	14.24
RELATIVELY	STABLE	3.76
RELIES	ON	14.92
REMAINED	STABLE	2.72
REMAINS	UNCLEAR	3.28
REPEATED	MEASURES	10.28
REPRESENTED	BY	29.64
REQUIRED	TO	65.84
RESEARCH	FOCUSED	3.56

RESEARCH	SUGGESTS	6.64
RESEARCHERS	HAVE	14.56
RESPECT	TO	107.16
RESTRICTED	TO	21.08
RESULTED	IN	55.68
RESULTING	FROM	24.64
RESULTS	INDICATE	25.28
RESULTS	OBTAINED	37.68
RESULTS	SHOW	28.36
REVEALS	THAT	14
REVIEW	LITERATURE	11.84
RIGHT	SIDE	20.24
SAMPLE	SIZE	27.76
SATISFACTION	WITH	7.44
SCOPE	THIS	7.16
SEE	ALSO	75.36
SEE	APPENDIX	10.76
SEE	DISCUSSION	10.36
SEE	FIGURE	23.88
SEE	TABLE	38.4
SELECTED	BECAUSE	2.96
SELECTED	FROM	15.36
SENSITIVE	TO	39.96
SERVE	AS	24.08
SERVED	AS	18.36
SERVES	AS	12.8
SEVERAL	AUTHORS	3.52
SEVERAL	STUDIES	16.68
SHED	ON	6.52
SHED	LIGHT	6.48
SHORT	LONG	12.16
SHOULD	CONSIDERED	10.48
SHOULD	NOTED	19.24
SHOW	THAT	127.84
SHOWED	NO	12.68
SHOWED	THAT	76.6
SHOWN	FIGURE	42.32
SHOWN	TABLE	39.56
SHOWS	THAT	86.6
SIDE	EFFECTS	16.48
SIGNIFICANT	BETWEEN	46.56
SIGNIFICANT	DIFFERENCE	37.96
SIGNIFICANT	DIFFERENCES	46.04
SIGNIFICANT	EFFECT	26.48
SIGNIFICANT	EFFECTS	12.16
SIGNIFICANT	IMPACT	4.76
SIGNIFICANT	INTERACTION	9.92
SIGNIFICANTLY	FROM	22.8

SIGNIFICANTLY	MORE	21.2
SIGNIFICANTLY	THAN	37.92
SIGNIFICANTLY	DIFFERENT	29.4
SIGNIFICANTLY	HIGHER	31.16
SIGNIFICANTLY	LOWER	19.08
SIMILAR	THOSE	25.84
SIMILAR	FINDINGS	5.04
SIMILAR	PATTERN	7.4
SIMILARITY	BETWEEN	5.72
SIMPLE	MODEL	9.2
SLIGHTLY	THAN	14.52
SMALL	NUMBER	20.4
SMALL	SIZE	11.88
SMALLER	THAN	34.24
SOME	EVIDENCE	11.04
SOME	RESEARCHERS	3.88
SPECULATE	THAT	5.64
STARTING	FROM	10.76
STATISTICALLY	SIGNIFICANT	57.84
STRONG	EVIDENCE	5.12
STRONGLY	ASSOCIATED	3.8
SUBJECTED	TO	42.44
SUBSET	OF	33.28
SUCH	AS	496.52
SUGGEST	THAT	129.32
SUGGESTED	BY	18.88
SUGGESTED	THAT	57.04
SUGGESTING	THAT	51.6
SUGGESTS	THAT	123.88
SUMMARIZED	IN	17
SUMMARIZED	TABLE	10.52
SUPPLEMENTED	WITH	13.8
SUPPOSE	THAT	29.64
SUSCEPTIBLE	TO	14.04
TABLE	PRESENTS	12.76
TABLE	SHOWS	34.64
TABLE	SUMMARY	9.08
TAKEN	ACCOUNT	16.32
TAKEN	TOGETHER	9.96
TAKING	INTO	17.92
TAKING	ACCOUNT	16.52
THAN	THOSE	63.28
THEIR	ABILITY	14.72
THEIR	COMMENTS	7.92
THEIR	COUNTERPARTS	13.16
THERE	DIFFERENCES	28
THERE	EVIDENCE	33.76
THERE	EXIST	14.32

