Lexical Intuitions and Collocation Patterns in Corpora

by

Iain David McGee

Centre for Language and Communication Research
School of English, Communication and Philosophy
Cardiff University

Thesis submitted for degree of Ph.D.
Cardiff University
June 2006

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed.....

Date.....28/6/06.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....

Date.....28/6/06.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signed.....

Date.....28/6/06.....

Thesis Abstract

Language teachers are often called upon by their students to provide examples of vocabulary usage in the classroom. Drawing on their experience of language, these teachers model lexical combinations and collocations, not only in their classes, but also in materials writing. However, corpus linguists have claimed that native speaker intuitions about the typical collocates of words are not reliable, because they do not align with the patterns observed in large corpora. These claims are critically evaluated, and an alternative explanation for the mismatch, the possibility that the corpora might not be representative of actual language in use, is also examined. Various linguistic and psycholinguistic explanations for the disparity between corpus data and elicited data are examined, and theories dealing with the mental representation of collocations are also discussed. Data from word frequency estimate research, and word association research are also analysed for relevant information on the subject. Five experiments are then reported, investigating the ability of native speakers (students and EFL teachers) and non-native speakers (Arab university teachers) to rank, recognize and spontaneously produce frequent adjective-noun collocations. The results indicate that a key factor affecting the 'quality' of lexical intuitions may be the employment of an 'availability heuristic' in judgements of frequency. It is argued that some collocates of words may be more hidden from memory searches than others, and that there may be a systematic bias in the respondants' lexical intuitions based on how words are stored in the mental lexicon. Conclusions are drawn that reflect the many facets of research relevant to the questions under discussion: corpus linguistics, frequency theory, word association research, learning theory and theories of lexical storage. The thesis ends in applying some of the key findings to language teachers.

Acknowledgments

I would like to thank Prof. Alison Wray for her help and support during the writing of the thesis. She has shown great patience and given much encouragement along the way: she has been a model supervisor. I would also like to thank Gaetanelle Gilquin for alerting me to some relevant research on elicited data and corpus data research.

The library staff at King Fahd University of Petroleum and Minerals, Saudi Arabia, have been very helpful in obtaining copies of hard-to-come-by journal articles and books – I am thankful for their help and hard work.

Without the respondents giving me their time, this research could not have been conducted and I gratefully acknowledge their willingness to participate in the experiments which are reported in this thesis.

Last, but not least, I should like to thank my wife and children for their patience and support, and for giving me the time to put in the many hours that have been spent on this project. This thesis is dedicated to them.

Table of Contents

Page No.

Declaration
Thesis Abstract
Acknowledgments

Chapter 1
The Claims of Corpus Linguists against Lexical Intuitions 1

1. Introduction
2. Background
3. Specific claims
    3.1. Collocation
    3.2. Frequency
    3.3. Semantic prosody and pragmatics
    3.4. Phraseology
    3.5. The true place for intuitions
4. Explanations for data differences
    4.1. Connotation, denotation and delexicalisation
    4.2. Saliency
5. Summary

Chapter 2
Corpus Representativeness: Establishing the Parameters of Usage 31

1. Introduction
2. Objection 1: Different registers different data
    2.1. Word frequencies
    2.2. Word meanings
    2.3. Collocations
    2.4. Spoken and written English
    2.5. National Englishes
3. Objection 2: Representativeness
    3.1. Size
    3.2. Content
    3.3. The BNC
4. Objection 3: The Internet as a super-corpus
5. Objection 4: The dating of a corpus
6. Objection 5: Corpus material and significance
7. Objection 6: Different corpora different data
    7.1. Comparisons of 'traditional' corpora
    7.2. Comparisons of 'traditional' corpora and the Internet.
        7.2.1. Frequencies of words
        7.2.2. Frequencies of multi-word items
8. Summary

Page No.

Chapter 3
Collocation Classification and Psycholinguistic Representation.....63

1. Introduction
2. Collocation: definition and classification
    2.1. Co-occurrence of words
    2.2. Frequency
    2.3. Restrictedness
    2.4. Selectional restrictions and embedded collocations
3. Categorising adjectives
4. Collocation representation: psycholinguistic perspectives
    4.1. Background
    4.2. Psycholinguistic explanations for data differences
    4.3. Wray's formulaic language model
    4.4. Evidence for the existence of fused language or formulaic language
5. Summary

Chapter 4
Word Frequency Estimation and Frequency Estimation Theory.....94

1. Introduction
2. Frequency estimation research
    2.1. Word frequency
        2.1.1. Methodological issues
        2.1.2. Different corpora different results?
        2.1.3. Priming effects
    2.2. Collocation frequency
    2.3. Miscellaneous
3. The coding of and access to frequency information
    3.1. Coding of frequency information
        3.1.1. Indirect and direct coding
        3.1.2. The automatic encoding of frequency
    3.2. Access to frequency information
4. Summary

Chapter 5
Ranking Restricted and Free Collocations.....116

1. Introduction
2. Statistical measures of collocation strength and raw frequency co-occurrence data
3. Experiment 1
    3.1. Rationale for experiment and hypotheses
    3.2. Experiment design
        3.2.1. Choice of adjectives and collocations

Page No.

3.2.2. Differences in frequency between test items
3.2.3. Collocation frequency and raw noun frequency
3.2.4. The sets for testing
3.3. Methodology
3.4. Subjects
3.5. Results and analyses
3.6. Discussion: testing the hypotheses
3.6.1. Evidence of embedding effects
3.6.2. Evidence of saliency effects
3.6.3. Miscellaneous explanations
4. Summary

Chapter 6
Insights from Word Association Studies.....151

1. Introduction
2. Review of word association testing
2.1. Background
2.2. Methodological issues
2.2.1. Controlled association and free association
2.2.2. Single and multiple responses
2.3. Types of stimuli and response classification
2.3.1. Stimulus type and its effect on response type
2.3.2. Response types: paradigmatic, syntagmatic, clang
2.4. Native speaker and non-native speaker responses
2.4.1. Native speakers
2.4.2. Non-native speakers
2.4.3. Developmental issues
3. Adjective-noun collocations in word association tests
3.1 Introduction and research questions
3.2. Analysis of Moss & Older (1996) data
3.2.1. Adjective stimuli and responses
3.2.1.1. Analysing the data
3.2.1.2. Results
3.2.1.3. Discussion
3.2.2. Noun stimuli and responses
3.2.2.1. Analysing the data
3.2.2.2. Results
3.2.2.3. Discussion
4. Summary

Page No.

Chapter 7
Testing Productive and Receptive Knowledge
of Noun Collocates of Frequent Adjectives.....192

1. Introduction
2. Experiment 2
    2.1. Main research question and hypothesis
    2.2. Subjects
        2.2.1. Group 1 – Native speakers (NSs)
        2.2.2. Group 2 – Non-native speakers (NNSs)
    2.3. Designing the experiment
        2.3.1. Justification for methodology adopted
        2.3.2. Criteria affecting the choice of word stimuli
        2.3.3. The issue of lemmatisation
        2.3.4. The adjective stimuli
    2.4. Method
    2.5. Quantitative analyses – justification for choice of test
    2.6. Results
        2.6.1. Native speakers
            2.6.1.1. Analysis 1 – Excluding ties
            2.6.1.2. Analysis 2 – Retaining ties
        2.6.2. Non-native speakers
            2.6.2.1. Analysis 1 – Excluding ties
            2.6.2.2. Analysis 2 – Retaining ties
        2.6.3. Comparison of native speaker and non-native speaker responses
            2.6.3.1. Quantitative analyses
                2.6.3.1.1. Number of different responses
                2.6.3.1.2. Comparison of NS and NNS group ranks
            2.6.3.2. Qualitative analyses
        2.6.4. The dominant responses
    2.7. Discussion
3. Experiment 3
    3.1. Main research interest and hypothesis
    3.2. Designing the experiment
    3.3. Method
    3.4. Subjects
    3.5. Results
        3.5.1. Native speakers
            3.5.1.1. Analysis 1 (Majority choice)
            3.5.1.2. Analysis 2 (Votes)
        3.5.2. Non-native speakers
            3.5.2.1. Analysis 1 (Majority choice)
            3.5.2.2. Analysis 2 (Votes)

Page No.

3.6. Discussion
4. Summary

Chapter 8
Testing Productive and Receptive Knowledge
of Adjective Collocates of Frequent Nouns.....252

1. Introduction
2. Experiment 4
    2.1. Rationale for experiment
    2.2. Research questions
    2.3. Choice of noun stimuli
    2.4. Method
    2.5. Subjects
        2.5.1. Native speakers (NSs)
        2.5.2. Non-native speakers (NNSs)
    2.6. Results
        2.6.1. Native speakers
            2.6.1.1. Analysis 1 – Excluding ties
            2.6.1.2. Analysis 2 – Retaining ties
        2.6.2. Non-native speakers
            2.6.2.1. Analysis 1 – Excluding ties
            2.6.2.2. Analysis 2 – Retaining ties
        2.6.3. Comparison of native speaker and non-native speaker responses
    2.7. Discussion
        2.7.1. Differences between the groups
        2.7.2. BNC – NS differences
        2.7.3. The dominant responses
            2.7.3.1. Embedding and choices
            2.7.3.2. Denotational meaning and polysemes
    2.8. Summing Up
3. Experiment 5
    3.1. Hypothesis
    3.2. Method
    3.3. Subjects
    3.4. Results
        3.4.1. Native speakers
        3.4.2. Non-native speakers
    3.5. Discussion
        3.5.1. Embedding
        3.5.2. z-scores
        3.5.3. Miscellaneous
    3.6. Comparing the results from experiments 3 and 5
4. Summary

Page No.

Chapter 9
Conclusion.....302

1. Overview of key findings
2. Limitations of the study and suggestions for further research
    2.1. Assumptions
        2.1.1. Elicitation tests
        2.1.2. Collocations
    2.2. Methodological issues
        2.2.1. Different instructions
        2.2.2. Extra clues
        2.2.3. Restrictions on the collocates permitted
    2.3. Subjects
        2.3.1. NS subjects
        2.3.2. NNS subjects
            2.3.2.1. Learning style
            2.3.2.2. Native language
            2.3.2.3. Different stimulus words in different languages
    2.4. Analyses
        2.4.1. Reconsidering the less frequent responses
        2.4.2. Embedding explanation
3. The last word

References.....313

Appendices

Appendix 1 – The Formulas of MI and z-score.....342
Appendix 2 – Task Sheet for Experiment 1.....344
Appendix 3 – Distributions in BNC and Brown Corpus
for Adjective Stimulus Words, Experiment 2..... 346
Appendix 4 – Stimulus Words (Experiment 2) and Collocate Details.....347
Appendix 5 – Task Sheet for Experiment 2.....350
Appendix 6 – NS and NNS Responses for Experiment.....351
Appendix 7 – Task Sheet for Experiment 3.....359
Appendix 8 – Distributions in BNC and Brown Corpus
for Noun Stimulus Words, Experiment 4.....360
Appendix 9 – Stimulus words (Experiment 4) and Collocate Details.....361
Appendix 10 – Task Sheet for Experiment 4.....364
Appendix 11 – NS and NNS Responses for Experiment 4.....365
Appendix 12 – Task Sheet for Experiment 5.....372

Chapter 1 – The Claims of Corpus Linguists against Lexical Intuitions

1. Introduction

In what situations do you use the word *little* rather than *small*? What is a typical noun collocate of the word *similar*? Which is more frequent in the English language *bad news* or *bad luck*? Which word most commonly fits into the slot in the following phrase: *a torrent of* _____[1]? Language teachers, in particular, are called upon by their students to answer questions similar to these, and to provide on-the-spot examples of vocabulary usage on a daily basis in the classroom. However, corpus linguists question whether native speaker EFL teachers, or any native speakers of English, are able to 'correctly' answer questions such as those provided above, if they depend only on their intuitions to do so. Corpus linguists view corpus data[2] as a more reliable source of language data: large corpora can be checked to investigate the common collocates of a word, to examine phrases and to identify the relative frequencies of collocations. The goal of this research is to compare lexical intuitions with corpus data in the area of collocations and to discover if, when and why they differ.

This view, that language intuitions are unreliable, seems odd, as there appears to be no obvious reason why, for example, native speaker language teachers should not be able to answer the above questions 'reliably'. After all, they have years of experience in listening to and engaging in conversation, and have read millions of words in such diverse formats from newspaper articles to billboard advertisements to medicine prescriptions. Their exposure to language is not only vast but also varied: one would think that this would ensure accuracy and reliability in describing the language to students and modelling the relevant patterns/combinations. One would assume, following Malmkjær (1993, pp.214, 215), that intuitions are formed and moulded by language use and exposure to language use. This being so, and because corpora are examples of language use, it does not seem unreasonable to suppose that the two sources of data should indicate the same kind of

[1] The 'correct' answers to all these questions are provided in this thesis.
[2] Definitions of corpora are noted in chapter 2, section 1.

things about language. However, corpus linguists do not, generally, hold this view. A key figure in corpus linguistics, John Sinclair, makes the following comment:

> It is almost impossible to invent an adequate example; attempts by language teachers, lexicographers and others to represent usage are often embarrassing and never reliable (1997, p.31).

If there is a mismatch between these two types of data there are two possible explanations for the differences. Either Sinclair is right and intuitions are not reliable in some way, or, alternatively, it may be the case that corpora do not reflect language usage, i.e. the 'problem' is with the corpus not with the intuitions. Corpus linguists have not, typically, considered whether this second explanation might account for the mismatch that they believe exists. This alternative explanation for the disparity between corpus data and intuitions is discussed in more detail in chapter 2.

2. Background

For corpus linguists, the authority which determines what is normal and typical in language is corpus data, not intuitively provided data. Some corpus linguists, e.g. Beaugrande (1996, pp.526, 527) and Sinclair (1991a, p.1), have seen corpus linguistics research as a return to the 'real stuff' of formal linguistics: a return to pre-Chomskian fieldwork - a return to examining language usage. In contrast to Chomsky's focus on the possible, the corpus linguist focuses on the performed - the normal, the frequent, the typical - where the corpus provides the performance data (Widdowson 1991, p.13). Of course, whether one wishes to focus on what is *typical*, rather than *possible* in one's description of language is a matter of choice and this is a fundamental difference between advocates of corpus linguistics and those who assign it a less important place. Fillmore (1992) caricatures the Chomskian linguist's response to all the data provided by the corpus linguist by referring to a comment by Michael Polyani: if the study of frequent phenomena was the prime mover in the world of science then most scientists would be studying interstellar dust (Fillmore 1992, p.37). Just because something is frequent it does not necessarily mean that it is worthy of study or particularly interesting.

It is important to point out that corpora *per se* are not a divisive issue in linguistics, for even critics of corpus linguistics concede that corpora are useful. For example, Owen (1993) concedes that "There is no doubt that computer-assisted corpus linguistics does reach some parts of the language other grammars fail to reach" (p.184); and Cook (1998) acknowledges that "Computerized language corpora have inspired some of the most important insights in recent linguistics" (p.57). The issue of controversy is more about 'significance of place' as Murison-Bowie (1996, p.182) puts it. In the context of language description and teaching, he distinguishes the *strong case for corpora*, (the idea that *no* work can be done without corpora) and the *weak case* (where corpora are viewed as *an additional* resource adding to our knowledge of language). 'Strong case' advocates assign corpora a very significant place and do not assign intuition a place at all in the description of language.

The term 'intuition' has been used in different ways, and in this research I shall adopt the following definition: "a judgment is termed intuitive if it is reached by an informal and unstructured mode of reasoning, without the use of analytic methods or deliberate calculation" (Kahneman & Tversky 1982, p.494). Linguistic intuition has typically been used with reference to grammar: grammaticality judgements, the noticing of ambiguity etc. (Fiengo 2003). These intuitions are important for Chomskian linguists, who, by and large, are quite happy to rely on them in the description of language. It is important to establish at the beginning of this research that it is not these kinds of intuitions that have been called into question by corpus linguists: Sinclair readily admits a place for intuition in recognizing the "well-formedness of sentences in isolation" (1997, p.32). However, if we push the concept of grammaticality to extend to patterns of grammar, and push it even further to word combinations, it is in just these sorts of areas that corpus linguists challenge the reliability of intuition. Such intuitions I shall term 'lexical intuitions' throughout this study. Regarding the unreliability of lexical intuitions, Hoey (2000, p.237) says "Intuition, even the intuition of the best lexical applied linguists, is likely to be flawed" and Francis (1993, p.139) makes a similar point: "Intuition may be useful to linguists in a number of ways, but for the purposes of saying exactly how language is used, it is notoriously unreliable". Regarding the combination of words Biber et al. (1996,

p.120) believe that, "Intuitions regarding lexical associations are often unreliable and inaccurate" and Beaugrande (1999, p.247) states that, "Intuition is only weakly predictive when speakers of a language are asked to state which selections and combinations can or cannot occur".

Others though, are not so convinced about the superiority of corpus data. For example, Owen (1996, p.219) has questioned the need for corpus data to verify intuitive prescription in language teaching, and Cook (1998, p.59) has argued that an individual's experience of language is broader and more complete than the material in a corpus: "Corpora are **only partial** authorities...This is why our intuition...can still tell us facts about the language which can not be evidenced by a corpus..." (1998, p.59, emphasis mine). The position that lexical intuitions are superior to corpus data has actually been held by some word frequency estimate researchers (e.g. Ringeling 1984; Carroll 1971; Frey 1981) and their views are discussed in further detail in chapter 4. While it is true that corpora have developed in significant ways since these word frequency estimate researchers expressed these views (especially in the matter of size, as noted by McEnery & Wilson 1996, p.164; Leech 1991, p.13; Svartvik 1996, p.9) doubts about corpora reliability remain, as will be discussed in chapter 2.

Between the *strong case* advocates and those who reject outright the view that corpus data is superior to intuitive data are the writers who have charted a 'middle way', those who argue that while lexical intuitions may not be 'infallible', they are still important. For example, Meijs (1996, p.102) believes that there is nothing wrong with intuition *per se*, but that corpus data can help in making more unbiased and more informed judgements. Aston (1997, p.54) and Takaie (2002, p.112, p.117) argue that intuitions and corpus data complement each other and help provide a more complete picture of language. Widdowson (2000) believes that different sources of language data provide different information, but that such differences should not simply be viewed as defects. He comments, "Corpus analysis reveals textual facts, fascinating profiles of produced language, and its concordances are always springing surprises.... But this achievement of corpus analysis at the same time necessarily defines its limitations. For one thing, since

what is revealed is contrary to intuition then **it cannot represent the reality of first person awareness**" (2000, p.6, emphasis mine).