THERE	SIGNIFICANT	52.96
THESE	ANALYSES	10.08
THESE	APPROACHES	7.32
THESE	ARE	237.72
THESE	ASSUMPTIONS	5.68
THESE	AUTHORS	7.68
THESE	CASES	20.24
THESE	CONDITIONS	20.96
THESE	CORRESPOND	2.64
THESE	CRITERIA	5.28
THESE	DEMONSTRATE	6.8
THESE	DIFFERENCES	22.8
THESE	ESTIMATES	6.8
THESE	FACTORS	23.04
THESE	FINDINGS	38.32
THESE	INDICATE	17.88
THESE	ISSUES	13.16
THESE	LINES	9.36
THESE	MEASURES	9.08
THESE	MODELS	15.52
THESE	PHENOMENA	3.52
THESE	REPRESENT	6.32
THESE	RESULTS	91.88
THESE	SUGGEST	33.24
THESE	VARIABLES	18.24
THIS	APPROACH	49.28
THIS	ARTICLE	69.64
THIS	ASSUMPTION	15.56
THIS	CASE	107.64
THIS	DIFFERS	2.84
THIS	FINDING	26.96
THIS	HYPOTHESIS	16.96
THIS	ILLUSTRATES	5.8
THIS	IMPLIES	21.28
THIS	INDICATES	19.16
THIS	ISSUE	27.12
THIS	LEADS	13.56
THIS	PAPER	163.68
THIS	PHENOMENON	14
THIS	PROCEDURE	16.72
THIS	REGARD	10.24
THIS	REQUIRES	9.08
THIS	STUDY	296.96
THIS	SUGGESTS	40.64
THIS	TECHNIQUE	14.76
THOSE	PREVIOUS	3.76
THUS	FAR	6.04
THUS	APPEARS	3.2

TO	ACCOMPLISH	5.04
TO	ANALYZE	19.2
TO	ASSESS	63.48
TO	ASSIGN	7
TO	CALCULATE	25.96
TO	CHARACTERIZE	16.56
TO	CLARIFY	11.24
TO	DELINEATE	1.76
TO	DETECT	33.68
TO	DETERMINE	121.08
TO	DIFFERENTIATE	6.16
TO	DISTINGUISH	18.48
TO	ELICIT	5.56
TO	ELIMINATE	13.24
TO	ENGAGE	18.76
TO	ENHANCE	18.24
TO	EVALUATE	56.4
TO	EXAMINE	54.36
TO	EXPLORE	30.84
TO	EXTENT	81.48
TO	FACILITATE	17.28
TO	GENERATE	30.64
TO	IDENTIFY	72.36
TO	ILLUSTRATE	15.56
TO	INCORPORATE	10.48
TO	INDUCE	17.36
TO	INFER	3.72
TO	INTERPRET	13.88
TO	INVESTIGATE	50.44
TO	MAXIMIZE	8.84
TO	MINIMIZE	19.32
TO	OBTAIN	73
TO	PARTICIPATE	26.28
TO	PERFORM	27.12
TO	QUANTIFY	12.68
TO	RECOGNIZE	15.56
TO	REDUCE	61.56
TO	REGULATE	9.08
TO	REPLICATE	3.96
TO	SIMULATE	11.28
TO	SOLVE	18.96
TO	UNDERSTAND	62.8
TO	VALIDATE	7.6
TO	VERIFY	14.96
TOP	BOTTOM	13.6
TOTAL	NUMBER	38.92
TOWARD	END	3.44
TRADE	OFF	8.96

TRANSFORMED	INTO	11.36
TRANSITION	BETWEEN	4.08
TRANSITION	FROM	12.76
UNCLEAR	WHETHER	3.56
UNDER	CONDITIONS	83.92
UNDERSTANDING	HOW	6.76
UNLESS	OTHERWISE	8.04
UPPER	LOWER	15.72
URBAN	RURAL	10.72
USEFUL	FOR	25
USING	SAME	15.6
USING	TECHNIQUES	7.36
VARIOUS	TYPES	8
VERTICAL	AXIS	7.72
VERY	LOW	22.12
VERY	SIMILAR	18.84
VIEWED	AS	24.4
WAS	ACHIEVED	13.48
WAS	CALCULATED	36.36
WAS	CARRIED	31.16
WAS	CHOSEN	14.96
WAS	CONDUCTED	29
WAS	CONFIRMED	17.92
WAS	DETERMINED	53.84
WAS	INITIATED	6.04
WAS	MEASURED	53.32
WAS	PERFORMED	84.8
WAS	USED	184.64
WE	ACKNOWLEDGE	5.44
WE	APPLY	11.24
WE	ASSUME	44.52
WE	CHOSE	7.28
WE	COMPARE	10.6
WE	CONCLUDE	22.88
WE	CONSIDER	49.72
WE	DISCUSS	14.64
WE	EXAMINE	17.96
WE	EXAMINED	22.6
WE	EXPECT	25.52
WE	EXPLORE	6.04
WE	FIND	61.24
WE	IDENTIFY	8.92
WE	INTERPRET	3.92
WE	OBSERVE	19.4
WE	PROPOSE	17.68
WE	REFER	10.64
WE	SPECULATE	3.12
WEEKS	AFTER	23.52

WELL	BEING	24.68
WELL	DEFINED	11.64
WELL	DOCUMENTED	6.48
WERE	ASKED	27.64
WERE	ASSIGNED	12.88
WERE	CARRIED	27.12
WERE	CHOSEN	11.68
WERE	COLLECTED	45.64
WERE	CONDUCTED	27.96
WERE	CONSIDERED	27.16
WERE	CONSTRUCTED	7.04
WERE	DETERMINED	37.36
WERE	DIVIDED	7.28
WERE	EXCLUDED	22.32
WERE	IDENTICAL	7.68
WERE	IDENTIFIED	34.76
WERE	PERFORMED	63.12
WERE	PREPARED	26.08
WERE	RECORDED	28.2
WERE	REMOVED	14.28
WERE	SELECTED	24.28
WERE	SIGNIFICANTLY	40.8
WERE	SUBSEQUENTLY	8.24
WERE	USED	132.04
WHAT	EXTENT	8.08
WHEN	COMPARED	22.44
WHEN	COMPARING	5.32
WHICH	CORRESPONDS	8.32
WHICH	INDICATES	8.12
WHICH	OCCURS	6.48
WHICH	TURN	16.08
WHO	PARTICIPATED	6.12
WIDELY	USED	18.2
WITH	EXCEPTION	17.84
WITH	MODIFICATIONS	4.48
WITH	REGARD	26.8
WITH	RESPECT	101.48
WITH	VARYING	10.04
WITHIN	CONTEXT	10.92
WITHIN	RANGE	12.8
WITHOUT	ANY	23.64
WORTH	NOTING	7.44
YEAR	PERIOD	13.6
YEARS	AGE	55.8
YOUNG	ADULTS	7.92

Appendix E: Collocations of academic keywords

keyword	collocate	frequency/ million words
ABSENCE	OF	106.4
ABSENCE	OR	16.96
ABSENCE	ANY	3.04
ABSENCE	EVEN	2.56
ACCORDING	TO	279.96
ACCORDING	CRITERIA	2.96
ACTIVE	MORE	7
ACTIVE	PASSIVE	2.96
ACTIVITY	DURING	7
ADDITION	IN	214.72
ADDITION	TO	86.88
ANALYSES	DATA	7.12
ANALYSIS	OUR	26.12
ANALYSIS	FURTHER	7.64
ANALYSIS	DATA	54.68
ANALYSIS	DETAILED	6.64
ANALYSIS	COMPARATIVE	6.36
ANALYSIS	USING	31.76
ANALYSIS	QUANTITATIVE	9
ANALYSIS	STRUCTURAL	7.2
ANALYSIS	PROVIDES	2.6
ANALYSIS	REVEALED	7.28
ANALYSIS	STATISTICAL	37.88
ANALYSIS	CARRIED	5.64
ANALYSIS	INCLUDED	9.32
APPEARS	TO	60.36
APPEARS	IT	22.36
APPEARS	THAT	22.96
APPEARS	BE	28.52
APPEARS	HAVE	8.16
APPEARS	THERE	3.88
ASSOCIATED	WITH	320.32
ASSOCIATED	ARE	33.84
ASSOCIATED	CLOSELY	3.24
ASSOCIATED	PROBLEMS	4.88
BASED	ON	408.72
BASED	UPON	12.4
BASED	APPROACH	15.44
BASED	ASSUMPTION	5.48
BASED	THEORY	6.24
BETWEEN	AND	1012.08
BETWEEN	DISTINCTION	21.16
BETWEEN	CONNECTION	12.84
BETWEEN	LINK	17.52

BETWEEN	RELATIONS	13.08
BETWEEN	RELATIONSHIPS	24.92
BETWEEN	GAP	12.64
BETWEEN	CONNECTIONS	7.2
BETWEEN	DISTANCE	25.32
BETWEEN	BOUNDARY	5.48
BETWEEN	DISTINGUISH	12.12
BETWEEN	LINKS	8.32
BETWEEN	SIMILARITIES	5.08
BETWEEN	CORRELATION	48.28
BETWEEN	ASSOCIATION	36.2
BETWEEN	BALANCE	8.56
BETWEEN	CLOSE	5.12
BETWEEN	INTERPLAY	3.6
BETWEEN	SIMILARITY	5.72
BETWEEN	BRIDGE	3.56
BETWEEN	COMMUNICATION	5.32
BETWEEN	INTERACTIONS	16.84
BETWEEN	OVERLAP	4.92
BETWEEN	STRONG	9.08
BETWEEN	ACTUAL	4.32
BETWEEN	COMPARISONS	8.76
BETWEEN	EXIST	3.4
BETWEEN	EXISTS	4.36
BETWEEN	AGREEMENT	11.04
BETWEEN	COMPETITION	4.72
BETWEEN	DISTINGUISHING	2.8
BETWEEN	DISCREPANCY	4.76
BETWEEN	ASSOCIATIONS	11.88
BETWEEN	CORRELATIONS	16.04
BETWEEN	INTERMEDIATE	3.24
BETWEEN	TRADE	6.6
BETWEEN	REVEAL	1.64
BETWEEN	REVEALED	3.44
BETWEEN	PERIODS	2.44
BETWEEN	TRANSITION	4.04
BETWEEN	LINKAGE	2.92
BOTH	AND	558.56
BOTH	SIDES	12.16
BOTH	TERMS	6.16
BOTH	SETS	2.28
BOTH	APPROACHES	2.28
BOTH	DIRECTIONS	2.76
BOTH	SIMULTANEOUSLY	1.8
CAPACITY	THEIR	5.28
CAPACITY	HAVE	4.28
CAPACITY	ITS	3.64
CASES	IN	179.76