It is suggested, therefore, that intuitions should not simply be jettisoned from their role in language description, but rather, they should be understood and studied. This research seeks to carefully examine and understand lexical intuitions, rather than simply criticize them.

Having made the above introductory comments, and having noted some of the issues surrounding the subject, it is now necessary to outline and analyse the specific claims made by corpus linguists against the reliability of our lexical intuitions. In addition, I also note the explanations which corpus linguists have forwarded to explain why there is a mismatch between corpus data and elicited/introspective/intuitive data.

3. Specific claims

Hunston believes that language intuition is weak in the areas of: collocation, frequency, prosody and phraseology (2002, pp.20, 21). Although Hunston does not expand upon her assertion, her taxonomy is a useful one and provides the framework for the following overview and critique. While this thesis is particularly concerned with collocations and frequency, comments about prosody and phraseology will also be examined, as there is a certain degree of overlap between these subjects. What follows, then, is a critical evaluation of the claims and counter-claims regarding lexical intuitions and corpus data.

3.1. Collocation[3]

Regarding collocation, Hunston (2002) argues that we can intuit some examples of common collocates (*play-game* is her example), but she goes on to say that native speakers are not conscious of such combinations as *fairly accurate*, *fairly certain*, *fairly small* and *fairly wide* (2002, p.21). Hunston suggests this on the grounds that such

[3] Note that a detailed discussion of collocation is found in chapter 3

combinations have not been included in language course books. Whether we can extrapolate from textbook exclusions to poor language intuition is rather questionable, and this subject is developed in section 3.2, below. Like Hunston, Stubbs (1995a) argues that examples of collocates can be given "sometimes accurately" (p.24), but on the whole he believes that the production of collocates on demand is weak: "[native speakers] certainly cannot document collocations with any degree of thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations" (1995a, pp.24, 25). Both Hunston and Stubbs fail to explain why intuition is only *sometimes* unreliable, and so their comments leave us none the wiser as to when exactly intuitions are supposed to fail.

Sinclair (1997) has similarly argued against the ability of native speakers to produce typical collocates of words out of context. To support this claim Sinclair suggests that even the most banal of words - *nice* - has patterns of association "that one can hardly imagine will be retrieved [from intuition]" (1997, p.33). For example, he notes that according to corpus data it attracts *a* and rejects *the*, attracts modifiers in predicative position, is used attributively with other adjectives and when preceding a noun without a modifier attracts, inter alia: *day, evening, boy, girl, surprise* etc. While Sinclair's comments are more specific (in that he actually forwards a specific case – see also his comments about *glad,* 1997, p.33), it is important to note that he provides no native speaker data to compare with the corpus data provided. This is not to say that Sinclair does not have evidence – indeed he speaks of the "impromptu reactions of hundreds of fluent speakers" as support for his position that intuitions differ from corpus data (1997, p.29). However, unless these reactions are documented, skeptics can easily view the claims as no more than anecdotal.

Fox (1987) reports on a (small-scale) study comparing intuitions about collocations and corpus data. In her experiment, 53 students (native speakers) were asked to give the 5 most likely collocates of *feet*. In analyzing her data, Fox notes the prevalence of semantic set words in the responses (that is, naming other parts of the body as collocates of *feet*) and also notes that several of the responses were found to be highly frequent i.e.

concurring with her corpus: e.g. *tall* and several numbers. Despite some correlation, Fox notes the absence of *high* and *long* from her native speaker data, items which she terms "extremely significant" (ibid, p.146) collocates according to her corpus. An additional stimulus word presented to the same subjects was the word *hint*. Fox reports responses which were highly frequent in her corpus such as *subtle, small, clue, give* and *take* (the last of these being the most frequent collocate according to her corpus), but notes that none of her native speakers provided the second most common collocate from her corpus – *no*. Unfortunately, Fox's report is lacking in many aspects. She does not mention whether respondants were asked to provide left or right collocates (and within a particular window) and we are not told in her discussion whether all 5 responses of the subjects were considered in her analysis (250+ answers). It is unclear why she chose to provide a polysemous noun (*feet*), in a task whose stated purpose was investigation of collocation production, and this perhaps confounds her stated research focus - collocation - with polysemy. The fact that the most frequent collocates of *feet* were 'measurement' collocates in her corpus, and 'body part' collocates from her subjects, while interesting, hardly establishes Fox's conclusions about poor knowledge of collocations. Further, the fact that *no* is not a 'full' lexical content word should not be overlooked in looking at the *hint* stimulus responses: it may be that respondants believed that they had to provide 'lexical' words – we do not know. Unfortunately, Fox provides no statistical analyses of her data, and when reporting that certain collocates were 'significant' she does not state what the basis of the significance was: pure frequency, z-score, MI score etc, (see chapter 5, section 2 on these statistical measures). In sum then, Fox is unable to provide, in her methodology and report, a robust justification for her conclusion that "intuition is not as good as evidence" (1987, p.146).

Beaugrande (1996) studied the verb *warrant*, focusing on the semantics of the word and its collocates in 392 lines of *warrant* concordance lines from the Bank of English. He observes that his pre-existing intuitions about the semantics of this word were narrow, i.e. he had only appreciated the legal meaning of the word (1996, p.523). On looking at the corpus data he notes his failure to realize that the verb also has several 'general' noun collocations (e.g. *situation* in the collocation *'....situation warrants.....'*), and this leads

him to question the validity of his intuitions (ibid, p.524). While his discussion and observations are interesting, Beaugrande's conclusions about his intuitions are perhaps a little premature, as he admits that he does not actually use the word *warrant* often, finding it "a bit stuffy or pompous" (ibid, p.523).

Although it was noted above that little explanation has been provided as to why it may be that some typical collocates can be provided for some words but not for others, there is some evidence that the typical collocating partner of some words in a particular type of collocation may be quite easy to produce, and it is to this subject that we now turn.

The limitations of her own study notwithstanding, Fox (1987) makes some interesting general comments about collocation. She argues that 'restrictedness' provides the key why intuitions about word partners are sometimes good[4]. Fox believes that the frequent/typical collocational partner of a word can be provided if it is the case that the stimulus word is the 'restricted' word in a 'frozen' collocation (1987, p.146). She believes native speakers could provide *image* as a collocate for *graven*, and *hair* for *blonde* without any difficulty, because of the severely limited set of collocations that these words have: *graven* describes very few words other than *image*[5], and the same is the case for *blonde* – typically describing hair[6]. For Fox, it seems that as soon as there are real choices (in the production of collocates), i.e. as soon as the network of associates of a word becomes larger, intuition about typical collocates from among these associates becomes suspect[7]. Unfortunately, Fox provides no data to establish this claim, yet her position seems justified (both in recognition, and production abilities) as noted in the two studies described below.

---

[4] A full and evaluative consideration of the definition of 'restricted collocation' is given in chapter 3, section 2.3.

[5] This is the case when it is a noun directly following the adjective in the BNC.

[6] Though of course there are a number of other collocations too. In an adjective search in a ±5 collocation window in the BNC there are many other nouns which collocate with blonde, but they are not nearly as frequent as *hair* (269 instances). The next most frequent noun collocate is *woman* (40 instances).

[7] Lewis (1997), though not using the term 'frozen', may have the same idea in mind when he argues that we can provide typical noun collocates of *golden* (p.30), namely: *opportunity, wedding, age, mean, boy/girl, handshake.*

Granger (1998) asked 56 French learners of English i.e. non-native speakers (NNSs) and 56 native speakers (NSs) of English to choose, from a list of 15 adjectives, the acceptable collocates of 11 amplifiers and to indicate with an asterisk which adjective was most frequently associated with the amplifier. She reports only on the responses which were asterisked. She notes that the NNSs were not nearly as uniform in their responses as the NSs. With regards to the NS uniformity, 43 of the NS respondants believed *available* to be the most frequent associate of *readily*; 33 believed that *aware* was the most frequent collocate of *fully*; 33 indicated that *significant* was the most frequent collocate of *highly*, and that there was a split response between *happy* (19) and *ignorant* (20) as being the most frequent collocates of *blissfully*. BNC corpus data confirms that these collocates are the most frequent of the options provided by Granger[8]. Benson et al. (1986) list *readily available*, and *blissfully ignorant* as restricted collocations, but they do not include *fully aware*, *highly significant*, or *blissfully happy* as restricted collocations in their dictionary[9]. There is then, some support for the hypothesis that typical restricted collocates can be accessed by respondants, but also, some frequent collocates in combinations not classified as restricted may also be accessible.

Greenbaum (1988, chapter 9) is one of the very few researchers who has actually used elicitation tests to help investigate collocation. He required British and American undergraduate students to complete preverb intensifier sentences with appropriate verbs, for example, *I entirely*..... He provided the principal collocates for six stimuli sentences elicited from the British and American students. The dominant responses (and the percentage of respondants who provided them) are given below.

1. I badly.....need (65% UK) need (48% US)

2. Your friend very much..... likes (29% UK) likes (19% US)

3. They all greatly.....admire (44% UK) appreciate (24% US)

4. I entirely.....agree (82% UK) agree (27% US)

---

[8] It should be noted that, according to the BNC, *successful* and *unlikely* are more frequent collocate partners of *highly* than *significant*. Also, *unaware* is more common than *happy* and *ignorant* as a partner of *blissfully*. However, these adjectives were not provided in the Granger study.

[9] Benson et al.'s (1986) collocation dictionary is discussed in more detail in chapter 3, section 2.3, and chapter 5, section 2.2.

5. They are utterly.....(no response having above 10% attestation from either group)

6. I completely.....forget (50% UK) forget (46% US)

Greenbaum (1988, chapter 9) also notes that though a particular verb may not have been produced by a large number of the respondants, there were strong 'semantic set' preferences, e.g. 85% of the responses produced by the British group gave *greatly* an approbatory verb. He also notes that though *utterly* elicited no dominant verb, the verbs produced were all negative in orientation/semantic prosody[10]. Greenbaum notes that there is general agreement between the two sets of respondants, though the case of *entirely* does seem to indicate differences (Americans including within this set 'failure' verbs e.g. *forget*).

Two questions naturally arise from this experiment. Firstly, are the elicited collocates different from those that would emerge from large corpora, as some of the claims mentioned earlier would predict? Secondly, are the collocations 'restricted'? Regarding the first question, a check of the BNC was conducted in a +1 right collocation window span for the preverb modifier, excluding adjective collocates, with words that could, semantically, be put into the incomplete sentences provided by Greenbaum (there was not enough data in the BNC to check the exact sentence for some of the combinations)[11]. The BNC confirms that the responses are the most frequent co-occurring verbs for sentence prompts 1,3[12], 4 and 6, though at times a different form of the verb had greater attestation in the BNC (e.g. more past tense than present tense instances). For the fifth prompt, *utterly*, the most frequent collocates in the BNC are *opposed, defeated, fail* and *condemn*, i.e. they are all negative in prosody concurring with the prosody of the responses from the subjects. Benson et al. (1986) classify the combinations for sentences 1, 3 (both options), 4 and 6 as restricted collocations. This research provides some empirical support for

[10] This is an important observation in connection with section 3.3 below.

[11] Sentence 2 was checked using *he very much* and *she very much*, but interestingly, *likes* was not present even once in the data. Further, there was a variety in the responses in terms of the semantic preferences for this word, e.g. *doubt, enjoy, want, hope.*

[12] There are more instances of *admire*, than *appreciate* in the BNC.

Fox's belief that typical partners in restricted collocations are open to our intuitions and can be provided in test-type conditions.

The corpus linguists Hunston, Stubbs, Sinclair, Beaugrande, and Fox, all question the reliability and accuracy of native speaker intuition concerning typical collocate partners. However, the evidence that has been produced to support this viewpoint is less than convincing. Fox has suggested that respondants can provide reliable intuitions about the partners of words in severely restricted collocations and there is some empirical support for this view, as noted above. The factor of restrictedness may be important in affecting the quality of lexical intuitions: clearly more research is required into this subject, and the subject of collocation is examined in considerably more detail in chapter 3.

3.2. Frequency

There are two key areas of interest with regards to frequency and lexical intuitions: the ability to document the relative frequencies of different words in the language; and consciously knowing the relative frequencies of different meanings of the same word. Hunston (2002) has only made comments about the former, though both are discussed here.

Hunston argues that "It is almost impossible to be conscious of the relative frequency of words, phrases and structures except in very general terms" (2002, p.21). She goes on to state that we can guess (correctly) that *take* is more frequent than *disseminate*, but not whether *fare* is more frequent than *fantasy* – presumably she makes this point on the basis of the differences in these two pairs of words' relative frequencies[13], but she does not elaborate on the matter any further, or give evidence to establish her case. Further discussion of her main point is provided in chapter 4.

[13] Hunston (2002, p.21) notes that there are approximately 4000 instances of *fare* and 10000 of *fantasy* in the Bank of English. In the BNC there are 69630 instances of *take* and 121 instances of *disseminate*.

Regarding the ability to identify or provide the most frequent meaning of a particular polysemic word, much more has been said by corpus linguists. In his overview of the Collins COBUILD English course, Willis (1990) makes a number of comments contrasting this work (based on the COBUILD corpus) with previous course books and pedagogic grammars. He notes that several word meanings have been neglected in pre-corpus grammars and textbooks. In particular, less frequent uses of words have been given more attention in such materials than the more frequent uses revealed by corpus data. For example, he provides corpus data to show that while *any* is indeed used in questions and negatives (as textbooks typically suggest), it occurs more frequently in affirmative sentences than in negative and question sentences put together (1990, p.49). He also argues that the past tense use of *will* (i.e. *would*) has been stressed in course books in reported speech, but the past habit usage of the word has been neglected (e.g. *the old man would walk down with me...*), though it has far more corpus data representation (21 percent versus 6 percent of data respectively, 1990, pp. 49, 55, 124). Willis believes that oversights such as the above are because of course writers' dependence on intuition in the development of their materials. With regards to the COBUILD series he is adamant that "Intuition alone would not have identified the most frequent words and phrases of the language, or recognized their importance" (1990, p.124).

So, is intuition to blame for the 'wrong' emphases in pre-corpus textbooks?[14] Firstly, it should not be taken as read that the goal of a language course book is the provision of the most frequent meanings of particular words. There is a general recognition that other factors too should play a role in what is, or is not included in a grammar, or textbook. For example, Barlow (1996, p.30), identifies complexity as an important consideration, stating that, "there is something to be said for presenting simple examples in a reference grammar that correspond to the semantic prototype" (1996, p.30; see also Biber et al.'s comments about how 'difficulty' plays a role in determining what is or is not included in course books 1994, p.174). Additional factors noted as influencing the creation of course materials are: 'tradition' (Byrd 1995, p.46) and 'teachability' (Widdowson 1991, p.21). It

[14] This discussion also addresses the Hunston comment noted in section 3.1.

may be then, that certain meanings of words are emphasized above others because of such considerations. Returning to Willis' examples, most teachers would recognize that contrasting *some* with *any* helps make teaching questions easier. Teaching the past habit use of *would* is difficult, as most teachers would attest, for it depends on subtle intonation and a particular evaluative stance to be used correctly. As Widdowson notes, teachers of all subjects are "economical with the truth in the interests of effective pedagogy" (1991, p.21). It would seem then, that the argument that lexical intuitions are weak in the area of meanings and frequency, if based on textbook omissions etc., may have failed to seriously consider the role of pedagogical considerations in influencing what material is included or excluded from these books.

Stubbs (2002a, p.9) notes that some meanings of some words are not given in pre-corpora dictionaries. In particular, he believes that connotative meanings have been overlooked, and the importance of the role of collocations in establishing meaning has been neglected. Sinclair (1991a), similarly, has set his sights on the lexicographer of yesteryear (working without the benefit of computerized corpora) and been critical of the outcome. For example, he notes that the Collins English dictionary (2nd Edition 1986) places the 'follow' meaning of *pursue* as the first sense meaning, relegating the most common usage 'to apply oneself to', to the fifth meaning (1991a, p.113). He also makes a strong connection between introspection affecting how different meanings are listed, a point explored in section 3.3 below. Cook (1998, p.59) is not very impressed with the above kinds of observations. He argues that it is not wrong to present learners with the 'prototypical' meaning first in a dictionary, as it is from this meaning that all other meanings derive. From a pedagogical viewpoint, therefore, one might argue that such a practice is justifiable.

Turning our attention from textbook and dictionary omissions to the more directly relevant intuition of language teachers, Renouf (1997, pp.259, 260), writing in the context of teacher training, argues that challenging teachers to think about word meanings can convince them about the benefits of using corpus data. She describes how she asks trainee teachers questions about, inter alia, the primary meanings of *keep* and *see*

and the use and meaning of *listen*, *affirm* and *confirm*. She notes the weaknesses of the explanations when compared to corpus data – but does not provide details. Renouf believes that the type of exercise she describes is a useful way of helping teachers realize their own limitations in terms of their explicit knowledge of language facts. Once again, it is unfortunate that the actual responses of such groups have not actually been recorded and critically analysed.

In their 1994 study, Biber et al. look in some detail at the use of the word *certain* in social science and fiction text types. They note that in "an informal survey" (1994, p.178) native speakers associated the word "with the condition of certainty" (ibid, p.178). In their view, and appealing to their corpora, they argue that this is a 'rare' sense, and that the word *certain* is "much more commonly used to mark a referent as named but not clearly described or known, as in 'a certain kind'" (ibid, p.178). They found (in their social science and fiction texts) that only a few of the collocates of *certain* indicated certainty. The 4 preceding collocates used with *certain* in this sense were: 1) *it BE* + ____; 2) *you/he/she/they BE* +_____; 3) *I / we BE* + _____; 4) *BE quite* + _____. The only collocate following the word that rendered the 'certainty' meaning was + *that.*

It is unfortunate that Biber et al. do not elaborate on what their 'informal study' asked, because how such a study was conducted could easily have influenced the responses. Were the respondants informed, for example, that their intuitions would be measured against social science and fiction texts? Considering that Biber continually stresses how register affects lexis in so many ways[15], it is surprising that his respondants seem to have been given such a 'vague' task. If the respondants were asked to define the meaning of *certain* (in isolation), then there is a methodological inconsistency. Intuitions about lone word meaning cannot fairly be compared with collocation data (for more on this, see the discussion in section 4.1, below).

Biber et al. arguably overstate their case in calling the '*certain* = *certainty*' meaning 'rare'. While it may be the case in terms of tokens, this is only half the picture. Biber et

[15] See chapter 2, section 2.

al. seem to pay little attention to the average frequency columns in their table (listing the number of times the *certain* collocates occur per 100,000 words of social science text and fiction text). While it is true that only 4 of the 14 preceding collocates of *certain* are consistent with a meaning related to certainty (it should be noted that *collocates* includes full stops as one category and commas as another), in 'Fiction', of the 14 different collocate types, the *you/he/she/they BE + certain* collocation is the second highest in terms of average frequency (6.4 instances in 100,000 words), coming after *a* + (7.3 examples in 100,000 words).