CASES	SOME	23.68
CASES	ALL	30.48
CASES	MANY	10.8
CASES	MOST	15.68
CASES	BOTH	16.52
CASES	SUCH	9.56
CASES	WHERE	12.92
CASES	THERE	7.8
CASES	FEW	3.6
CASES	SEVERAL	2.56
CASES	MAJORITY	2.56
CHARACTERISTICS	SUCH	6.24
CHARACTERIZED	BY	42.12
CHARACTERIZED	IS	14.36
CHARACTERIZED	BEEN	3.88
COLLECTED	WERE	46.16
COLLECTED	FROM	26.28
COLLECTED	DATA	23.24
COMMONLY	IS	10.84
COMMONLY	ARE	7.44
COMMONLY	MOST	7.48
COMMONLY	USED	13.92
COMPARE	TO	44.96
COMPARE	WE	10.52
COMPARE	WITH	12.36
COMPARISON	WITH	38.16
COMPARISON	BETWEEN	21.2
CONSISTENT	WITH	123.48
CONSISTENT	IS	49.64
CONSISTENT	ARE	29.96
CONSISTENT	THIS	26.72
CONSISTENT	RESULTS	13.52
CONTENT	ITS	5.56
CONTRAST	IN	135.12
DEFINED	AS	101.24
DEFINED	IS	64.56
DEFINED	BY	55.04
DEFINED	WELL	11.72
DEFINED	CLEARLY	3.72
DEGREE	SOME	8.2
DEGREE	WHICH	12
DEGREE	HIGH	9.04
DEGREE	GREATER	2.56
DEGREE	HIGHER	3.88
DEGREE	CERTAIN	2.44
DEMONSTRATE	TO	24.4
DEMONSTRATE	THAT	29.2
DERIVED	FROM	66.96

DERIVED	ARE	10.76
DETERMINE	TO	125.92
DETERMINE	WHETHER	23.8
DETERMINE	IT	6.6
DETERMINE	WHAT	3.32
DETERMINE	WAS	16.92
DETERMINE	WHICH	7.36
DETERMINE	HOW	5.24
DETERMINE	IF	10.08
DETERMINE	DIFFICULT	2.36
DETERMINE	ORDER	6.6
DETERMINE	USED	13.44
DIFFERENCE	BETWEEN	92.72
DIFFERENCE	THIS	21.64
DIFFERENCE	THERE	23.4
DIFFERENCE	NO	33.32
DIFFERENCE	ONLY	6.6
DIFFERENCE	LITTLE	2.16
DIFFERENCE	MAKE	4.2
DIFFERENCE	MAIN	2.08
DIFFERENCE	STATISTICALLY	7.84
DIFFERENCES	BETWEEN	84.4
DIFFERENCES	THERE	27.76
DIFFERENCES	NO	34.04
DIFFERENCES	SOME	6.92
DIFFERENCES	DESPITE	2.76
DIFFERENCES	AMONG	10
DIFFERENCES	INDIVIDUAL	15.12
DIFFERENCES	SMALL	3.64
DIFFERENCES	SIMILARITIES	3.08
DIFFERENT	FROM	103.28
DIFFERENT	ARE	66.2
DIFFERENT	BETWEEN	27
DIFFERENT	MANY	8.52
DIFFERENT	SEVERAL	6.52
DIFFERENT	ACROSS	8.76
DIFFERENT	VERY	17.08
DIFFERENT	WAYS	11.36
DIFFERENT	QUITE	8.96
DIFFERENT	FORMS	6.28
DIFFERENT	TIMES	9.64
DIFFERENT	SAME	11.16
DIFFERENT	SLIGHTLY	5.48
DIFFERENT	PARTS	4.72
DIFFERENT	SOMEWHAT	3.24
DIFFERENT	AMONG	10.24
DIFFERENT	POINTS	5.92
DIFFERENT	COMPLETELY	3.36

DIFFERENT	LEVELS	16.84
DIFFERENT	PERIODS	3.08
DIFFERENT	ASPECTS	2.88
DIFFERENT	COUNTRIES	4.8
DIFFERENT	AREAS	4.76
DIFFERENT	CATEGORIES	3.12
DIFFERENT	CLASSES	2.92
DIFFERENT	INDIVIDUALS	3.4
DIFFERENT	LOCATIONS	3.76
DIFFERENT	SETS	4.56
DIFFERENT	STRATEGIES	4.48
DIFFERENT	CONDITIONS	12.44
DIFFERENT	ROLES	2.24
DIFFERENT	METHODS	8.16
DIFFERENT	VARIETY	1.88
DIFFERENT	STRUCTURES	3.96
DIFFERENT	APPROACHES	5.32
DIFFERENT	COMPARE	2.24
DIFFERENT	REGIONS	5.64
DIFFERENT	REPRESENTING	1.48
DIFFERENT	VARY	1.96
DIRECT	BETWEEN	7.52
DIRECT	INDIRECT	4.8
DIRECT	EVIDENCE	4.56
DIRECT	CONTACT	3.56
DIRECTLY	FROM	13.44
DIRECTLY	OR	7.84
DIRECTLY	INTO	4.28
DIRECTLY	CAN	7.12
DIRECTLY	THROUGH	1.8
DIRECTLY	EITHER	2.64
DIRECTLY	NOT	11.68
DIRECTLY	INDIRECTLY	3.4
DIRECTLY	LINKED	2.12
EFFECTS	ON	114.28
EFFECTS	HAVE	22.08
EFFECTS	SIDE	16.6
EFFECTS	LONG	6.36
EFFECTS	MAIN	6.96
EFFECTS	SHORT	4.48
FACTORS	OTHER	25.28
FACTORS	SUCH	16.36
FACTORS	MAY	9.96
FACTORS	IMPORTANT	330.28
FACTORS	INCLUDING	5.72
FACTORS	AFFECTING	6.24
FIGURE	SEE	22.84
FIGURE	LEFT	3.96

FIGURE	ILLUSTRATED	5.52
FUNCTION	AS	101.16
FUNCTION	PROBABILITY	7.44
FURTHERMORE	THERE	2.8
GROUPS	BETWEEN	45.8
GROUPS	OTHER	17.84
GROUPS	WERE	37.84
GROUPS	BOTH	22.72
GROUPS	ALL	17.56
GROUPS	INTO	15.16
GROUPS	ACROSS	8.64
GROUPS	DIFFERENT	18.28
GROUPS	DIFFERENCES	12.68
GROUPS	AMONG	13.08
GROUPS	CONTROL	13.32
GROUPS	DIVIDED	4.12
HIGHLY	IS	27.12
HIGHLY	ARE	16.52
HOWEVER	THIS	61.44
HOWEVER	THERE	33.04
HOWEVER	DOES	12.76
HOWEVER	DID	11.2
HOWEVER	CANNOT	4.04
HOWEVER	DESPITE	3.04
HOWEVER	NEITHER	2.04
HOWEVER	NOT	55.4
HOWEVER	IMPORTANT	6.72
HOWEVER	SEEMS	3.12
HOWEVER	CLEAR	4
HOWEVER	SUGGEST	3.32
HOWEVER	UNLIKE	2.96
HOWEVER	RECENT	4.08
HOWEVER	REMAINS	2.92
HOWEVER	DIFFERENCES	5.28
IDENTIFIED	AS	35.04
IDENTIFIED	BY	25.84
IDENTIFIED	HAVE	15.12
IDENTIFIED	BEEN	13.56
IDENTIFIED	HAS	7
IDENTIFIED	WERE	35.08
IDENTIFIED	CLEARLY	1.36
IDENTIFY	TO	75.92
IDENTIFY	CAN	5.48
IDENTIFY	WE	8.8
IDENTIFY	WHICH	4.16
IDENTIFY	ABLE	2.68
IDENTIFY	ORDER	2.88
IDENTIFY	POSSIBLE	2.52