If we consider now the words that follow the word *certain*, it is the one indicating certainty that is the most common – at least in fiction (+ *that*, 1.2 instances per 100,000 words). The next highest is + *amounts* with 0.8 instances per 100,000 words. Admittedly, the *certain that* collocation is the only one (of the 6 collocates following the word *certain*) that indicates 'certainty': however, if we note its frequency (1.2 counts per 100,000 words) and compare this figure with the 5 other collocations, **combined** they come to a total of 1.05 per 100,000 i.e. less than the lone 'certainty' collocate.

With reference to the social science texts, Biber et al's view about the typical meaning of certainty (noted above) seems correct, but one might assume that native speaker intuitions would be closer to fiction because of the higher dialogue content etc., (see more on this in chapter 2). In sum, the 'certainty' meaning of *certain* is far from rare in fiction; and can only be described as such if we seek to look at the number of different collocation patterns which Biber et al. find (5 of the 20 collocations give the word its 'certainty' meaning) rather than the actual number of corpora-cited instances in these categories.

Kennedy (1991), in his study of *between* and *through* (and their semantic differences), begins his study by quoting native speakers on their thoughts about these words' similarities and differences. He found that the locative meaning of the words seems to be uppermost in most native speakers' minds. He also notes that grammars and teaching materials have identified a variety of functions of these words and that there are

"considerable differences" (1991, p.96) in the books about which functions should be highlighted. After completing his statistical analysis of the LOB corpus, detailing these words' functions and their relative frequencies in corpora, he notes that the locative senses of the words are "quite frequent" (ibid, p.109), but that there are a number of other functions, for example in comparison, agency and causation; however, for *between,* he states that "non-locative uses constitute a majority of the tokens" (ibid, p.109). While this is true, his data indicate that the locative use has the greatest percentage of attestation in his corpus. This somewhat weakens his claim about the "possible arbitrariness and unreliability" (ibid, p.110) of intuition about the meanings of these words.

3.3. Semantic prosody and pragmatics

Louw (1993) argues that intuitions are weak in the area of semantic prosody and pragmatics. He defines semantic prosody as, "a consistent aura of meaning with which a form is imbued by its collocates" (1993, p.157). There are problems with this definition, as discussed by Whitsitt (2005) – particularly the idea that words 'instil' other words with meaning. However, in this discussion I shall pursue Louw's argument, as a number of studies challenging the quality of intuition have come out of his work.

Louw believes that Sinclair's study (1987) of the phrasal verb *set in* and its (negative) subjects (e.g. *rot, disease* etc.) was the first time a corpus was used to uncover semantic prosody, indeed he questions whether the existence of semantic prosodies was really appreciated before the advent of corpus linguistics (1993, p.173), further suggesting that semantic prosody is one of the areas where human intuition about language is distinctly lacking (ibid, p.173). In an effort to substantiate his claim, Louw invites his readers to provide as many phrases, or collocates as possible containing *'....without feeling....'* and then refer to the concordance lines he provides so that the results can be compared. The challenge is an interesting one, and yet had he actually provided data from native speakers along with his concordance lines, his case that the semantic prosody of *without*

feeling is not consciously known by native speakers would have been made much more effectively[16].

But what of Louw's claims about the 'discovery' of semantic prosody through corpus research? As Partington (2004) points out, the idea that this is a revelation available to us only through corpus data and concordance lines is probably an overstatement, for pre-corpus dictionaries note that words such as *commit* and *perpetrate* have unfavourable collocates (p.155). Fox (1998) also acknowledges this overstatement, arguing with regards to *break out* and its negative prosody that, "Information of this kind about words is not necessarily new. But corpus evidence allows us to make statements with greater confidence than we could if we had to rely totally on our own intuition" (p.30). Partington (2004, p.153) makes the important observation that prosodies are rarely as strong as that seen in the *set in* example provided by Sinclair, indeed he notes that prosody may, for some words, be genre specific[17].

Channell's (2000) work is centrally concerned with positive and negative evaluation in the choice of particular phrases and expressions. She is convinced that intuitions are unable to detect the pragmatic force of words or phrases, arguing that "many pragmatic phenomena...are not accessible to introspection" (ibid, p.40); "evaluative polarity is not usually accessible to intuition" (ibid, p.41). Channell acknowledges that it is not clear why it should be the case that intuitions are weak in this area (ibid, p.55), but provides concordance data to support her comments about the polarities of various words, for example, that *par for the course* is overwhelmingly negative, *off the beaten track* is positive, *out in the sticks* is negative etc.

Channell forwards two justifications for her claim that such prosodies are indeed inaccessible to intuition. The first is that they have not always been captured by

[16] Examples from his concordance lines include: *without feeling guilty, without feeling tense, without feeling foolish* and *without feeling embarrassed.*

[17] Partington (2004) notes that *lavish* in the press is usually used in a negative way, but that in the entertainment field it is usually 'neutral-to-good' (p.153).

lexicographers[18]. For example, she notes that the Oxford Advanced Learners' Dictionary (OALD, 1995) defines *right-on* without giving it a negative evaluation, and her corpus indicates that this is its typical prosody. As noted earlier, she states that *par for the course* is typically negative and the same dictionary fails to note this. Her second justification is made in the context of her description of the negative prosody of *regime*. She comments, "I have shown these data [on *regime*] to several hundred people in different audiences and it seems that while people readily accept (and add to their stock of conscious knowledge) that the word *regime* is negative, many report that they had not consciously realized it until they saw the data" (2000, pp.45-6).

Nevertheless, Channell **does** believe that intuitions can detect polar prosody at times. Regarding the word *fat*, Channell comments that, "In a British context, it is clear that the word *fat*, because of the learned prejudices of British culture in regard to body weight, is neither a neutral descriptor, nor a compliment" (2000, p.41). Importantly, Channell gives this commentary before providing us with the corpus evidence. Similarly, with regards to *self-important*, before looking at her corpus data Channell comments, "It is hardly necessary to show examples to convince readers of the awful disapproval with which British English speakers use the expression *self-important*" (2000, p.43). Channell is not inconsistent in her claims: she states that prosodies are **sometimes** not appreciated: however, she does not seek to explain why this is the case. In addition, it should be noted, in fairness to the OALD, that some of the collocates it provides for the word *regime*, concur with the corpus used by Channell and do appear implicitly negative (at least for democratic nations), e.g. *fascist, totalitarian, military.*

There is a tendency among corpus linguists to suspect pre-corpus dictionary entries because introspection or small-scale collections of data have played a large role in influencing what material has been included. A logical corollary of this position is to 'trust' dictionary entries written on the basis of findings from large corpus data. However, this position is flawed as it fails to take into account the important intermediary role of the lexicographer – the 'go-between' between the corpus data and the dictionary,

[18] This case is also made by Sinclair (2004, p.142) with reference to *budge* and the LDOCE.

who, it should not be forgotten, also has intuitions about typicality. Summers (1996) notes that lexicographers using corpus data do use their intuitions, and cannot blindly follow corpus data, because of the 'oddities' that a corpus can contain (p.266), and Cowie makes a very similar point (1981, p.224). However, the best evidence that can be forwarded to challenge the view that post-corpus dictionaries are the best mirrors of real language is the omission in these dictionaries of certain detail about semantic prosody. Stubbs (1995a, p.27) notes that the Collins COBUILD dictionary, with Sinclair at its helm, provided a neutral definition of *cause* though it has a strong negative prosody, and that the same dictionary also failed to record *cronies* as pejorative (2002c, p.72). The fact that these prosodies were not highlighted, though corpus data was used to help make the entries, should temper corpus linguists' criticisms of how pre-corpus dictionaries define words, or list meanings: dictionaries depending on corpora can have similar omissions because of the role of human judgement.

In his discussion of the phrase, *didn't mean to* _____, Beaugrande (1999) notes that, "My unaided intuition failed to anticipate the significant pragmatic and performative constraints that the collocation almost always carries pejorative and apologetic connotations" (p.253). In addition, he notes similar limitations of his own intuition in identifying verbs which typically collocate with *couldn't help* _____ expressions – e.g. *noticing, thinking, wondering* (ibid, pp. 249, 250). He comments, "my intuitions could certainly not have predicted, but could "retrodict" by noting that these Verbs represent Processes which might well be judged not properly subject to conscious control and which might lead to emotions, perceptions, and thoughts people might feel self-conscious about" (ibid, p.250). While these comments of Beaugrande should be interpreted as legitimate attempts to document corpora and intuition differences in the realm of prosody, it is unfortunate that Beaugrande does not actually clearly document his intuitions before expressing surprise about what the corpus reveals, for it is all too easy to express surprise about the contents of corpora without putting in the intuitive work beforehand, with which to compare the findings.

In a detailed study of the word *cause*, Stubbs (1995a) makes a passing comment concerning native speaker intuition and corpus data with regards to the dominant prosody of this word (which, as noted above, is negative). He comments that in his informal testing of native speakers about this word "one or two" of the collocates provided by some of the respondants were "unpleasant collocations, but such native speaker data are very sparse and unreliable indeed" (ibid, p.26). Stubbs used the LOB corpus (a 1 million word corpus) in his study of the prosody of this verb and found that around 80% of occurrences of *cause* are with negative (noun) collocates, 18% with neutral ones and 2% with positive ones. Stubbs finds confirmation in his findings by appealing to the 120 million word COBUILD corpus (ibid, p.42). It is unfortunate that Stubbs does not elaborate any more on his informal study for it would give much greater support to his belief that, "It is...well known that attested data are required in collocational studies, since native speaker intuitions are not a reliable source of evidence" (ibid, p.24). Work recently conducted by Nordquist (2004), which actually required 25 respondants to use the verb *cause* in a sentence, found that 70% of the sentences had a negative semantic prosody (similar to Stubbs' 80% finding in his corpus). This suggests that respondants do have 'prosodic knowledge' which informs their choices of the collocates of words. Further support for the idea that semantic prosody is consciously known comes from the research of Greenbaum (1988), about the collocates of *utterly*, as discussed in section 3.1, in this chapter.

It seems that the case against semantic prosody intuitions is far from proven, and rather anecdotal in nature. General claims, in particular, seem to lack empirical justification and support.

3.4. Phraseology

With regards to lexical intuition weakness in the area of phraseology, Hunston (2002) has in mind the inability to explain why certain things are atypical: for example, why some verbs do not fit into grammatical patterns. Her reference to Owen's difficulties with the phrase *require to be done* supports her position. Owen (1996) looked at the (hypothetical

non-native speaker) sentence "Many more experimental studies require to be done before we can say that...." (1996, p.222). He hypothesizes that a native speaker teacher would correct this and offer alternatives, (i.e. substitute *require* with *need*, or rewrite *experimental studies are required*). However, Owen notes that if the student were to check the COBUILD corpus there are examples of passive *require to be...*, and he notes how the concept of 'normalcy' (an intuition) is as important as corpora, in dealing with this case, i.e. the usage is unusual, though it has corpus support. Hunston picks up on this, but argues that Owen's problem can be solved on phraseological grounds, namely that "the past participle that follows [REQUIRE to be] is usually that of a verb with a specific meaning, not a general verb such as *do"*, offering *pruned*, as an example in the sentence *These roses require to be pruned each spring* (2002, pp. 21, 22). Hunston believes that Owen was not consciously aware of this 'fact' i.e. that he lacked sufficient conscious awareness of this usage pattern rule.

However, Hunston's position is weakened somewhat in comments she makes later about marking a student's paper. The paper stated that an author 'is under the influence of Halliday' (Hunston 2002, p.214, emphasis mine). She found this phrase odd, and, relying on her intuition, made a note to the student commenting that the phrase *under the influence of* was used only for 'bad' things like alcohol and drugs. She checked corpus data which confirmed her intuitions about these collocates being highly frequent, but she was surprised to find that there were examples of people being under the influence of other people in the corpus data. Hunston provides concordance data of forms expressing the notion of a person *having been* under the influence of another (e.g. *...girl was said to have been under the influence of an older woman*) and also of people *coming* under the influence of others. However she fails to provide an example of the type 'person X is under the influence of person Y'. This suggests that her intuitions may have been more accurate than she thought – the phrase 'Person X is under the influence of Person Y' is odd and she forwards no corpus evidence to substantiate its usage.

Fox (1987) also argues that native speakers lack phraseological awareness, in particular that they are unable to produce typical missing words from phrases. Fox believes that

when presented with *a torrent of NOUN*, intuition (presumably her own) suggests that *abuse* would be the most typical noun collocate filling the slot (p.139). She checked this phrase in the COBUILD corpus[19] and found that the noun provided by her intuition was too general: more specific nouns had more corpus support, namely: *outrage*, *confession and explanation* and *invective*. She argues that there are, "far more [concordance] lines [for these specific examples] than there are for 'a torrent of abuse'" (ibid, p.139). What are we to make of this case? Stubbs (2001), in discussing what is possible (in language), attested (in corpora) and probable, states that, "Corpus linguistics is not concerned with what happens to occur (at least once)...it is concerned with a much deeper notion: **what frequently and typically occurs'** (p.151, emphasis mine). Fox's example phrase *a torrent of NOUN* occurs 76 times in the BNC, a corpus of 100 million words, at least five times the size of Fox's corpus[20]. Though Fox doesn't actually say how many examples she was dealing with, extrapolating down from the number of instances in the BNC (on the dangers of this see chapter 2, section 7.1) we can suppose that there were no more than 20 instances in her corpus[21].

Of the 76 concordance lines found for *a torrent of*____ in the BNC, on 9 occasions it is followed by *abuse*, on 3 by *verbal abuse*, and once each by *personal abuse*, *foul mouthed racist abuse* and *the vilest abuse imaginable*. So there are actually 15 instances of *abuse* as a noun head following *a torrent of* in this corpus. In contrast, there is not one single line for any of Fox's examples. *Abuse* is the single most frequent noun following *a torrent of*___ in the BNC, which suggests, contra Fox's findings, that her intuition about the most common noun to fill this slot was, actually, correct[22]. Of course, the BNC corpus is different to the corpus used by Fox, both in content and size. However, this is an interesting case, for it shows how an experienced lexicographer was quick (perhaps too quick) to abandon her intuitions on the basis of 'marginal' data. The alternative position, as noted earlier on in the chapter, would have been to hold on to that intuition,

[19] Unfortunately it is unclear whether she refers to the main corpus (7.3 million words at the time according to Sinclair 1987, p.150; Renouf 1987a, p.171) or the main *and* reserve corpus combined, totaling 20 million words at the time (Renouf 1987b, p.12).
[20] 13.6 times if she used the smaller 7.3 million corpus in her research.
[21] There would be approximately 6 instances if she used the smaller corpus.
[22] Hanks (2004) terms *a torrent of abuse* "a stereotypical phrase" (2004, p.246).

and question the corpus. This interesting case highlights the need to be even-handed in our treatment of both corpus data and intuition, and not dismiss one or the other as 'unreliable' too quickly.

Sinclair (1997), also, seems reluctant to concede a role for intuition in the matter of phraseology. In particular he argues against the ability of intuition to pick up the typical function words 'around' an expression. The example he provides is *naked eye*. He comments, "A preposition in front of *the* is a safe bet, but it is quite likely that a person will retrieve one of them (e.g. *with*) and forget the other one (*to*)" (1997, p.34). He does not explain why. Perhaps it is because *with the naked eye* is a unit (a prepositional phrase) whereas *to the naked eye* is not: it requires *visible* to make it an adjective phrase containing a prepositional phrase. However, to speak of probabilities about intuitions is not enough: data is required.

There are, on the other hand, writers who argue for the existence of intuition ability in the provision of missing words in phrases. Regarding *PREP the NOUN of the NOUN* phrases Stubbs (2002b) argues that, "Native speakers can make intuitive judgements as to which words would be acceptable in the frame" (p.233). Stubbs is not saying here that the most frequent words can be provided, but simply that filling the slots with suitable candidates (e.g. *at the end of the day*) is not particularly difficult, and as such this does not constitute strong counter-evidence to challenge phraseological blindness. It is Mackin's (1978) work, in particular, which seems to provide evidence for the sort of ability that Hunston, Fox and Sinclair question the existence of, though it should be noted that this is not a corpus based study. Mackin (writing in the context of how items were chosen for inclusion in the Oxford Dictionary of Current Idiomatic English), describes how, as part of the dictionary making process, he asked 10 university graduates to fill in the missing word or words in various phrases which provided very little context, and which the compiler had chosen for inclusion in the dictionary because he believed that they would be completed in a predictable way. Mackin records that for the first fifty items (requiring the provision of one word to fit the slot, e.g. *for old times*_____), a typical score was 47/50 i.e. there was strong agreement with the compiler's judgement. For the second fifty

(requiring the provision of 2 words, e.g. *He told her off in* _____ _____ *terms*) he noted that agreement with the compiler's judgment was forty-plus from 50. He comments, "in nearly every case there was a consensus of opinion regarding the most likely ways of completing the expression" (1978, p.157). It would be interesting to be able to check these responses against corpus data, but unfortunately there is very sparse data in the BNC for some of the phrases. For example, there is only one example of *all dressed up and nowhere to go*, two examples of *screamed blue murder* and no examples of *take it at its face value* in the BNC[23]. While it would be possible (theoretically) for a corpus to provide us with data contradicting the frequency of the intuitively provided words missing from Mackin's phrases, e.g. that *he told her off in very strong terms* is more common than *he told her off in no uncertain terms*, for this to be 'established' there would need to be a large number of examples to make a valid comparison – just how big such a corpus would have to be is a matter of speculation. Mackin's work is a reminder that intuitions, at times, are unassailable. If we take to heart the comment that corpus linguistics has nothing to say on whether things are possible or not, but only about whether they are frequent (as noted above), then this should temper broad claims excluding a role for intuitions in the description of language. As Howarth (1998a) notes, "it must be recognized that decisions about the acceptability of combinations that occur individually at very low frequencies must continue to rely heavily on human judgement" (p.29).

3.5. The true place for intuitions

Before moving on to the explanations forwarded by corpus linguists to explain why they believe intuitions differ from corpus data, it is worth noting that corpus linguists do see a place for intuition, but not in any of the above four mentioned areas. In general terms, intuition is 'relegated' to the role of non-data providing activities, i.e. deciding upon what should be studied, designing the analysis, and interpreting/confirming the findings (see Stubbs 1995a, p.48, 1995b, p.388; Beaugrande 1999, p.247; Sinclair 1991a, p.39;

[23] These are all phrases used in Mackin's study.

Tognini-Bonelli 2001, p.91). Nation (2001, p.56) argues that without intuition being utilized in the above ways corpora are of little use.