IMPACT	ON	76.04
IMPACT	HAD	8.36
IMPACT	HAVE	17.84
IMPACT	ITS	6.52
IMPACT	HAS	8.16
IMPACT	LITTLE	1.6
INDICATE	THAT	72.68
INDICATE	THEY	4.12
INDICATE	RESULTS	24.96
INDICATES	THAT	58.72
INDICATES	THIS	19
INDICATES	WHICH	8.2
INFLUENCE	ON	49.76
INFLUENCE	UNDER	4.52
INFLUENCE	HAD	4.72
INFLUENCE	MAY	9.4
INITIAL	AFTER	9.88
INITIALLY	WAS	7.28
INITIALLY	WERE	5.64
INTERACTION	BETWEEN	37.08
INTERACTION	THROUGH	2.92
INTERNAL	EXTERNAL	9.16
LARGER	THAN	44.8
LARGER	MUCH	9.72
LARGER	NUMBER	5.28
LIMITED	THEIR	5
LIMITED	ONLY	8.16
LIMITED	BECAUSE	3.2
LIMITED	VERY	4.64
LIMITED	RANGE	2.08
LIMITED	INFORMATION	3.28
LITERATURE	REVIEW	11.96
LOCATED	IN	30.88
LOCATED	IS	16.6
LOCATED	ARE	11.44
LOCATED	AT	13.84
LOCATED	WERE	6.8
LOCATION	ITS	3.24
MODEL	OUR	35.28
MODEL	FIT	11.96
MODEL	PROVIDES	4
MODEL	BASED	24.24
MODEL	SIMPLE	9.36
MODEL	PROPOSED	12.24
MODELS	DIFFERENT	6.92
MODELS	USED	10.28
MULTIPLE	SINGLE	4.64
MULTIPLE	INCLUDING	2.68

NEGATIVE	EFFECTS	6.68
NEGATIVE	EFFECT	11.36
NOTED	THAT	46.64
NOTED	AS	24.08
NOTED	IT	32
NOTED	BE	26.96
NOTED	HAS	7.28
NOTED	SHOULD	19.28
NOTED	ABOVE	5.96
NOTED	ALSO	7.28
NOTED	HOWEVER	4.16
OBSERVATIONS	ARE	10.2
OBSERVATIONS	FROM	9.44
OBSERVED	HAS	13.52
OBSERVED	BEEN	16.44
OCCUR	THAT	16.28
OCCUR	CAN	9.8
OCCUR	WHICH	5.6
OCCUR	DOES	4.4
OCCUR	THEY	2.96
OCCUR	DID	3.8
OCCUR	WILL	5.2
OCCUR	MAY	9.52
OCCUR	WHEN	5.56
OCCUR	WHERE	1.96
OCCUR	WOULD	3.84
OCCUR	IF	2.92
OCCUR	BECAUSE	2.44
OCCUR	COULD	2.44
OCCUR	DURING	3.08
OCCUR	MIGHT	1.96
OCCUR	NOT	9.12
OCCUR	LIKELY	3.36
OCCURS	THAT	14.76
OCCURS	WHICH	6.52
OCCURS	WHEN	9.4
OCCURS	AT	9.92
OCCURS	AFTER	3.12
OCCURS	BEFORE	1.36
PATTERN	THIS	12.68
PATTERN	SAME	3.8
PATTERNS	DIFFERENT	6.56
PATTERNS	OBSERVED	2.48
POSITIVE	NEGATIVE	31.32
POSITIVE	VALUE	5.72
POTENTIAL	ITS	7.36
POTENTIAL	HAS	10.56
POTENTIAL	BENEFITS	2.52

PRESENCE	OF	241.12
PRESENCE	ABSENCE	18.52
PRESENTED	ARE	54.44
PRESENTED	HERE	11.2
PRESENTED	ABOVE	2.56
PRIOR	TO	98.68
PRIOR	WITHOUT	1.84
PROCESSES	COGNITIVE	5.24
RELATED	TO	203.4
RELATED	BE	29.04
RELATED	OTHER	9.08
RELATED	CLOSELY	12.12
RELATED	DIRECTLY	6
RELATED	ISSUES	5.56
RELATION	BETWEEN	29.96
RELATIONSHIP	BETWEEN	115.24
RELATIONSHIP	THERE	5.44
RELATIVE	IMPORTANCE	4.84
RELATIVE	POSITION	3.44
RELATIVELY	LITTLE	4.2
RELATIVELY	FEW	4.32
RELATIVELY	SMALL	14.2
RELATIVELY	SHORT	3.52
RELATIVELY	HIGH	12.48
RELATIVELY	STABLE	3.8
RELATIVELY	LOW	12.64
RELATIVELY	EASY	1.8
RELATIVELY	NUMBER	3.08
RELEVANT	ARE	12.72
RELEVANT	MOST	4.16
RELEVANT	INFORMATION	4.96
RELEVANT	PARTICULARLY	2.48
REPRESENT	THEY	6.2
REPRESENT	MAY	7.04
REPRESENT	DOES	2.32
REPRESENT	WOULD	1.8
REPRESENT	DO	2.44
REPRESENT	NOT	6.56
REPRESENTS	WHICH	6.56
REPRESENTS	EACH	7
RESPECTIVELY	AND	245.4
RESPONSE	TO	166.48
RESULTING	FROM	24.76
ROLE	IN	168.24
ROLE	PLAYED	15.52
ROLE	PLAY	43.32
ROLE	PLAYS	28.48
ROLE	IMPORTANT	25.16

ROLE	CENTRAL	4.76
ROLE	PLAYING	4.28
ROLE	CRUCIAL	4.36
ROLE	KEY	5.36
ROLE	MAJOR	5.16
ROLE	CRITICAL	5.64
ROLE	UNDERSTANDING	2.24
ROLE	ESSENTIAL	3.6
ROLE	DETERMINING	2.84
SHOWN	AS	123.92
SHOWN	HAS	55.12
SHOWN	THAT	84.08
SHOWN	HAVE	67.28
SHOWN	ARE	98.64
SHOWN	BEEN	51.36
SHOWN	TABLE	37.88
SHOWN	FIGURE	51.96
SHOWN	ALREADY	1.84
SHOWN	STUDIES	14.12
SHOWS	THAT	87.8
SHOWS	TABLE	33.16
SHOWS	FIG	66.92
SHOWS	FIGURE	50.36
SIGNIFICANT	WAS	74.24
SIGNIFICANT	THERE	51.64
SIGNIFICANT	WERE	57.32
SIGNIFICANT	NO	60.92
SIGNIFICANT	BETWEEN	46
SIGNIFICANT	DIFFERENCES	45.84
SIGNIFICANT	FOUND	17.92
SIGNIFICANT	DIFFERENCE	37.92
SIGNIFICANT	EFFECT	25.96
SIGNIFICANT	STATISTICALLY	57.92
SIGNIFICANT	HIGHLY	5.84
SIGNIFICANT	INTERACTION	9.16
SIGNIFICANT	IMPACT	4.56
SIGNIFICANT	AMOUNT	3.76
SIGNIFICANT	EFFECTS	11.32
SIGNIFICANTLY	MORE	21.12
SIGNIFICANTLY	FROM	22.68
SIGNIFICANTLY	THAN	37.92
SIGNIFICANTLY	BUT	7.72
SIGNIFICANTLY	WERE	40.92
SIGNIFICANTLY	DID	13.76
SIGNIFICANTLY	DIFFERENT	30.72
SIGNIFICANTLY	HIGHER	31.2
SIGNIFICANTLY	LOWER	19.08
SIMILAR	TO	177.4

SIMILAR	ARE	42.24
SIMILAR	OTHER	10.76
SIMILAR	THOSE	25.68
SIMILAR	VERY	18.72
SIMILAR	MADE	4.12
SIMILAR	FOUND	9.52
SIMILAR	QUITE	4.68
SIMILAR	WAY	4.84
SIMILAR	PATTERN	7.32
SIMILAR	FINDINGS	5
SIMILAR	SITUATION	1.88
SIMILAR	MANNER	4.12
SIMILAR	OBSERVED	10.04
SIMILAR	PATTERNS	3.8
SIMILAR	RESULTS	24.08
SOURCES	FROM	19.32
SOURCES	OTHER	8.72
SOURCES	SUCH	3.52
SOURCES	INFORMATION	6.76
SOURCES	AVAILABLE	1.56
SOURCES	DIFFERENT	5.24
SOURCES	MAIN	1.24
STRONGLY	MORE	4.8
SUBSEQUENT	ANALYSIS	2.96
SUGGESTS	THAT	125.92
SUGGESTS	THIS	40.36
SUGGESTS	SOME	3.96
SUGGESTS	THERE	3.84
SUGGESTS	ALSO	5.64
SUGGESTS	EVIDENCE	8.88
SUGGESTS	STRONGLY	2.4
SUGGESTS	RESEARCH	6.48
SUGGESTS	STUDY	5.32
TERM	LONG	86.08
TERM	SHORT	29.16
TERM	LONGER	4.28
TERM	EFFECTS	7.68
THEREFORE	IT	37.32
THEREFORE	BE	32.6
THEREFORE	WE	39.52
THEREFORE	CAN	15.68
THEREFORE	MUST	5.68
THEREFORE	SHOULD	6.52
THEREFORE	MAY	10.84
THEREFORE	CANNOT	3.08
THEREFORE	MIGHT	2.92
THEREFORE	POSSIBLE	4.04
THEREFORE	DIFFICULT	1.44