In addition, it is important to note that while corpus linguists have criticized first person introspection judgments, they believe that, retrodictively, corpus evidence is confirmed by intuition (e.g. Louw 1993, p.173; Francis & Sinclair 1994, p.191; Fox 1987, p.146). It is unfortunate that this ability has not actually been 'tested', as it is not very hard to imagine respondants agreeing with a reasonable option forwarded by a person 'in the know' (i.e. who has checked the corpus data) about what is or is not frequent/typical. For example, let us imagine that a respondant, asked to provide a frequent noun collocate for the adjective *appropriate*, provided the noun *conditions* (there are 11 instances in the BNC of *appropriate conditions*). If the respondant were then told that *measures* is the most frequent collocate, he or she may well nod his or her head, concur, and say, 'well, yes, of course, why didn't I think of that!' (there are 40 instances of *appropriate measures* in the BNC). Actually though, according to the BNC, the most frequent noun collocate is *action* (115 instances). 'Retrodictive' ability should be tested, not assumed, and the testing of this ability is incorporated into the design of the experiments reported in chapters 7 and 8.

One final area in which corpus linguists concede a position for intuition is the ability to define 'lone word' meaning. Summers (1988) is happy to label teachers and parents as "excellent...dictionaries" (p.113), in the context of the help that they give to children looking for explanations of new words, and Sinclair echoes this belief (1997, p.32). In no way is Sinclair inconsistent in saying this, for this belief can sit quite comfortably alongside the idea that intuitions are not good in relation to usage.

4. Explanations for data differences

On more than one occasion Stubbs specifically calls for research into the disparity between intuitions and linguistic data (e.g. 2001, p.168; 2002a, p.10; 2002b, p.226). As noted above, it is unfortunate that so little research has been conducted in this area, and

so the opinions of corpus linguists noted below should perhaps be viewed more appropriately as 'hypotheses', rather than post-experimental 'explanations'.

4.1. Connotation, denotation and delexicalisation

Stubbs stresses the importance of context in determining word meaning (1995b, p.381), citing with approval Hunston & Francis' comment that "most words have no meaning in isolation, or are at least very ambiguous" (2000, p.270). As noted earlier, Stubbs (1995a, p.24) argues that our intuitions about collocations and dominant semantic prosodies are weak. If context establishes meaning, and we are indeed unable to provide typical collocates (which includes a knowledge of prosody) then, logically, it follows that our intuitions about typical word meanings and uses will differ from corpus data. Stubbs emphasizes the significance of the denotational meaning of a word in affecting our ideas about word meaning, but notes that often, idiomatic, rather than literal uses of a word may be more frequent (2002b, p.221). He also argues for connotation blindness, i.e. Stubbs believes that native speakers are not consciously aware of the connotative meanings of (apparent) synonyms such as *little* and *small*, i.e. *little* has cuteness connotations and *small* does not; *small* has pejorative / derogatory associations (e.g. *small man*), whereas *little* does not[24]. The idea of delexicalisation is also important for Stubbs in explaining why intuitions may differ from corpus data. He observes that frequent words typically undergo delexicalisation, i.e. lose their 'meaning' (presumably dictionary meaning) in usage. For example, he notes that *way* is typically used in phrases where its meaning is delexicalised, i.e. bears little or no relationship to the '*path* or *track*' meaning (i.e. dictionary meaning) as used in the phrases *the other way round, the correct way of holding it* etc. (2002b, pp.228, 229). Stubbs also believes that "delexicalisation is a logical consequence of [a word's] frequent use in phrases, where meaning is dispersed across the phrase as a whole" (2002b, p. 230). As a result of this, the idea of word

[24] See also Stubbs (2002c, p.167) "Work on recurrent collocations suggests that many more words have evaluative connotations than is often realized".

meaning is diluted, and in such cases a word may have a purely supporting function e.g. the word *take* in *take a decision* (Stubbs 1995b, p.381)[25].

Sinclair (1991a) argues, in a similar way to Stubbs, that frequent words in particular are prone to being delexicalised[26]. For Sinclair, this is a key reason why intuitions about typical word meanings differ from corpus data. He states that, "The 'core' meaning of a word -- the one that first comes to mind for most people -- will not normally be a delexical one. A likely hypothesis is that the 'core' meaning is the most frequent independent sense" (Sinclair, 1991a, p.113). This seems a reasonable explanation, but it would have to be tested (as Sinclair himself acknowledges, ibid, p.113). However, how it is tested is critically important. If a subject were asked what *take* means it would only seem natural for a person to ascribe meaning to that word (denotative, dictionary meaning), rather than bleach the word of its meaning, as happens when the word is delexicalised in usage; i.e. in a situation where the question is not about cotext or usage, but meaning, one should hardly be surprised if someone gives the meaning of the word in isolation: 'the dictionary meaning'. If though, computer software were used to investigate the most frequent meaning of a word in a corpus, then cotext, patterns, and collocations would inform the result. It would be questionable practice indeed to compare such findings with respondants' definitions of words. A single word cannot be delexicalised – there is nothing for it to be delexicalised across. So, can we be expected to give delexicalised meanings to words in isolation? To enable a 'fair' comparison of corpus data and elicited data, respondants must be allowed to access cotext (e.g. provide collocates) before providing the meaning, or be allowed to produce a sentence containing the word[27] under investigation. In such a scenario the resulting intuitions can (legitimately) be compared with corpus data, because both 'searches' access cotext first and provide meaning later. For example, if when asked to provide a high frequency noun collocate for *take*, a respondant provided *care* or *advantage*, then the respondant is aware of a highly frequent collocate (according to the BNC). In such uses (*take care, take*

[25] See chapter 3, section 2.3, for more discussion of delexicalised words in collocations.
[26] Kennedy (1991) may have the same thing in mind when he argues that our ideas about the typical functions of frequent words will be inaccurate, because of their complex semantic structure (p.97).
[27] A methodology adopted by Nordquist (2004) and Gilquin (2005a and b) to elicit data.

advantage) *take* is delexicalised. If, after taking part in such a task, the respondant were then asked to give his/her judgement on the most frequent meaning of *take* the respondant could 'correctly' answer, that it has very little independent meaning **in such contexts**. If, on the other hand, a respondant provided *a book* as a highly frequent (right) collocate of *take*, then, in such a case we can fairly confidently say that the denotative meaning of *take* seems to be driving the response. In sum, I argue that we cannot test notions about typical meaning unless we allow respondants to provide cotext/collocates first. This is the procedure adopted in the research reported in this thesis.

Perhaps though, one might argue that the provision of 'lone word meaning' includes (of necessity) the accessing of some cotext. Malmkjær (1993, pp.228, 229), for example, argues that when asked about the meaning of a word, some cotext must be accessed before a meaning is provided, otherwise we could only respond 'x' to the question 'what does x mean?' But is this really the case? If when asked to explain the meaning of *small*, or to provide a synonym for *small*, the word *little* is provided, it would perhaps be equally valid to argue that a semantic feature of the lone stimulus word, rather than typical cotext instances influence the response. But on what basis would a particular semantic feature become the most typical/dominant, if not through its usage? A key factor may be saliency.

4.2. Saliency

It has been argued that saliency is a cognitive counterpart of prototypicality (Williams 1992, p.208; Radden 1992, p.519-520). With regard to the word *see*, Sinclair & Renouf (1988) argue that the meaning that people associate with this word, i.e. the salient meaning is 'seeing through one's eyes' rather than 'understanding' (which, they argue, is the most frequent 'meaning' because of the use of the word in the very frequently used phrases *I see*, and *you see* (1988, p.152)[28]. In contrast to Sinclair's 1991a explanation

[28] Sinclair & Renouf's *see* example can be criticized. In a BNC search the instances of *see p., see pp. and see page*, if combined, come to 12637 instances. There are 11655 instances of *you see*. In addition, it seems rather strange that Sinclair defines *see* as 'know', when actually this is not the case. It is peculiarly in the cases of *I see* and *you see* that the **phrase** means 'I / you understand'. *They see*, in contrast, does not appear

(that lone word meaning determines notions about typical use), Sinclair & Renouf explain the above 'mismatch' by appealing to Lyon's (1977, p.247) argument that saliency (either biological or cultural) affects what is noticed about a particular word. Whereas delexicalisation is a linguistic explanation, saliency is a psychological one. Such saliency might be universal, in the case of *see* for example, which Malmkjær believes is biologically salient, or, one might assume, it could be environmentally or culturally determined, and be particularly evident when polysemous words are provided as stimulus words. A different meaning may be more salient for different types of people: for some people the word *interest* may be a 'finance' word, and for others it may be a 'hobby' word. Such differences in saliency might affect notions of typicality. To give a rather 'extreme' example we could expect a mismatch in the intuitions of a cohort of Californian surfers asked about the typical collocates of *surf*, when their intuitions are compared to a corpus composed of books about the Internet. The idea that saliency and frequency of exposure are somehow interconnected is made by Giora (2002, p.491) who argues that saliency, "is a matter of degree, determined primarily by frequency of exposure and experiential familiarity with the meaning in question". However, as soon as we open up the possibility that saliency is determined environmentally, then there is a problem. The 'problem' is that if frequency of exposure is a key criterion in affecting saliency, and if the corpus is 'representative', and the respondant is 'representative' (i.e. has been exposed to material comparable to that in the corpus) then we should not find that intuitions and corpus data differ. As Malmkjær (1993) notes, it is difficult to explain why there is, at times, an apparent conflict between saliency and frequency. This subject is examined in more detail in chapter 4.

In the above section it has been noted that some corpus linguists have forwarded two different factors to explain the differences between intuition and corpus data with regards to meaning and usage: the favoured status of the independent form (i.e. denotational, non-delexicalised form of a word), and/or the psychological saliency of a particular meaning of a word driving intuitions about its meaning/collocates.

---

to have the same meaning: a brief scan of the BNC concordance lines suggests that it means 'they look at / observe'. Sinclair is, then, somewhat inconsistent in arguing generally for the importance of delexicalisation but, here, that a particular very frequent word has a particular meaning.

5. Summary

The discussion in this chapter demonstrates that, by and large, corpus linguists have not provided an overwhelming case against lexical intuitions in the four areas of collocation, word frequency, semantic prosody and phraseology[29]. The 'discovery' accounts, in which writers express surprise about corpus data are of questionable scientific value because no rigorous recording of intuition pre-exposure to corpus data has been provided. Similarly, the appeal to textbook material can be countered, in that textbooks have a different agenda to follow than just the recording of frequent uses/meanings. As noted, prosody omissions in corpora-based dictionaries highlight the vital intermediary role of the lexicographer, and as such it is not necessarily the case that corpus-based dictionaries will be any more reliable than their pre-corpus cousins.

It is not enough to say, as Barlow does (1996, p.6), that our intuitions will 'probably' provide certain verbs as collocating with a certain word: these intuitions must be documented, otherwise, we are recording our intuitions about intuitions. There is, then, a real need for empirical research to investigate intuition-corpus differences in the areas discussed above: collocation, frequency, prosody and phraseology. The research reported in the remainder of this thesis touches on all four areas, though the main focus is on frequency and collocation.

[29] Comments have been made on other issues in the literature, e.g. N. Ellis has made comments on intuitions and the use of tenses in conversation (Ellis 2002a, p.317); however, this subject is not so directly related to the lexical focus of the present study.

Chapter 2 – Corpus Representativeness: Establishing the Parameters of Usage

1. Introduction

As noted in chapter 1, the goal of this research is to compare lexical intuitions and corpus data in the area of collocations. In this chapter six objections against the validity of this enterprise are formulated. Rather than just stating these objections and examining them, the discussion below attempts to find areas where these objections are less valid. The reason for this is simple: corpus data will be used in the research, but, how it can best be used must be determined by addressing problems such as those noted below.

Saying that intuitions differ from corpus data is one thing: saying that they differ from 'language' is a much harder claim to make, and yet it should be noted that the underlying assumption in much of what corpus linguists have said against lexical intuitions, is that the corpus data is an accurate reflection of 'language'[1]. Corpus linguists have made the step from corpus data to 'language' look like a small one, but it is not. It is a step that can only be made if the corpus is 'representative' of the language that it purports to sample[2], and, as noted below, the concept of representativeness is not well understood in the field of corpus linguistics. Rather than jump to the conclusion that lexical intuitions differ from corpus data because intuitions are defective in some way, the possibility that the corpus against which intuitions are measured may be 'inadequate' must also be acknowledged. This is one of the key issues that is addressed in this chapter.

---

[1] Note the underlying assumptions in the comments in chapter 1, section 1. While Hunston recognizes the problems with extrapolating corpus data to language: "A statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample' (2002, 23), she is not consistent on this point: "Intuition is a poor guide to at least four aspects of **language**" (Hunston 2002, p.20 emphasis mine). If corpus data is being used to establish the latter point (which it is) then, implicitly at least, Hunston is making the step from corpus data to language.

[2] Interestingly, some writers have made representativeness a part of the definition of corpus. For example, Francis (1982) defines a corpus as "a collection of texts **assumed to be representative** of a given language, dialect or other subset of a language, to be used for linguistic analysis" (p.7, emphasis mine), whereas others e.g. Kilgarriff & Grefenstette (2003) have questioned that a corpus needs to be representative, arguing that this confuses the issue about what is a corpus and what is a good corpus (p.334).

There are other problems too which must be addressed before any rigorous empirical investigations into corpus data and intuitions can be carried out. Research suggests that the distribution of words and collocations differs across different registers. If this is so, then against what register(s) should intuitions be compared? Further, it has been suggested that even large corpora can indicate different things about the relative frequencies of fairly common words in the language. If this is so, then this is another obstacle to overcome. Should intuitions be compared with corpus X or corpus Y in such cases, or should the enterprise be abandoned?

2. Objection 1: Different registers different data

It makes no sense to compare lexical intuitions with corpus data as the distribution of words and collocations across different registers varies. Against what then can lexical intuitions be compared?

The idea that we can compare lexical intuitions about word combinations with 'general' English language data as captured in a large mixed corpus is open to criticism, for the simple reason that, in the words of Biber et al. (1996) "there is no single register that can be identified as 'general English'" (p.129). It could be argued that language from a large mixed corpus can, by virtue of its sampling of different registers, be called 'general'. However, this does not see off the real problem, namely, "corpus-based analyses show that linguistic association patterns are generally **not** valid for the language as a whole" (Biber et al, 1996, p.117 emphasis in original; see also Biber 1994, p.186). If this is the case then problems may well arise in comparing lexical intuitions with corpus data. For example, if respondants are asked to produce frequent collocates of a word, collocates might be produced which are frequent in spoken language, but not in the large mixed corpus, or are frequent in fiction, but not academic writing. While Biber and his associates have said much about differences of lexical associations and word frequencies in different registers and across different media (explained in greater detail below), it can also be argued that there is a certain degree of commonality in language – across register

types and media types[3]. This being so, it is defensible to focus on this 'generic' material when designing research into lexical intuitions about general English usage. Provided the position is valid, it will be possible to sideline, to a certain extent, the fact that there are indeed important language differences in different registers. The following discussion concerning this first objection is divided into 5 areas where register/media play a part in affecting the language present in a general corpus.

2.1. Word Frequencies

The fact that words are not spread evenly across the wide variety of register types seems hardly surprising. Our experience of language tells us that we are unlikely to see the words *happy* or *glad* in an academic journal, and are more likely to see them in a novel; and such is one of the many findings of Biber (1996, p.176). Biber, together with his associates, has conducted extensive research investigating differences in words, word frequencies and collocations across different register types. In a 1994 paper Biber et al. note, for example, how the 'apparent' synonyms *certain*, *sure* and *definite* have different distributions in social science and fiction texts in the Longman/Lancaster corpus. *Sure* is around four times more common in fiction than in social science texts; *certain* is found twice as often in social science texts than in fiction texts; and *definite* is around ten times more common in social science texts than in fiction texts (1994, pp.175, 176). Differences tend to be more apparent when the corpora or text types being compared contain noticeably different language. For example, a comparison of the most frequent nouns in the COBUILD corpus (a large mixed corpus) and a biology corpus, reveals that there are no common items in the top 20 noun frequency lists of the two corpora (Gavioli 1997, p.87). Clearly then, the notion of 'word frequency' requires, at times, more precision, and the fact that general frequency figures about words (from a mixed corpus) mask individual register differences has been noted by Kilgarriff (1995, p.4). However, the 'masking' can be uncovered, for as Kilgarriff goes on to note, there are ways of measuring the 'burstiness' of words, i.e. checking their distributions throughout different

[3] See the discussion in chapter 7, section 2.3.2 and chapter 8, section 2.3, which provides empirical support for this position.

documents and within documents in a corpus. *Word frequencies in Written and Spoken English* (Leech et al., 2001) a book based on the British National Corpus (BNC), not only indicates how many text sectors include a word in the BNC (termed the 'range'), but also provides a dispersion ('burstiness') statistic (Juilland's D), which enables us to see how evenly distributed a word is across sectors of the corpus. The dispersion statistic is a more useful tool than the range statistic in that it can indicate more accurately how widespread and 'general' a word is (Leech et al. 2001, pp.18, 19). It has been noted that grammar words seem somewhat immune from 'text type' effects (Hunston 2002, p.3), and the range and dispersion statistics of the words *the, a* and *of* from the BNC substantiate this, as indicated in Table 2.1 below. However, it should be noted that it is not only grammar words that are uniformly represented across the BNC. Content words, particularly frequent content words, can also have a broad range as indicated in Tables 2.2 and 2.3 below. Also, it should be noted that fairly frequent phrases may be quite evenly spread across different language types (Table 2.4). In addition, a fifth table is provided (Table 2.5), giving statistics about some words that are not well distributed throughout the corpus. In the tables below the maximum range possible is 100, and the maximum score for Juilland's D is 100. The data is from Leech et al. (2001).

Table 2.1. Dispersion statistics of some common grammar words in the BNC

Word / no. of instances per million words	Range	Juilland's D
The (61847)	100	98
A (21626)	100	99
Of (29391)	100	97

Table 2.2. Dispersion statistics of some frequent nouns (+100 / million) in the BNC

Word / no. of instances per million words in brackets	Range	Juilland's D
News (145)	100	94
Body (255)	100	95
Office (300)	100	96

Table 2.3. Dispersion statistics of some infrequent nouns (10 or less / million words) in the BNC, with quite wide representation

Word / no. of instances per million words in brackets	Range	Juilland's D
Boost (10)	94	86
Waiter (8)	79	85
Pillar (5)	93	90

Table 2.4. Dispersion statistics of some multiword items in the BNC (various frequencies)

Word / no. of instances per million words in brackets	Range	Juilland's D
With regard to (17)	94	91
For example (239)	100	92
And so on (49)	99	92

Table 2.5. Dispersion statistics of some items *not* well distributed in the BNC.

Word / no. of instances per million words	Range	Juilland's D
Polymer (7)	43	45
Parameter (6)	49	65
Mucosa (11)	15	35

The above data are interesting in that Tables 2.2-2.4 seem to indicate that, in addition to grammar words being fairly genre independent, certain nouns (frequent ones more so than less frequent) and also certain multi-word strings are fairly 'generic'. These findings should caution us from adopting a 'default difference' approach to word/phrase distribution across text types.

If we are seeking to compare lexical intuitions with corpus data, it would seem to make sense to provide stimulus words that are fairly well distributed throughout a corpus, in order to minimize the effects that exposure to different register types may have on people's lexical intuitions. This approach has been adopted in word frequency estimate research[4]. For example, Ringeling (1984) when discussing how he chose words for inclusion in the word frequency ranking task that he conducted comments, "Care was taken that the words occurred in as many of the sub lists that the corpus consists of as possible" (p.63).

As a consequence of adopting the 'uniform range' approach in the present research, it will not be possible to investigate intuitions about infrequent words and their collocations. However, this is not a major problem. As will be noted in chapter 6, section 3, there is existing research about infrequent words and their collocate associates from word association research.

[4] This is discussed in greater detail in chapter 4.

2.2. Word meanings

In Biber et al's. (1994) view, Sinclair's observation that *back* used adverbially is more common than the body part meaning of the word, while true in an absolute sense, is not true in a register sense. They note that in social science the adverbial sense is more frequent; however, in fiction, the body part meaning is more common (1994, pp.174, 175). Many words in English have multiple meanings, and the more frequent a word is the more meanings it tends to have (Gernsbacher 1984, p.271; Schmitt 2000, p.73). Hence, a consequence of the focus on frequent words in this study is that an additional variable is brought into the study - polysemy. In Fox's *feet* experiment (reported in chapter 1, section 3.1), it was noted that the collocates provided by the respondants often gave the word *feet* its body part meaning, as opposed to its measurement meaning (the meaning which had the most frequent collocates attracted to it in the corpus). It may be that respondants are not sensitive to the different frequencies of different polysemes, and, as a consequence, their collocate responses differ from corpus data. If so, how and why a particular polyseme might have greater saliency in the minds of respondants is an important and necessary part of our investigation: polysemy as such is not an obstacle to the research focus.

2.3. Collocations

A number of writers have argued that different types of text reflect different collocation patterns. Among them is Partington (1998), who argues that, "Collocational *normality* is dependent on genre, register and style" (p.17). In a similar vein, Lewis (2000) believes that, "different kinds of text have radically different collocational profiles" (p.186). There is plenty of evidence to substantiate this. For example, Murison-Bowie (1996, p.194) notes that the different meanings of *heart* in romantic and medical texts are indicated clearly by the different collocations (*Heart throb* - romantic novel: *heart failure* - medical

text) and Gavioli (1997, pp.87, 88) notes the differences for the most frequent collocates of *criminal* in newspapers and academic texts of social science[5].

Acknowledging the fact that certain collocations tend to appear only in certain registers does not mean, however, that there is not an element of commonality about collocations too. Three important points should be remembered.

Firstly, it would be wrong to think that collocational prosodies *usually* differ across text types[6]. Stubbs (1995a, p.30) notes that 4 different corpora: LOB + LUND (1.5 million words), a 700,000 written word corpus, a 725,000 word written and spoken corpus, and a 425,000 word corpus on environmental issues, all seemed to show negative prosody for the word *cause*. In addition, he notes the tendency for unpleasant adjectives to follow *get*, and *get* passives to refer to unpleasant events across different corpora (2001). He found, from an examination of the BNC spoken corpus, that *get* passives typically have unpleasant subject-referents (over 60%), whereas *be* passives have them only around 25% of the time. Noting that corpus studies are replicable (2001, p.165) he refers to two other studies that show similar findings: Collins (1996) showed that *get* passives were adversative around 70% of the time; and Carter & McCarthy (1999) concluded that the figure was around 90% in their spoken corpora. While one might wish to make something of these differences, Stubbs notes that "there is no doubt about the direction of the strong regularities which emerge from three independent studies of three independent corpora" (2001, p.165).

Secondly, it should be remembered that certain collocations are simply the 'idiomatic' way of saying something, no matter what the genre/text type. Stubbs (1995b) looked into the collocations of synonyms - *little, small; big, large* - and used a 2.3 million-word corpus of contemporary English and the Oxford English Dictionary (OED) on CD-ROM in his study. He found that the two corpora contained the same kind of collocations, e.g. a

[5] The most frequent collocates (in order of frequency) in newspapers are: *war, act, law*; the most frequent collocates in academic social science are: *law, liability, English.*

[6] There are some exceptions. For example, Biber et al. (1999, p.509), note that *poor* is generally descriptive in academic prose, and emotive in fiction, see also the comments of Partington (2004) on *lavish* in chapter 1, section 3.3.

preference for *little girl*, over *small girl*; fixed phrases e.g. *big toe*; and metaphorical meanings e.g. *big brother* etc. While noting the very different composition of his corpora (both in terms of type of sources and age of sources), he comments, "they differ very little in what they reveal of the behaviour of the...words...If two completely independent and very different samples of the language provide the same results, then this is an indication that these findings are features of the English language as a whole" (1995b, p.380). This fact, that there are commonly acknowledged ways of expressing meaning, should not be forgotten. Any corpus might contain the term *large quantity* but not (typically) *big quantity*[7].

Thirdly, Stubbs (2002b) and Stubbs & Barth (2003) make comments about both the similarities and differences across text-types in the matter of 'chains' of language. Stubbs & Barth, in their study of three different text types (fiction, aesthetic literature and academic writing from LOB/LOB/Brown/Frown), note that there are often differences in terms of the most frequent two, three, four and five word chains in these small corpora. However, they also note that a chain such as *at the end* cannot be used as a text-type discriminator, because of its high frequency and generic use (2003, p.82). They point out that Biber et al. (1999, p.1015) found the chain *at the end of the* to be the most common 5 word chain in academic prose genre. It is also the most frequent 5-word chain in the spoken LUND corpus (Stubbs 2002b, p.233). Stubbs (2002b) compared the most frequent *PREP the NOUN of the* chains in LOB/FLOB/LUND, plus a one million word corpus of written language and a 500,000 corpus of spoken language with the most frequent *PREP the NOUN of the* chains in the BNC (100 million words), normalized to 1 million words. Although he gives no correlation statistic, he comments, "The frequencies are remarkably similar. So, we can be confident that these 5-word chains are not an artefact of the small corpus which I started with, but are frequent 5-word chains **in general English**, though some are more frequent in written genres" (2002b, p.234, emphasis mine).

The above three observations, should, therefore, warn us against adhering to a 'default difference' position in the matter of collocation and phraseology across genre and text

[7] In the BNC data there are 97 instances of *large quantity*, but no instances of *big quantity*.

types as perhaps unwittingly implied by Partington and Lewis. While it is undoubtedly true that certain collocations are present or not present in certain genre/text types[8], it should also be recognized that some collocations and phrases are fairly 'generic': they may be found in quite different genres, and in both written and spoken corpora.

Unfortunately, unlike the case with words (as noted in section 2.1 of this chapter), readily available range and dispersion statistics are not available for collocations. Kjellmer (1994) does indicate the dispersion of the collocations from the Brown corpus in his dictionary, though only by indicating which, or how many of the 15 text categories contain the collocation. For example, *full amount* (occurring 6 times in the Brown corpus) is found only in the 'Press: Reportage' section, but *a good deal* (occurring 27 times) occurs in six of the 15 categories. The fact that this resource is readily available may seem to open up the possibility of identifying which collocations are more generic in (American) English as a whole. However, this is not really the case. As Kjellmer candidly acknowledges, a larger corpus than the Brown corpus (1 million words) would be more representative in terms of the collocational uses of native speakers (1991, p.117) and Granger (1998, p.154) points out that even familiar combinations such as *highly significant* and *seriously ill* are not present in this corpus. A fairly simple check can be conducted when searching and using the BNC for how well distributed a collocation is: the number of documents containing the collocation is given along with the number of instances. For example, there are 47 instances of the collocation *dark matter* in the BNC, but these are from only eight sources. Further checks indicate that just two sources: *Nature* and *The Economist* magazine account for around three quarters of the data. There are, on the other hand, 50 instances of *easy matter* found in 49 sources. When considering the most frequent collocates of a word, an eye must be kept on their distribution. Those which are not well distributed, could, on principled grounds, be excluded from tables drawn up to show the most frequent collocates of a word, when designing lists against which to compare intuitions. For more on this issue see chapter 7, section 2.3.4 and chapter 8, section 2.3.

[8] E.g. Smadja (1994) notes that the rather typical collocation *eat food* is not present in her Dow Jones corpus, as *food* is *traded, sold, offered* and *bought* at the stock exchange, not *eaten.*

2.4. Spoken and written English

How significant is the difference between spoken and written English, as evidenced in corpora? Crystal believes that the difference is very significant, arguing that the differences between writing and speech "go well beyond the contrast in medium" (1998, p.1), citing vocabulary and grammar as examples. 'Spoken language' is a wide term, a monologue being very different from a conversation. Svartvik (1993), when comparing corpora, suggests that planned monologue is closer to writing than to spontaneous speech in terms of lexis (p.22). Even different types of spontaneous speech have been shown to vary in the relative frequencies of their most common words. O'Keeffe and Farr (2003) report the most frequent words used in shop encounters and chatting (i.e. both spoken and spontaneous) in their 2 corpora. They list the 10 most frequent words in both corpora and only four of the words are common to both lists: *I, you, it* and *the*. The relative frequencies of even these words are quite different.

Biber et al. (1996, p.119) have observed differences in collocations in comparing an academic written corpus and a conversational spoken corpus. For example, common collocates of *big* in conversation are *one(s), thing(s), house(s), rooms(s)*, compared to the common collocates in academic texts: *change, increase* and *difference*. While these differences should be noted, it should also be remembered that these two register types (academic prose and conversation) are probably at opposite ends of the spectrum in terms of the kinds of lexical material they contain, indeed Biber et al. (1996, p.128) acknowledge this.

Ideally, if given the choice, a comparison of lexical intuitions with spoken data, rather than written data, would seem preferable for two reasons. The first, is that the majority of people probably speak more than they read (Kilgarriff & Grefenstette 2003, p.341; Svartvik 1993, p.15), and so one might expect intuitions to be more directly affected by spoken language. The second reason is that it may be the spoken language in particular that helps us understand how language is organized (Sinclair 1991a, p.16). Unfortunately, spoken corpora are still very small (compared to written corpora), and, following on from

the comments made in chapter 1, section 2 (and developed below in section 3.1), it makes more sense to use a large, rather than a small corpus. This is particularly the case when investigating multi-word items. Data scarcity problems will arise if we try to compare intuitions about collocations with a spoken corpus.

Using a large mixed corpus (including spoken and written material) as a baseline against which to compare intuitions, as will be done in this research, is not as problematic as it might seem. The intended focus of this study is adjective-noun collocations and Biber et al. (1999, p.506) have indicated that adjectives are rather uncommon in conversation, in comparison to fiction, news, and academic registers. They also note that the academic prose register is the register which has the greatest number of attributive adjectives, and the respondants used in the research reported in this thesis are all university based. There is also evidence that intuitions may actually be closer to written data rather than spoken data – at least in the area of word frequency estimation (as Schmitt & Dunham 1999, p.392 note, with reference to research by Richards 1974).

2.5. National Englishes

If one were to test intuitions against a British corpus (BNC) what of different Englishes? If working with informants overseas, and if deliberately targeting native and non-native speakers, the issue of native speaker origin and non-native target variety might be important, hence the following discussion.

O'Keeffe & Farr (2003) note the effect of regional/national differences in the production of question tags, and in the use of the word *would*, which is more common in Irish English, than in New Zealand, or British, or American English (p.405). Biber et al. (1999, p.545) have also found differences in national varieties in English, for example that American English and British English have different frequent amplifier-adjective collocations (e.g. Americans prefer to use *really good* and Britons prefer *very good*).

However, it would be wrong to conclude that there are great contrasts between British English and American English, with regard to word frequencies and typical collocations. Hofland & Johansson (1982, p.18) in comparing the Brown Corpus (American) with the LOB corpus (British) found very few differences in the frequency rankings of the top 50 words of these small written corpora. They comment "only one of the 50 most frequent words in the LOB corpus has a rank lower than 50 in the Brown Corpus. This is *so,* which has rank 46 in the LOB Corpus, compared with 52 in the Brown Corpus...The correspondence in rank is very close indeed" (1982, p.18)[9]. Kjellmer (1994, p.x) also, does not seem to think that American / British differences are very significant with regards to collocation. Of course, this is an assumption, though as noted in chapter 1, section 3.1, the evidence from Greenbaum (1988) suggests that this belief is more often justified than not in the matter of preverb intensifiers – but it would be unwise to generalize this finding. In chapter 7, section 2.3.2 and chapter 8, section 2.3, where there is discussion about the words chosen for inclusion in the research, Brown and the BNC are compared on the most common collocates of the stimulus words used in the experiments, and the differences are not very great.

To sum up, we can say that the objections to there being no such thing as 'general English' are not insubstantial. However, it has also been argued that there may be a window of opportunity to exploit in this research, by focusing on very high frequency items in the language, that are well distributed. All the same, checks should be made on the stimulus items to ensure that their most frequent collocates are well distributed, and, as far as is possible, are similar among different varieties of English and in written and spoken corpora.

3. Objection 2: Representativeness

***There is no such thing as a representative corpus*[10].**

[9] It should be noted that the compilers of the LOB were trying to put together a British corpus which matched the American corpus (Hofland & Johansson 1982, p.1).

[10] This was one of Chomsky's objections to corpus use, and this challenge refuses to be laid to rest (see e.g. Takaie 2002).

It has been readily acknowledged in the field of corpus linguistics that 'representativity' is not very well defined (e.g. Meijs 1996, p.103; Kilgarriff & Grefenstette 2003, p.343). While it is true in an absolute sense that, "true representativity is illusory" (Williams 2002, p.44) and that, "the concept of a representative sample of the English language makes little sense" (Stubbs 2002c, p.223), this should not dissuade us from striving towards that (unattainable) goal. As McEnery & Wilson (1996, p.64) point out, random sampling in any field (sciences and social sciences) is an established research practice. While there is a possibility that unusual elements may be over- or under-represented in a sample (vis à vis the population), this does not mean that scientists abandon their research. For any sample to be called representative of a larger population, it must be of adequate size and contain non-biased content. These two issues are discussed in some detail below.

3.1. Size

Engwall's (1994) comment, that "no scientific criteria exist for determining the size of any corpus" (p.51), provides a necessary balance to researchers who suggest that the large size of corpora today can cope with Chomsky's (1962) doubts about corpus skew. Chomsky believed that corpus data would inevitably be skewed, "some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural will be so wildly skewed that the description would be no more than a mere list" (1962 p.159). These comments were made many years ago and Leech (1991, p.13) and Svartvik (1996, p.9) both believe that the large corpora available to us today deal with Chomsky's concerns. But the case is not quite so simple. We can say with confidence that the corpus should be large enough for the purpose for which it is being used; however, what this actually means is still unclear. While there are equations to help ensure reliability, even these have their failings (Biber 1993, p.248). As noted in the comments made about the Fox *torrent of NOUN* study (chapter 1, section 3.4), generalizations about phrases, made on the basis of 'inadequate' corpus data (however that may be defined) are highly questionable. It was suggested in chapter 1 that Fox's corpus was probably not big enough for the uses to which she put it.

Regarding usage purposes and the size of a corpus, Francis (1982) points out that the Linguistic survey of Scotland, comprising a list of only 1000 words pronounced by informants, sufficiently highlighted the phonological system under investigation (p.11). Similarly, he notes that a small corpus could help in determining the relative frequencies of the letters in the alphabet, their acceptable combinations and the use of punctuation marks (1982, p.11). Sinclair has suggested that a corpus of 1 million words may suffice to adequately document the grammar of a language (1991a, p.100). The call for large corpora is typically made in the context of rare words[11] and in investigating collocations and multi-word items. Regarding collocations, Halliday & Hasan (1976) believed that a corpus of 20 million words was needed for collocation analysis (p.159[12]). Few would view this figure as high enough today; for example, Beaugrande (1999, p.256) bemoans the inability of his 200 million-word corpus to help identify the 'feeling' words filling the slot *couldn't help* _____ . Indeed, some of those wishing to investigate more infrequent data (e.g. Kilgarriff & Grefenstette 2003) have argued that the use of the Internet as a 'super' corpus is the only solution to these kinds of problems (see below, section 4).

3.2. Content

The second requirement of representativeness is that the corpus should reflect the material in the language population, i.e. not be skewed in content. Is this a realistic expectation? An example from the BNC illustrates this problem well. The BNC indicates that the compound nouns *sewing machine* and *knitting machine* are of comparable frequency[13]. However, this corpus contains extracts from 6 copies of *Knitting Machine Monthly* totaling over 150,000 words. There are no texts from *Sewing World*, or any other sewing machine magazines. Clearly then, we can have no confidence that these

[11] Biber (1993, p.252): "Rare linguistic features show much more distributional variation within texts and thus require longer text samples for reliable representation"; Hunston & Francis (2000, p.16): "As a corpus gets bigger, it is possible to describe more and more accurately items of less and less frequency".
[12] It should be noted that what they meant by the term 'collocation' is rather different from most writers' understanding of the term: See chapter 3, section 2 for more details on this.
[13] There are 88 instances of *sewing machine* and 87 instances of *knitting machine*.

compounds are really of equal frequency in the English language, despite the figures from the BNC[14].

It is quite easy to say what sort of corpus will *not* be representative: a 'one-sided' corpus, i.e. a corpus made up of material from one specific register or genre, for example a newspaper corpus or a corpus of car manuals etc. However, saying what *is* representative is much more difficult. How texts should be sampled and the criteria for selecting them are a large research interest (see Engwall 1994 for an overview), though perhaps not one in which a linguist should engage: Sinclair (1991a) argues that this field is "more appropriate to the sociology of culture" (p.13). In compiling a 'representative' corpus issues such as genre, text-type and register must be discussed and understood[15], and it is clearly not possible that every specialism will be represented in the corpus, a point made by Aston & Burnard (1997, p.40). Some corpus compilers have endeavoured to 'ensure' representativity by having book sale statistics influence both the type of texts to be included and their number (on the BNC, see Leech et al. 2001, p.2; Aston & Burnard 1997, p.29).