THEREFORE	NEED	2.24
THESE	ARE	232
THESE	WERE	136.88
THESE	HAVE	70.56
THESE	HOW	18.32
THESE	HOWEVER	29.72
THESE	MANY	15.56
THESE	QUESTIONS	13.52
THESE	EXAMPLES	6.88
THESE	CASES	19.08
THESE	LINES	8.24
THESE	FIGURES	6.88
THESE	STUDIES	35.96
THESE	SUGGEST	33.04
THESE	GROUPS	15.32
THESE	CHANGES	12.72
THESE	ISSUES	12.48
THESE	DIFFERENCES	21.92
THESE	SHOW	13.12
THESE	CLEARLY	6.04
THESE	TOGETHER	11.76
THESE	FACTORS	21.84
THESE	NONE	6.36
THESE	INCLUDE	15.08
THESE	INCLUDED	9.36
THESE	CATEGORIES	5.72
THESE	CONDITIONS	19.16
THESE	ELEMENTS	5.4
THESE	FEATURES	7.48
THESE	ASPECTS	4.92
THESE	RESULTS	88.28
THESE	SEEM	3.64
THESE	SOURCES	5.44
THESE	DESPITE	6.4
THESE	INDICATE	17.76
THESE	FURTHER	10.36
THESE	AREAS	9.52
THESE	REPRESENT	6.16
THESE	AUTHORS	7.2
THESE	PROVIDE	10.2
THESE	REASONS	5.12
THESE	APPEAR	4.4
THESE	RELATIONSHIPS	6.36
THESE	EFFORTS	3.52
THESE	DATA	44.4
THESE	FINDINGS	37.72
THESE	GENERALLY	5.36
THESE	ANALYSES	9.56

THESE	PHENOMENA	3.32
THESE	APPROACHES	6.76
THESE	DEMONSTRATE	6.92
THESE	REVEAL	2.88
THESE	ASSUMPTIONS	5.36
THESE	MEASURES	8.2
THESE	POSITIONS	2.36
THESE	SHIFTS	2
THESE	EXPLAIN	4.2
THESE	HELP	4
THESE	MODELS	14.12
THESE	PROGRAMS	3.68
THESE	REFLECT	3.96
THESE	LED	2.44
THESE	CONSISTENT	9.8
THESE	LOCATIONS	2.4
THESE	TECHNIQUES	3.92
THESE	INTERACTIONS	5.2
THESE	CONSIDERATIONS	3.28
THESE	IMPLICATIONS	3.72
THESE	SITUATIONS	2.4
THESE	CRITERIA	4.84
THESE	SETS	3.28
THESE	VARIATIONS	2.2
THESE	CORRESPOND	2.64
THESE	ESTIMATES	6.2
THESE	IMPLY	2.72
THESE	MECHANISMS	4.48
THESE	DIFFER	3.6
THESE	RELATE	1.48
THESE	VARIABLES	17.44
THESE	PROPERTIES	6.44
THUS	CANNOT	2.64
THUS	FAR	5.88
THUS	APPEARS	3.08
THUS	ALLOWING	1.72
TYPES	OF	145.12
TYPES	THESE	8.88
TYPES	OTHER	11.72
TYPES	BETWEEN	5.68
TYPES	ALL	8
TYPES	BOTH	9.12
TYPES	CERTAIN	2.52
TYPES	SEVERAL	3.16
TYPES	DIFFERENT	26.64
UNIQUE	ITS	2.16
UNIQUE	EACH	2.52
VALUES	ARE	63.88

VALUES	DIFFERENT	14.8
VARIOUS	FORMS	4.72
VARIOUS	ASPECTS	2.8
VARIOUS	DIFFERENT	4.44
VARIOUS	AMONG	3.44
VARIOUS	GROUPS	3.12
VARIOUS	TYPES	8.04
VARIOUS	INCLUDING	4.8
VARIOUS	LEVELS	4.36
VARIOUS	SOURCES	1.8
VARIOUS	FACTORS	2.96
WITHIN	CONTEXT	10.92
WITHIN	FRAMEWORK	8.96
WITHIN	PLACE	3.8
WITHIN	COMMUNITY	5.32
WITHIN	POSITION	3.72
WITHIN	EACH	19.96
WITHIN	SAME	8.2
WITHIN	FRAME	3.48
WITHIN	FIELD	4.76
WITHIN	INDIVIDUAL	4.24
WITHIN	SINGLE	3.72
WITHIN	GROUP	10.4
WITHIN	ACROSS	5.24
WITHIN	CONTAINED	2.76
WITHIN	RANGE	12.72
WITHIN	LIMITS	2.96
WITHIN	GROUPS	7.88
WITHIN	DAYS	8.52
WITHIN	AREA	4.76
WITHIN	CATEGORY	3.52
WITHIN	LOCATED	4.16
WITHIN	OCCUR	2.24
WITHIN	OCCURS	2.08
WITHIN	LOCATION	2.04
WITHIN	VARIATION	3.72