Perhaps the greatest failing to date with large mixed corpora is the under-representation of spoken material. It has been admitted quite readily, in various quarters, that this under-representation is a product of pragmatic financial constraints, rather than anything more principled (Aston & Burnard 1997, p.32; Beaugrande 1996, p.530; Takaie 2002, p.120; Schmitt 2000, p.70; Gavioli & Aston 2001, p.238; Svartvik 1996, p.10). As a direct result of this imbalance, Beaugrande (1996) argues that it is wrong to expect congruence between intuition and corpora. He argues that the heavy emphasis of corpora on crime, politics and entertainment material, can in no way be assumed to reflect what people are actually interested in. He comments, "What people can and do talk about in normal life is not necessarily what receives intensive mass media coverage in newspapers and magazines" (1996, p.530). He believes that a solution to the imbalance would be for

[14] An 'advanced exact phrase search' of Altavista (8/9/04) reveals startlingly different figures: 945,000 *sewing machine* hits and only 60,700 *knitting machine* hits. While the ratio of *sewing machine* to *knitting machine* hits is 1:1 in the BNC, it is approximately 15:1 on the Internet according to Altavista.
[15] For a discussion on the minefield of problems in this area see Lee (2002).

corpora to contain a more representative (i.e. larger) chunk of everyday (conversational) language. If intuitions are more affected by conversation than anything else, then the under-representation of conversational material in large mixed corpora may play a (significant) part in the mismatch between intuitions and corpus data[16].

Another way in which corpus material differs from an individual's experience is the broadness of the former. While it does not include everything, it contains material from a wide range of areas: general and specialist. It would perhaps be unusual to find a single individual who reads *Knitting Machine Monthly*, converts old buildings, uses a skateboard and has a rottweiler, and yet the BNC contains material written for people with such interests. In arguing for the usefulness of corpora, Svartvik & Quirk cited in Aston & Burnard (1997, p.5), state that it is unrealistic to expect a certain individual to have 'an adequate grasp' of all language types and specialisms. This is a valuable point, but it should be noted that the limitations which Aston & Burnard highlight are with regards to 'specialist' types of English (i.e. not general English usage), the examples they forward being drafting a law and commentating on a football match. So long as we are testing 'generic' English usage, then this problem does not arise. It does not follow, as Kilgarriff (2001) claims, that because we do not read corpora, that we cannot have accurate intuitions about the language contained in them.

3.3. The BNC

There are a large number of mixed corpora in use today, varying in size (see Singleton 2000, pp.54, 55 for a brief overview). Three of the best-known large corpora in the UK are the Bank of English (BoE, formerly COBUILD), the Oxford English Corpus[17] and the British National Corpus (BNC). The BNC has received a (generally) positive reception by those working with corpora and it has been extensively used by corpus linguists, because of its large size and carefully chosen material. Biber et al. (1999, p.27) comment "The British National Corpus...is exceptional in that it is fairly 'balanced' yet very large-

[16] As noted in chapter 1, section 2.4 above, focusing on adjective-noun collocations may minimize this 'problem' as they are more typically seen in written registers.
[17] http://www.askoxford.com/oec/

100 million words". The BNC is used in this research because of the claims made for its representativeness[18].

Perhaps more than any other corpus the BNC has been marketed as a representative, balanced corpus (Leech et al. 2001, p.1). Very importantly for our purposes, that representativity has been interpreted as giving sanction to the idea that it can provide us with **reliable** frequency statistics, unlike its bigger rival (Bank of English) (Leech et al. 2001, p.1). Because of this, it has been claimed that the BNC is a, "microcosm of current British English in its entirety" (Aston & Burnard 1997, p.29); "a finite, balanced, sampled corpus" (Leech et al. 2001, p.1); and, "large enough and varied enough... to represent an adequate cross-section of written and spoken language" (Leech et al. 2001, p.xi). Because the corpus has been constructed in this way, Leech et al. make the bold claim that frequency data from the BNC can be extrapolated to "inferences about the language as a whole" (2001, p.1). This is, of course, a rather optimistic view to hold in the light of the comment made above in section 3.2, about the *knitting machine* example. However, it might be true with regards to its most commonly represented material.

4. Objection 3: The Internet as a super-corpus

The Internet cannot be used as a super-corpus. Though it is large it is not balanced in terms of content. Internet searches are limited in what they can reveal about language and their searches are prone to errors.

Is the web a corpus? Stubbs (2000) says it is not[19] and Kilgarriff & Grefenstette (2003, p.334) say it is. This difference of opinion is definitional in origin: perhaps the more important point is that Stubbs and Kilgarriff & Grefenstette have used the web in 'corpus-like ways'.

[18] Much has been written about the composition and compilation of the BNC, and details are readily available elsewhere (e.g. Aston & Burnard 1997, Leech et al. 2001), and so not elaborated on here.
[19] Stubbs (2000) prefers to call it a 'text collection'.

As noted in section 3.1 above, some writers have suggested that the Internet, because of its size, can be used as a super-corpus, and the Oxford English Corpus is breaking new ground in using web pages in its corpus. As our interest is collocations or multi-word items, then we might wish to consider whether the Internet would be of use – as a secondary check on the BNC data. The need for 'independent verification' of corpus data has been argued by Stubbs (2002c, p.72), and such a procedure has been adopted in previous research comparing corpus data and intuitions about word frequency. Schmitt & Dunham (1999) ensured that the Bank of English and the BNC concurred on the relative frequencies of words in synonym sets before requiring respondants to rank the words according to their frequencies.

In terms of size, Keller & Lapata (2003, pp.466, 467) have calculated that an Altavista search accesses between 55.0 and 69.1 billion words, and a Google search around twice as much (106.4 and 139.6 billion words) or, put another way, that Altavista accesses approximately 500 times more material than in the BNC, and Google accesses 1000 times more material than in the BNC[20]. If these figures are correct, then it is clear that the web contains considerably more language than that of any individual's lifetime experience[21].

Although size is not the only attraction in using the web[22] it is perhaps the most important. The potential of using the internet as a corpus has been readily acknowledged by corpus linguists[23] and its perceived value in researching language typically scarce in more conventional corpora has resulted in its exploitation in research on: forenames and slang (Blair et al. 2002), certain syntactic constructions (e.g. *who I like*, compared to *whom I like* (Meyer et al. 2003), and bigrams (Keller and Lapata 2003, see section 7.2.2 in this chapter, for more on their study).

---

[20] Similar figures have been forwarded by Renouf (2003, p.40) who claims that an Altavista search is searching over 50 billion words. These figures are probably on the small side today, as the number of pages on the Internet is now considerably larger than when these studies were conducted.

[21] Human language experience has been variously estimated at: 8 ¾ million words a year (Francis 1982, p.9); 10 million words a year (Aston & Burnard (1997, p.28); up to 20 million words a year (Stubbs 1995a, p.49).

[22] See, for example Resnik & Smith (2003) and Cromm (2001) and their use of the web as a parallel corpus, and Renouf (2003) and her comments on neologisms and new uses of words.

[23] See Computational Linguistics vol 29:3 2003, special issue on the web as corpus.

However, using the Internet as a corpus *purely* on the basis of its size, would, according to Biber (1994, p.179), risk the danger of content bias error, simply because unlike 'traditional' corpora, nobody actually knows what is on the net exactly. Biber (1994, p.180) argues that large size alone is insufficient to ensure representativity – diversity of text types is also required[24].

Is the material on the Internet adequately diverse? Blair et al. (2002, p.287) believe that Internet material is representative on the grounds that: it is comprehensive (i.e. covering all subjects, with postings by all sorts of different people); it covers material comparable to spoken corpora (e.g. newsgroup postings); and it is a reflection of current use of English[25]. This is perhaps a rather superficial treatment of the subject as the heavy commercial bias of the web has been noted by Lawrence & Giles (2000). They found that 83% of web content was commercial, the next most significant category being scientific/educational (at 6%). While this does not seem particularly 'balanced', Meyer et al. (2003, p.244) quite rightly note that many .com sites would include some 'non-overtly' commercial type text, e.g. newspaper sites, which do have a significant representation in more traditional corpora. The discussion in section 7.2 below investigates how similar traditional corpus data is to Internet data.

Researchers who use the Internet as a corpus have not been slow to point out some of its weaknesses (see e.g. Kilgarriff & Grefenstette 2003). Firstly, what the Internet can be used for (i.e. its search options) is very limited compared to a traditional (tagged) corpus, though Sketch Engine (Kilgarriff et al. 2004), using the Oxford English Corpus (which contains a large amount of web material), has made this less of a problem because of the software it uses. Focusing solely on search engine capabilities, we cannot readily search for patterns when one (or more) word/s in the pattern is variable (e.g. a *PREP the NOUN of the NOUN* search). In addition we are not able to employ a collocation span search – being limited to searches of strictly adjacent co-occurring words with no intervening

[24] It should be noted that Biber does not make these comments in discussing the Internet, but they are, obviously, of relevance to the subject.
[25] Meyer et al. (2003, p.253) add to this, the fact that the material on the web is unedited and hence "reflective of how people actually use language".

words separating the collocation: a search engine will pick up *old man* (in an exact phrase search), but not *old (other adjective) man*, which BNC will do, if the collocation span window is adjusted accordingly in the search: the internet is only useful in 'exact phrase' searches[26]. Another weakness of Internet searches is that a part of speech (POS) discriminator cannot be used: this is not such a problem in collocation searches, but is more of a problem in looking at individual word frequencies[27]. While information on the distribution of words or collocations in a traditional corpus can be readily obtained (as noted in section 2.1 and section 2.3 above), it is almost impossible to investigate how well distributed a particular word or collocation is on the Internet[28]. Further, the Internet is not helpful in the giving of 'absolute' frequencies. While corpus software can be used to identify the most frequent collocate of a particular word, the Internet cannot. Its strength, as noted by Blair et al. (2002, pp.289, 290), is in the information that it can provide about relative frequencies. One cannot find out what the most frequent collocate of *young* is, by using an Internet search engine, but one can compare the frequency of *young people* and *young man*, by comparing the number of hits when using a search engine in an exact phrase search. When used alongside a traditional corpus, the absolute frequency search problem becomes less of a problem.

There are search anomalies with the Internet, in terms of the material that is picked up in a search. Keller & Lapata (2003) comment, "Google (but not Altavista) will sometimes return pages that do not include the search term at all. This can happen if the search term is contained in a link to the page (but not in the page itself)" (p.468). Another search limitation, noted by Kilgarriff & Grefenstette (2003, p.338), is that search engines only give page hits, not word hits. In addition, it should not be forgotten that search engine searches are punctuation insensitive (Keller & Lapata 2003, p.468) so, if looking for the collocation *various difficulties*, *...various. Difficulties....* will also be picked up. Finally, Kilgarriff & Grefenstette (2003, p.339) have noted that vast tracts of the Internet are not

[26] It is probably fair to say that it is only a matter of time till many more of the features of the other corpora are available for the web. The only thing holding it up is that linguists' searches are not really the same as the searches that most people want to do.

[27] For example, if we want to compare the frequency of the modal verbs *can* and *would*, this is problematic as *can* is a noun as well as a verb.

[28] It should also be remembered that there is a lot of repetition of material on the web, as noted by Stubbs (2000).

searched by search engines, items in the so-called 'hidden web' e.g. databases requiring password access.

5. Objection 4: The dating of a corpus

Corpora are products of their times, and as such, the material they contain will be different from current language usage and so intuitions will differ from the corpus data for that reason.

In their discussion of limitations of corpora, Aston & Burnard (1997, p.40) note the problem of the presence of 'buzzwords' in a corpus: words which are present only because of the time period within which the data was collected. One example of this, (again with regards to the BNC), and limiting ourselves to collocation, is readily provided in the form of a collocate search of *round*. The highest z-score collocate (see chapter 5, section 2 and Appendix 1 for more details of this statistic) for this word is *Uruguay*. The *Uruguay round* of talks in the WTO lasted from 1986-1994, and so it is perhaps not surprising that there are 136 examples of *Uruguay round* in the BNC, as this corpus contains material collected within this period. However, its presence is purely a reflection of when the corpus was compiled, and not indicative of its general frequency or usage today. Similarly, Quirk & Stein (1996, p.29) note that *glasnost* and *perestroika* are present in 1980s corpora (including in an extended (non-literal) sense), but they question whether such words will be found in corpora post 1994. If it really is the case that corpora age so quickly, then, just as Leech et al. (2001, p.x) question the ability of the corpora of yesteryear to represent current usage, so too, we might begin to question the ability of the BNC to represent usage of today in 2006 (collected as it was between 1985 and 1994). Further light is shed on this matter in section 7 below and in chapter 4, section 2.1.2.

6. Objection 5: Corpus material and significance

The material in a corpus should not be given equal significance: some content is more important than other content in shaping our intuitions, but research into intuitions cannot be sensitive to this.

One of Cook's (1998) criticisms of corpora is that they are products, and tell us nothing about how often a message has been received, or by how many people. He comments, "Occurrence, distribution and importance...are not the same" (p.58). This is a very good point and similar observations have been made by Francis (1982) and Biber (1993). Francis (1982, p.10) notes that 1 person or 50 million may have heard something in a corpus, and yet this fact is missed by giving equal weighting to the data in the corpus. Arguing a similar line, Biber (1993, p.248) believes that the relative proportion of books, newspapers and broadcasts in a corpus probably *under*estimates the importance of these types of material in their influence on the language.

The effect of different media in shaping our intuitions of language is largely unknown. Might it be the case that what we hear influences our intuitions more than what we read? Might it be the case that if *thing* is elicited as a frequent collocate for *real*, it is not perhaps due to our exposure to *real thing* in the English language, but rather because of our exposure to the collocation in a variety of different formats (e.g. billboard, TV advert, magazine advert, T-shirt, Coke bottle, key ring etc.)? Is an audible enforcement equal to reading the same words twice, a hundred times? One piece of research which is relevant to this subject is that of Balota et al. (2001)[29]. They required subjects to rank words according to their estimates about how often they were encountered in general (overall) and also specifically in reading, in hearing, in writing and in saying them, on a seven-point scale of frequency. They note that the 'hearing' and 'overall' estimation scores matched each other. This is tentative evidence that what we hear may be particularly important in affecting how our frequency intuitions are formed.

[29] This study is described in more detail in chapter 4.

7. Objection 6: Different corpora different data

Large mixed corpora vary in the information they provide about the relative frequencies of words and collocations.

Do large mixed corpora (and we will now include the Internet in this term) indicate similar information about word frequency and collocation frequency? This is a very important question, and is, in some ways a question that is logically connected to a number of the objections noted above (particularly objection 1).

7.1. Comparisons of 'traditional' corpora

Takaie (2002) believes that large mixed corpora do indeed give different information about words and their frequencies, arguing in particular that the distribution of 'frequent' words is not stable across different corpora. In support of his claim, Takaie provides 2 analyses which are discussed below. If Takaie is right, and his research findings can be generalized, then this appears to be a serious obstacle to the stated aim of comparing intuitions with objectively collected language data: if large mixed corpus (A) differs from large mixed corpus (B) in the information that it provides about words and their collocations, then against what should intuitions be compared?

In his first analysis, Takaie compares modal auxiliary verb frequencies across different corpora. He used Biber et al.'s (1999) data based on the LSWE (40 million words), and two other corpora: Brown/Frown/LOB/FLOB combined (1 million words each i.e. 4 million words), and COBUILD*Direct* (57 million words), norming the findings to occurrences per 1 million words.

Though he does not provide a correlation statistic Takaie reports differences in the relative frequencies of the verbs *will*, *would* and *could* across the above mentioned corpora (2002, p.122). In Table 2.6 below, the rankings of the words in Takaie's research are listed in columns 1, 2 and 3. In addition, in the fourth column research by Mindt

(1996, p.234) is also provided on the same words, based on the first 12 conversations of the CEC corpus (London-Lund corpus)[30]. Also added (in the 5th and 6th columns) is data from the BNC, taken from Leech et al. (2001, p.79). Internet search engine data has not been provided because of part of speech problems (i.e. *may*, *can* and *will* are also nouns and Internet search engines are not sensitive to this fact in their searches). The fourth, fifth and sixth columns, are then, additional data on the subject (not used by Takaie) and are shaded to indicate this.

Table 2.6. A comparison of the frequencies of modal verbs across different corpora. Figures indicate normed instances per million words, (with the exception of column 4) and are listed from most frequent to least frequent.

LSWE	**Brown / Frown / LOB / FLOB**	**Cobuild**	**CEC (relative frequency rank)**	**BNC whole corpus**	**BNC conversation**
Will (3600)	Would (2600)	Will (2700)	Would (27.5%)	Will (3357)	Will (6726)
Would (3000)	Can (2400)	Can (2500)	Can (22.4%)	Would (2904)	Can (5573)
Can (2500)	Will (2300)	Would (2200)	Will (18.3%)	Can (2672)	Would (3737)
Could	Could	Could	Could (10.4%)	Could	Could
May*	May	May	Must (6.5%)	May	Should
Should*	Should	Should	Should (5.4%)	Should	Might
Must	Must	Must *	May (3.4%)	Must	Must
Might	Might	Might *	Might (3.3%)	Might	Shall
			Shall (2.9%)		May

* = Equal

While the first three words are 'common' (i.e. the same words are among the 3 most frequent words) in the Takaie data, they are not of the same relative frequencies, e.g. in

[30] She provides her figures as percentages (relative frequencies) unlike the other columns which provide normed instances per million words.

Brown/Frown/LOB/FLOB *would* is the most frequent modal verb, but this is not the case for LSWE or Cobuild. When the other corpora data are taken into account, it should be noted that there is a considerable amount of agreement between the corpora on the relative frequencies of the words in the table and this might be considered surprising, particularly because some of the additional data comes from two corpora containing *only* conversational material (columns 4 and 6). However, Takaie makes little of the similarities in his analysis: he focuses on the lack of agreement across the corpora he used (columns 1-3 above) in the relative frequencies of the 3 most frequent words. This is, in some ways, a challenging finding, in the sense that high frequency words should have a more 'accurate' sampling than would be the case for infrequent words, i.e. it is more difficult to explain this disparity on the grounds that the samples are too small. The LSWE and the BNC (both of which claim to be representative[31]) have exactly the same rank orderings of the items (though it should be noted that Takaie did not use the BNC in his study). The other large corpus (the COBUILD corpus) does differ from the two other large corpora on the position of the second and third items in the list (*can* is higher than *would*) and this corpus did start out with representativity being a goal[32]. It may be that the relatively higher proportion of spoken material in the COBUILD corpus (10 million of the 57 million words compared with 10% spoken data in the 100 million word BNC), pushed *can* up higher than *would*, since *can* has a higher representation in the conversation corpora (see columns 4 and 6). As such then, the differences can be explained quite reasonably, by noting the different percentages of spoken and written material in the different corpora. Takaie does not consider this as a possible explanation for the findings in his research. Kennedy (2002) notes that different genre types contain different percentages of the modal verbs. In addition to noting that more modals are used in speech compared to written language (2002, p.86, also note column 6 above and the numbers per million words of text), he further notes that the distribution across different texts is not the same because of the different functions of the verbs (2002, p.81). This

[31] Regarding the LSWE corpus (40 million words) Biber et al. (1999) comment "the corpus includes a representative sampling of texts across multiple registers" (p.28); note also the claims of Leech et al. (2001) and Aston & Burnard (1997) in this chapter, section 3.3 about the BNC.

[32] Renouf (1987b) comments, concerning the COBUILD corpus, "When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalizations and against which hypotheses can be tested" (p.2).

being so, if different corpora contain different types of material, or/and different proportions of different types of material then, as a logical consequence, word frequency statistics will throw up differing figures, and sometimes, as is the case here, the differences may be large enough to affect the relative frequencies of words vis à vis another corpus. Takaie's study is helpful in indicating that there may not be agreement between large corpora about the lexical items of interest in this thesis – frequent words. It would be wrong to assume that a perfect correlation between relative word frequencies is a given when comparing different data sources[33].

Does the smaller amount of data in the 'compilation' corpora (column 2) make a difference? Are these corpora less reliable because of their smaller size? The answer to this is a not very satisfying 'maybe'. Certainly, the words' frequencies are so close that it would not be wise to make a judgement on the frequencies in language on the basis of the data. It is a fact that larger corpora will not simply reflect the same findings as smaller corpora: they should 'disperse' the frequencies. Beaugrande (1999) argues that 'norming' is problematic in some ways, as it fails to take into account the possibility of a non-linear directly proportional increase of tokens to corpora size: just because there are 5 occurrences of a word in a 1 million word corpus does not necessarily mean that a 50 million word corpus will contain 250 occurrences. Beaugrande believes that larger corpora will not just reflect the findings of smaller corpora, but be different, be 'more delicate', and this is simply a consequence of dealing with a bigger sample (1999, p.256).

In addition to comparing different corpora in relation to the frequencies of modal verbs, Takaie also investigated inter-register agreement on word frequency across corpora. Takaie compared Biber et al's (1998) study on the distribution of *big, large* and *great* in academic prose and fiction (using the Longman-Lancaster corpus), with Brown/Frown/LOB/FLOB data on the same registers. He found, inter alia, contra Biber et al. (1998), that in the 'compilation' corpora (Brown/Frown/LOB/FLOB), *great* had a

[33] This finding should, incidentally, warn us against expecting that lexical intuitions and corpus data should ever *perfectly* correlate.

higher frequency in academic prose than in fiction. The differences (normed to 1 million words) are presented in Table 2.7.

Table 2.7. A comparison of Longman-Lancaster and Brown/Frown/LOB/FLOB data on the distribution of the word *great* within two specific registers (Academic and Fiction) Numbers = normed instances per million words.

	Longman-Lancaster	Brown/Frown/LOB/FLOB
Academic (*great*)	284	788
Fiction (*great*)	490	514

It is clear that the two corpora are showing different things, both in terms of which register has a higher number of *great* tokens, and in terms of the actual number of tokens per million (this can be compared to the fairly similar number of modal verb occurrences in Table 2.6 when normed to 1 million words – though the BNC conversation data figures are clearly higher than the other data). Two comments need to be made about Takaie's analysis. The first is that in making these comparisons Takaie makes little of the fact that Biber et al. deal with 3 to 4 times as much data as he deals with in his analysis. Biber et al. used 2.7 million words from the Academic content of Longman-Lancaster, and 3 million words of Fiction from the same corpus. Takaie used 0.66 million words from the Academic data in Brown /Frown / LOB/FLOB and 1.01 million words of Fiction data. Takaie has to norm his Academic sample *up* to a million words (rather than norming down) to make his comparison, and many writers have noted the danger of making much of lexical differences on the basis of findings in small corpora [34]. The second point concerns the material in the corpora. Hunston (2002) notes that the concept of 'academic prose' (one of Biber et al.'s registers) is perhaps an "overly blunt instrument" (p.103) since some research indicates that academic material can be quite variable, e.g. history research articles are more similar to narrative than ecology articles are (2002, p.201). This being so, it is not particularly surprising that academic corpora might differ in what

[34] McEnery & Wilson (1996) note that small corpora, "tend only to be representative for certain high frequency linguistic features" (p.64). See also Aston & Burnard (1997, p.15); Sinclair (1991a, p.24) who make similar comments.

they show about different words, depending on their content. This point, taken together with the point made above about the small size of the corpora used in the analyses, should be remembered in considering Takaie's study.

Takaie is keen to find differences across corpora and he makes little of the similarities between the corpora[35]; however, his research is helpful in that it puts us on our guard against blindly accepting the idea that corpus data can be extrapolated to the language/specific sub-language as a whole without some kind of secondary corpus check. If at all possible, at least two large ('representative') corpora should concur on their findings before we proceed to measure intuitions. If two corpora do not agree on details, then it would seem wise to test intuitions more in the area of generalities upon which they do agree.

7.2. Comparisons of 'traditional' corpora and the Internet

The discussion thus far suggests that, subject to certain precautions, the Internet could be used as a corpus. But how 'reliable' would the data collected from the Internet be? Recently, researchers have begun to compare the data returned by Internet search engines about words and collocations with more 'traditional' corpora. Such research enables us to note the degree of correlation between traditional corpora and the Internet on word frequencies and collocation frequencies. How similar are the findings?

7.2.1. Frequencies of words

Blair et al.'s (2002) research was specifically set up to test whether the Internet could be used to provide valid data about word frequencies, and this was done by comparing the data returned from 4 search engines (Altavista, Northern Light, Excite and Yahoo) with word frequency data from Kučera & Francis (1967), based on the Brown corpus, and

[35] See the comments above regarding the relative frequencies of *all* of the modal verbs across the different corpora. In addition, the figures Takaie provides indicate that the relative frequencies of the three words *large, great* and *big* in the academic registers from the different corpora indicate the same relative frequencies: *large>great>big*.

CELEX (Baayen et al. 1995). They used two groups of words to test the correlations between the search engine data and the other two corpora: 250 'standard' words (a mixture of adjectives, nouns and verbs varying in frequency) and 132 'nonstandard' words (slang terms, forenames and African American names). Not surprisingly, the frequency range of the words in the nonstandard group was lower than the standard words. The search engine data (average) correlated with Kučera and Francis data at r = .79 and CELEX at r = .72, both significant at p <. 0001. (The correlation between Kučera & Francis data and CELEX data was .92.) Test-retest findings between the search engine data were (mean average) .92 after a period of six months. The correlations obtained between the different 'corpora' were higher with the standard words than the non-standard words. Blair et al. offer no reason why this may be the case, but a feasible reason is that chance plays a more significant role in whether these items are present in the traditional corpora, as they tend to be less frequent. Blair et al. argue that their research, "demonstrates that Internet search engines provide word frequency estimates that are both valid and reliable" (2002, p.289). It is a key finding that the correlations of the word frequencies between an old corpus (Brown) and the search engines are statistically significant. This challenges the idea that corpora collected at different times (noted above in section 5) will not broadly agree on word frequency information.

7.2.2. Frequencies of multi-word items

In addition to work which has compared Internet data with traditional corpus data on the subject of word frequency, research comparing data on bigrams and bigram frequency has also been conducted across different corpora and the Internet.

Keller & Lapata's (2003) key research interest is data sparseness and how the Internet can help in this area. Of particular interest to us here are the correlations they found between the BNC, NANTC (an American news corpus containing 350 million words) and Altavista and Google search engine data in the matter of bigrams. Keller & Lapata used the BNC to compose a list of 90 word combinations of various frequencies: 30

adjective-noun bigrams, 30 noun-noun compounds and 30 verb-object bigrams[36]. Log-transformed counts were also obtained for these bigrams from NANTC, Altavista and Google search engines. Using these counts they found a correlation between BNC and Altavista for the adjective-noun bigrams of r = .847, the noun-noun bigrams r =. 720 and the verb-object bigrams r = .762 (Pearson's *r* all significant at $p < .01$). BNC and Google correlations were very similar to BNC and Altavista correlations. The correlation between NANTC data and Altavista for the three types of bigram was mean average .722 (significant at p <. 01.) and for NANTC against BNC the correlation (mean average) for the three types of bigrams was .720 (significant at $p < .01$). What is interesting here is that the mono-source corpus (NANTC) correlated less highly with the BNC than did the Altavista data. This suggests, at the very least, that the Internet material is more 'representative' of actual language use than the Newspaper corpus[37]. This is not a particularly startling claim, but adds empirical support to the argument that concern about content-bias, as noted in section 4, is not as well grounded as it might seem, with regard to the Internet. Keller & Lapata's research is important in showing that there are high correlations between web counts and traditional corpora. Their finding is particularly significant as Keller & Lapata are keenly aware of the shortcomings of the Internet being used as a corpus. Despite the problems with search anomalies, etc. (see discussion in section 4 above), Keller & Lapata comment "it seems that the large amount of data available for web counts outweighs the associated problems (noisy, unbalanced, etc.)" (2003, p.470).

The results from the two studies described above suggest that the use of the Internet as a corpus can be justified in terms of its correlation with more established corpora in the areas of word frequency and collocation frequency. This gives us an empirical rationale for using the Internet as a secondary check on BNC data in the research to be reported.

---

[36] So, for example, *guilty verdict* was classified as a high frequency bigram, *guilty secret* medium and *guilty cat* low frequency.

[37] This claim is made on the assumption that the BNC is 'representative' and that when a corpus has a higher correlation with it on frequency data information, this is indicative that that corpus is therefore also more 'representative' than another corpus with a lower correlation.

8. Summary

Aston & Burnard (1997) note that "The BNC was designed to characterize the state of contemporary British English in its various social and generic uses" (p.28). If we wish to compare intuitions about collocation data with the BNC it seems only right that we ask respondants about generic uses, avoiding low frequency items and specialist vocabulary, for the reasons outlined above. While there are important differences in the distribution of different words and collocations throughout different registers, I have argued that there may be a degree of similarity too, and that the research design procedure should endeavour to focus on this area of commonality. The fact that large corpora do not necessarily agree on frequency information - even for frequent items - has been noted. As a consequence of this, a secondary check on BNC data seems sensible. Because a high correlation exists between BNC data and Internet search engine data in our specific area of interest - adjective-noun collocations - it makes sense for BNC data to be confirmed by Altavista data in this research. Unless Altavista and the BNC agree on the data, then items cannot be used in elicitation/intuition tests. This should help ensure that the objective base against which elicited data is compared is a truer reflection of language 'at large', than data from a single source. As a consequence, we can have more confidence that any differences discovered between corpus data and elicited data are not differences resulting from problems with corpus representativity, but due to 'failings' with intuition.

Chapter 3 - Collocation Classification and Psycholinguistic Representation

1. Introduction

There are two sections in this chapter of the study. The first deals with issues of collocation classification (including a subsection on adjective classification), and is a more systematic treatment of the subject of collocation, touched on in chapter 1, section 3.1. The second focuses on the psycholinguistic representation of multi-word items and collocations. Though the second section is our main interest, some background knowledge of definition and classification issues helps inform that discussion. The issues that are discussed here, and the findings from the relevant research, contextualise the later account of how the research was designed, how the hypotheses were made, and how the results were interpreted.

2. Collocation: definition and classification

Definitions of collocation range from the very general to the very specific, and what different researchers mean by the term, is, unfortunately, not the same[1]. It is not an aim of this section to analyse the numerous definitions in depth, but rather to provide an overview of the key criteria used in defining and classifying collocation. Such an analysis sets the groundwork for the psycholinguistic focus section that follows and helps us see more clearly whether there are connections between defining and classifying criteria and psycholinguistic representation.

With the notable exception of Halliday & Hasan's (1976) use of the term collocation[2], it tends to be the case that the same words turn up in the definitions, these words being 'co-

[1] Some have argued that this term is generally understood in the same way, e.g. Zughoul & Abdul-Fattah (2003) believe that, "The definition of a collocation is not a matter of controversy among linguists" (p.61), and Bahns & Eldaw (1993) state that, "There is quite broad agreement on some kind of working definition" (p.102).

[2] Halliday & Hasan (1976) term collocation a "part of lexical cohesion" (p.284) and they used it to refer to cohesion not covered by "reiteration" (1976, p.287). They have been criticized for their definition of the term *collocation* by Hoey, (1991) who believes that the meaning of the term was extended beyond its Firthian sense in their work, and that it became, "an all-purpose lexical cohesive device" (p.154). Herbst

'frequency' and 'restrictedness'. Beyond this commonality[3], however, there is very little agreement among writers. As Kita & Ogata (1997, p.230) point out, researchers have been more interested in defining the term with reference to their own specific research interest, rather than working on a universally acceptable definition. The call for the latter, however, continues to be made (Cowie 1981, p.225; Schmitt 1998a, p.30; Krenn & Evert 2001). Below I provide a brief overview of the above noted 'key concepts' and discuss how they are understood by different writers.

2.1. Co-occurrence of words

Those who have given us definitions of the term collocation typically refer to the 'grouping together' of 'words'. The notion of co-occurrence is, though, not as simple as one might assume. Sinclair (1991a, p.117) argues that we need look no further than a span of 4 words either side of the node word (i.e. the word under investigation) in typical investigations of collocation. Some, however, have noted that this window does not always catch collocating items of interest (e.g. Nattinger & DeCarrico 1992, p.22; Stubbs 1995a, p.47). Kjellmer (1994, p.xiv) on the other hand, in compiling a collocation dictionary based on the Brown corpus, was only interested in strictly adjacent word co-occurrence. Still others have abandoned the pre-defined span position of Sinclair in favour of a linguistic unit focus (e.g. sentence, phrase) including Cantos & Sánchez (2001), who argue that, "what matters is not the span, but the lexical hierarchy which collocates form within a linguistic unit (sentence, phrase, concordance line, etc.). **The lexical hierarchies are neither predetermined nor assumed**" (p.223, emphasis mine).

Regarding 'words' there are also differences of opinion – both in terms of how many and what types. While it has been noted that the traditional focus of collocation has been on two-word combinations (Partington 1998, p.16), word combinations of more than two words are also termed collocations. For example, Firth (1957), sometimes termed the

---

(1996) terms their approach, "a text oriented approach" (p.381), and he also is critical of their definition of the term, as it is so wide. Examples from a text which Halliday & Hasan term "chains of collocational cohesion" (1976, p.287) are *'wallowing...sinking...buried...imbedded'*.

[3] This is more marginal than one might suppose given the fact that these same terms are understood differently.

'father'[4] of the term collocation, called strings such as: *You silly ass!*, *There was an old man of...* and *frittered away time* collocations (p.195) and Kennedy (1991) notes that "Collocations, of course, frequently are more than two words in length" (p.98), providing '*flashed through her/my mind*' as an example. Renouf & Sinclair (1991) coin the term 'collocational frameworks', to describe the phenomenon of "a discontinuous sequence of two words, positioned at one remove from each other" (p.128). Examples that they forward of such frameworks are '*a + ? + of*', '*too + ? + to*' etc. (e.g. *a lot of*, *too late to*). Other writers have also discussed 'slot filling' phrases or expressions and called these collocations, e.g. Nattinger & DeCarrico (1992, pp.36, 41,42)[5] and Smadja (1994, p.148, 149)[6].

A classification issue arises with regard to the types of words in collocations - the 'grammatical'/'lexical' distinction. Benson et al. (1986, p.ix) state that lexical collocations contain combinations of nouns, adjectives, verbs and adverbs, e.g. *warmest regards* (1986, p.xxiv), and that a grammatical collocation is, "a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause" (1986, p.ix), e.g. *account for*. Both classes would exclude the most frequent co-occurring items in the language, for as Nation's (2001, p.334) collocation frequency table (based on the BNC) shows, the most frequent co-occurring items contain *only* grammatical words e.g. *out of, such as*. Most of the research conducted on collocations has been on lexical collocations, and verb-noun collocations in particular.

2.2. Frequency

A requirement fairly consistently forwarded for a string of words to be termed a collocation is that the combination of words be 'common'. Corpus linguists, in particular,

---

[4] For example, Partington (1998, p.15) and Carter (1983, p.174) believe this. Others suggest Palmer's important role in giving the term prominence in linguistics, e.g. Kennedy (2003, p.468); Nation (2001, p.317); Singleton (2000, p.52); Cowie (1998, pp.210, 211).

[5] They term these 'Phrasal constraints', e.g. *a _____ ago.*

[6] He terms these 'phrasal templates', e.g. *the NYSEs' composite index of all its listed common stocks rose _____ to _____* (numbers filling the slots, Smadja 1994, pp.148, 149).

seem to have stressed this frequency aspect of collocation - what Nesselhauf (2005) calls the 'frequency based approach', rather than the 'phraseological approach' to collocations (p.12). Stubbs (1995a, p.23), Biber et al. (1999, p.988), Hunston (2002, p.12) and Hoey (1991, p.7) all stress this aspect of collocation: the combination must be 'habitual' (Stubbs 1995a, p.23), or have some kind of greater than expected statistical basis (Biber et al. 1999, p.988; Hunston 2002, p.12). Kjellmer (1994, p.xiv, xv), on the other hand, abandons the need for a combination to be frequent for it to be termed a collocation, noting that if a frequency requirement is adopted, important instances will be missed. However, it should be noted that this is probably not a principled approach, but rather one which Kjellmer has to adopt in working with a small set of data. In his collocation analysis of the Brown corpus (1 million words), he required only that a combination occur more than once and that it qualify the requirement of "grammatical well-formedness", excluding "inorganic groupings", e.g. *but too, day but* etc. (1994, p.xv)[7]. 'Frequency' as a requirement for collocation is problematic, for, as Schmitt (1998a, p.30) notes, the problem with (an exclusive) frequency criterion is knowing where to draw the line before a combination is frequent enough to be termed a collocation.

2.3. Restrictedness

Restrictedness is, for some writers, a classifying criterion of collocation and for others a definitional one: some writers divide collocations into 'free' and 'restricted' (e.g. Howarth 1998a[8], Aisenstadt 1981), and others prefer to reserve the term collocation for cases when a word is 'restricted' in terms of what it can partner and how. Restrictedness is a subject in which those adopting a 'phraseological approach' to the subject of collocation have a particular interest (e.g. Cowie 1981, 1992). The term 'restricted' has been used in different ways and Herbst (1996, p.385) notes that, "one looks in vain for precise or even coinciding criteria for restricted collocations". Unfortunately, little has been said on this subject of restricted collocation in relation to adjective-noun

[7] Kjellmer (1994, xv) acknowledges that data paucity in a small corpus led him to adopt this approach.
[8] Note though, that he seems to use 'free combination' interchangeably with 'free collocation' at times, e.g. 1998a, pp. 28, 35.

collocations; considerably more work has focused on verb-noun collocations[9]. Below I outline Mel'čuk's understanding of the term 'restricted collocation', partly because he offers several adjective-noun collocations as examples, and partly because of his clear categorization. I then note how two other writers' approaches to the subject are rather different. As noted in chapter 1, section 3.1, it has been argued that native speakers can provide typical/frequent collocates of words if the stimulus word is the 'restricted' word in a 'frozen' collocation. It is important to note that these frozen collocations are only one type of restricted collocation, as discussed below.

Mel'čuk (1998) believes that there are 2 major classes of restricted collocation – each differing from the other in the 'type' of restriction that is present. His two major categories each contain 2 subclasses. In the first category the meaning of one of the collocating words does not have its dictionary meaning (i.e. prototypical meaning). In one subtype, the meaning of one of the words in the collocation is empty, i.e. it plays a supporting role only, e.g. the word *give* in *give a look*, or the word *take* in *take a step*. In these examples, *give* and *take* are delexicalised. Mel'čuk does not forward an adjective-noun example, but Tognini-Bonelli (2001) does offer one that she sees as comparable to these verb-noun collocation examples[10]. In discussing *real*, and its uses, she argues that in a combination such as *real problem*, *real* is delexicalised and has lost its original meaning of 'existing in reality' (2001, p.118, 119). In both cases (i.e. the Mel'čuk examples, and the Tognini-Bonelli example) one of the words in the combinations is delexicalised; however, there are differences between the examples they forward. In the verb-noun cases the verb is *purely* supportive, whereas in the adjective-noun example the adjective emphasises a characteristic of the noun. Returning to Mel'čuk, in the second subtype (first major categorisation) one of the words in the collocation has its particular collocational meaning *only* when combined with the other collocating word or "with a few other similar lexemes" (Mel'čuk 1998, p.31). The examples that Mel'čuk forwards here, include *black coffee*, where the meaning of *black* is 'without milk' in this collocation. Mel'čuk's second major category (again containing 2 subcategories) allows

[9] See, for example, Cowie (1981,1992); Howarth (1998a and 1998b)
[10] She speaks of *take* in *take a photograph* as being a delexicalised use, and she uses the same term (delexicalisation) for the use of *real* in *real problem* (2001, p.117,118).

for both words in the collocation to have their dictionary meaning. In one subtype synonym substitution is not possible, e.g. *weighty* cannot replace the synonym *heavy* in *heavy smoker*; *powerful* cannot replace its synonym *strong* in *strong coffee*. In the second subclass the collocating word has a strong connection with its node word, e.g. *rancid butter, artesian well*. Mel'čuk requires for this latter subcategory that the bond be "utterly specific" (1998, p.31). Presumably, Mel'čuk means that the relationship is exclusive; however, whether there really are 'exclusive' rather than just very strong tendencies in collocation is open to question in the light of corpus data – at least for the examples that he forwards[11]. It might be better to view certain collocates as strong prototypes from which other collocates are derived.

Of these four types of restricted collocation, it would seem that the closest to Fox's 'frozen' collocation (discussed in chapter 1, section 3.1) is the last category forwarded by Mel'čuk. Assuming then, that respondants would be able to provide *well*, to the stimulus word *Artesian* or *butter* to *rancid*, as frequent collocates of these words, what of the other classes forwarded by Mel'čuk? Would or could respondants provide *problem* to the stimulus word *real*[12], *coffee* to the stimulus word *black*, or *heavy* to the stimulus word *smoker*, for example? In chapter 1, section 4.1 it was noted that corpus linguists believe that the fact of delexicalisation works against 'good' intuitions, arguing that intuitions are more likely to be affected by the denotational meanings of words. It would seem reasonable to argue, therefore, that productive lexical intuitions would not be very strong in the other 3 types of restricted collocation noted above. If it is the case that the denotational meaning of a word affects our intuitions then we would expect respondants to produce collocations such as *real gold*, rather than *real problem*, when presented with the stimulus word *real*, where the denotational meaning of *real* (i.e. 'genuine') drives the association. Similarly, we might expect respondants, to produce *black hair*, rather than *black coffee*, as a typical combination containing the word *black*, as the meaning of the

[11] Of the 14 instances of *artesian* in the BNC, on 11 occasions it is followed by *well/s*, and once by *basin* (in *Great Artesian Basin of Australia*), once by *supply* and once by *tube*. *Rancid* has many collocations, e.g: *air, odour, smell, meat, oils, atmosphere, fish* and *fat*, and as such, it is far from 'utterly specific' in its connection with *butter*.

[12] Note that this is not a combination forwarded by Mel'čuk, though it is a case where one of the words is 'delexicalised'.

word *black* in the former combination is more prototypical – it is more obviously connected to colour. It has also been suggested by corpus linguists that a particular meaning of a word or a particular feature of a word may be psychologically salient (see chapter 1, section 4.2). With regard to the class of collocations where synonym substitution is not permitted, it may be that respondants would have problems providing typical collocates where the dominant salient meaning of the stimulus word is not present in the collocation. For example, one would not expect respondants to provide *rain*, or *smoker* when presented with the stimulus word *heavy* but rather a response connected to the prototypical salient feature of *heavy*, connected to 'weight'. Respondants might then produce a combination such as *heavy burden*, in which the word *heavy* has its salient 'weighty' meaning. The research findings of chapter 6 and 7 engage with these hypotheses.

Other writers have developed definitions of restricted collocations which have some overlap with Mel'čuk's classification – but are either simpler or more exacting. In a simpler approach, Benson argues that a collocation is 'restricted' when there are only a limited number of words with which a particular word can combine – collocations are "arbitrary recurrent word combinations" (Benson 1989, p.3). So, for example, *commit murder* is a restricted collocation because *commit*, "is limited in use to a small number of nouns" (Benson et al. 1986, p. xxiv) but *condemn murder* is not, because many things can be *condemned.* The word *commit* has its normal dictionary meaning but it has few (perhaps arbitrary) partners[13].

A definition that is more demanding than Mel'čuk's is provided by Aisenstadt (1981), who argues that in a restricted collocation two conditions must be fulfilled: firstly the words must be used in an unidiomatic way (by which she means "often secondary, abstract, figurative" 1981, p.54), and secondly, either one or both words must be restricted grammatically, semantically and/or by usage patterns. The adjective-noun example that Aisenstadt provides (*cogent argument*) qualifies, in that *argument* as used in

[13] 'Arbitrary' in the sense that not all crimes are "committed" * *libel, theft, break in* (see Herbst 1996, p.388). In addition it should be remembered that other things *committed* are not, typically, crimes – e.g. *adultery, sin, a tort.*

this collocation does not have its prototypical 'angry disagreement' meaning and because *cogent* is restricted, in the sense that usage/semantics restrict it to several nouns only[14]. The first of Aisenstadt's requirements is similar to the first of Mel'čuk's major categories i.e. the meaning of the word (functioning in the collocation) is not its (primary) dictionary meaning. The second of Aisenstadt's requirements is similar to Mel'čuk's third and fourth categories (i.e. the non admission of synonyms, or the requirement of utterly specific (arbitrary) partners). However, it is important to note that Mel'čuk's requirements are either/or; i.e. a combination is termed a collocation in either case. For Aisenstadt on the other hand, there is a necessary combination, i.e. one of the words must be used in an 'unidiomatic way' (like Mel'čuk's first major group) *and* there must be semantic restrictions (like Mel'čuk's second major group). Aisenstadt's definition, is then, tighter and it would seem to exclude more combinations from the 'restricted' class.

The terms 'free collocation' and 'free combination' are used to designate word combinations that are not restricted. Aisenstadt defines free collocations as "combinations of two or more words with free commutability within the grammatical and semantic framework of the language" (1981, p.54; note also Benson et al. 1986, p.ix, who provide a very similar definition). Benson et al. (1986) go on to state that in free lexical combinations, "the two elements do not repeatedly co-occur" (p.xxiv). What this means is rather hard to understand, unless it is meant simply as part of their definition. If, on the other hand, it is meant as a *descriptive* comment it is wrong, for many free collocations (according to their definition and excluded from their dictionary) are high frequency, i.e. they have strong corpus attestation (see chapter 5, section 2.2.4).

Another way of looking at collocations is from a 'cline' perspective. The belief that there is some sort of restrictedness 'cline' with free combinations at one extreme of the cline and very restricted collocations at the other is fairly widely held (see e.g. Howarth 1998a and 1998b, Carter 1987, p.63). Schmitt (2000, p.79) drawing on the work of Cowie & Howarth classifies restricted collocations according to their invariability. The most restricted are invariable, e.g. *from head to foot*, and lesser restricted collocations allow

[14] These are *argument, criticism, reason, evidence* according to the BNC.

limited choice at one point, e.g. *give/allow/permit access to [noun phrase]*. Less restricted still are collocations where there is two point limited choice - *as dark/black as night/coal/ink*. Presumably after this, come free combinations, where multiple choices for either word in the frame can be used to achieve the desired meaning. The cline approach, which, to a certain extent, resists the boxing of collocations, overcomes some of the problems in classifying collocations, though even the above classification can be challenged on the basis of corpus evidence[15].

2.4. Selectional restrictions and embedded collocations

While writers interested in studying collocations have spent much time discussing the issues noted above, there are two additional features that may be of relevance to our current research focus[16]. The first is the matter of the (semantic) selectional restrictions of a word with regard to its collocates, and the second is the 'embedding' of two word collocations in larger chains of language.

Clark (1970) and Nesselhauf (2005) discuss the matter of the semantic selectional restrictions of words when discussing collocations. Nesselhauf defines selectional restrictions as, "conditions for the combinability of elements which are a consequence of the meaning of a word and expressed by means of semantic features" (2005, p.19). Nesselhauf forwards the example of *kill*, noting that the object must be [+ animate][17], and Clark provides the example of *young* and notes how its noun selectional restriction [+ animate], helps explain the responses to the word *young* in word association tasks (1970, p.281)[18]. It would seem reasonable to expect that lexical intuitions would be affected by such selectional features, so that frequent collocates of words with narrow selectional restrictions would be more easily produced than frequent collocates of words with less narrow selectional restrictions, simply because the range of options open to a respondant

[15] The BNC indicates variability on *from head to foot* (especially with *toe* or *heels* substituting for *foot*). As such, this seems to be a one point limited choice collocation, not 'invariable' as classified by Schmitt.
[16] There are other issues of interest too, e.g. collocations with/without pragmatic import (Nattinger & DeCarrico 1992) and non-technical and domain specific classifications (Smadja 1994).
[17] Although when used in an idiomatic way this is not always so, e.g. *killing time*.
[18] Chapter 6, section 3 examines word association responses, and this issue is discussed further there.

is smaller in the former situation. Should selectionally-restricted collocations be classed 'restricted' or 'free'? The answer to this depends on how one views the restriction. Nesselhauf (2005, p.19) notes that if the selection is viewed as originating from within the word itself, then it would seem best to classify it as free, rather than restricted. If, on the other hand, there is an arbitrary element about the selectional restriction, for example that *commit* collocates with some serious crimes, but not all or that *rancid* collocates with some dairy products/fats but not all, then such selectional restrictions might be termed restricted. Bley-Vroman (2002) fails to recognize this arbitrary element in collocation, when he argues that 'meaning' rather than statistics is more important in investigating collocation. He comments "The chief reason why *profound* modifies words like *ignorance* or *admiration* more often than it modifies words like *roof* or *telephone* is because of what *profound* means.... the statistical facts are secondary and derivative" (2002, p.210). This is only partly true. For example, *regard* can and does function as a synonym for *admiration* in some contexts, as can be seen in some overlap in their collocates (e.g. *mutual regard/admiration, particular regard/admiration* etc). *Profound* can, and does collocate with both words: one's admiration or regard for a person can be profound. However, usage statistics are important here, in establishing whether these semantically possible combinations are typical. In an Altavista exact phrase search (5/02/06) there were 44,900 examples of *profound admiration* on the web, but only 1320 example of *profound regard.* One might suppose that this difference can be explained by the relative frequencies of the nouns, but it cannot[19]. Meaning is important: the selectional restrictions of *profound* disallow combinations with *telephone*, but an *exclusive* focus on meaning and selectional restrictions cannot account for the seemingly arbitrary preferences that words have for certain semantically possible partners.

In section 2.1 above, passing reference was made to the number of words in a collocation. Adjective-noun collocations (whether classified as free or restricted) sometimes occur within a larger chain. In compiling his collocation dictionary, Kjellmer (1994) listed not only two word collocations, but also larger collocations that also co-occur in the Brown corpus. Sometimes it is the case that a 'bare' collocation (by which I mean the adj-noun

[19] Noun search in BNC reveals the following: *regard* 1320, *admiration* 927.

combination) e.g. *great importance,* is recorded not only as a bare collocation, but also as part of a larger 'chain' in the Brown corpus, e.g. the prepositional phrase *of great importance*. Interestingly, some of the 'bare' adjective-noun collocations occur **exclusively** in the larger chain according to the data from the Brown corpus, e.g. *similar fashion* **only** occurs in the prepositional phrase *in a similar fashion* in that corpus[20]. At other times the presence in a particular 'chain' is less exclusive as, for instance, with the embedding patterns of *various parts*. In the Brown corpus there are five instances of *various parts of* (i.e. NP containing a prepositional phrase), three *in various parts of* (i.e. prepositional phrase containing a prepositional phrase) and two *the various parts*.

The failure to appreciate that certain adjective-noun collocations (whether restricted or free) occur in particular chains of language has largely been overlooked by writers such as Lewis (1997, p.79) who has advocated the use of 'collocation boxes' in the teaching of collocations, or by dictionary writers who include collocation boxes in their dictionaries (e.g. Cambridge Advanced Learner's Dictionary, Macmillan Essential Dictionary). In these boxes, a word (e.g. a noun) is kept constant and some of its collocates are listed (e.g. the most frequent adjective collocates of that noun). Such an approach highlights the typical combinations, which is a good thing, but also, by its very design, blocks out an appreciation of the different ways in which a particular collocation is typically used. An unintentional consequence of focusing on the adjective-noun combination may be that the learner believes the combinations in the boxes have the same 'chain' preferences. However, this is not the case. For example, Lewis provides 5 collocates for *prospect* (*bleak, daunting, dismal, exciting* and *vague*) to illustrate a collocation box for adjectives and nouns (1997, p.79). These collocations were checked in the BNC and one readily apparent difference between them is that *bleak prospect* rejects *a* as its determiner and the others, in contrast, reject *the*[21]. In some senses it seems that the learner is being 'shortchanged' in the information required to use the collocation in a typical way, when the focus is on the 'bare' two-word collocation. While Kjellmer (1982, p.31) notes that concordance lines may show that the same combination has quite different meanings,

[20] In the BNC there are 81 instances of *similar fashion*. 77 of these occur in the prepositional phrases *in similar fashion* or *in a similar fashion*.
[21] *Vague prospect* could not be checked as there are no instances in the BNC.

e.g.: '*monkeys are capable **of developing** bronchiolitis*', '*within a context **of developing** autonomy and initiative*', concordance lines *sometimes* show that a particular combination can also be invariable syntactically and semantically. For example, *in a similar fashion* only means similarly, and *similar fashion* only occurs within this chain in the Brown corpus. Whether this typical 'embedding' of a 'bare' collocation in a fixed chain of language may play a part in influencing lexical intuitions about the frequent collocates of a word should be investigated.

3. Categorising adjectives

In the overview below, I briefly note the way in which adjectives have been categorised. In particular, descriptions by Quirk et al. (1972) and Biber et al. (1999) are outlined. These brief taxonomies are provided here as they are referred to later on in the research, in chapters 7 and 8.

In their analysis of adjectives, Quirk et al. (1972) note that a particular adjective may be found exclusively in predicative position, e.g. the adjective *loath* in *the woman is loath to admit it*; exclusively in attributive position, e.g. the adjective *utter* in *an utter fool*, or in both positions, e.g. the adjective *hungry* in *the hungry man/the man is hungry*. They note that most adjectives can occupy both positions. They also note that attributive adjectives (our specific interest) may be classified as inherent or non-inherent. In the former the adjective "characterise[s] the referent of the noun directly" (ibid, p.259), and in the latter case the adjective does not describe the noun, but rather something else. They offer the example of *old friend* to illustrate this latter point, in which, if the meaning is 'a longstanding friendship', then *old* refers not to the *friend*, but rather to the duration of the friendship. They note that most adjectives are inherent (ibid, p.266). They classify adjectives into four categories:

- Intensifying:
    - Emphasisers (e.g. *a clear winner*)
    - Amplifiers (e.g. *a complete victory*)

- Down toners (e.g. *a feeble joke*)

- Restrictive: "restrict[ing] the reference of the noun exclusively, particularly or chiefly" (ibid, p.261) e.g. *the precise reason.*
- Related to adverbs: non-inherent, e.g. *an old friend.*
- Denominal: derived from nouns and restricted to attributive position, e.g. *an atomic scientist.*

In terms of semantic classification they use the following taxonomy:

- Stative/dynamic: They note that adjectives are 'characteristically' the former (e.g. *tall*), but sometimes the latter (e.g. *helpful*)
- Gradable/non-gradable: They note that adjectives are typically gradable, i.e. can be modified by adverbs.
- Inherent/non-inherent: See above

Biber et al. (1999) make many of the same points as those noted above. In addition they note that attributive adjectives are more common than predicative across all registers (1999, p. 506), and, as noted earlier (in chapter 2, section 2.4) are dominant in written registers – particularly academic prose. They divide adjectives semantically into descriptors and classifiers (ibid, pp.508-509).

- Descriptors:
    - Colour
    - Size / quantity / extent
    - Time
    - Evaluative / motive
    - Miscellaneous descriptive

- Classifiers
    - Relational / classificational / restrictive
    - Affiliative

- Topical / other

They also note that, "very common adjectives typically designate a range of meanings" (ibid, p.509) and state that descriptors are more commonly used in conversation than classifiers (ibid, p.513).

A rather different approach from the above two classifications is that of Tognini-Bonelli (1993). Her classification is noted here as it is a bridge between this section on classifying adjectives, and the next which looks at psycholinguistic issues in representation. Tognini-Bonelli, drawing on Sinclair (1992), distinguishes instances where she believes the adjective is chosen *independently* of the noun, (which is how she views previous work, such as that of Quirk et al. 1972), and instances where the adjective and noun are co-selected. In this latter class she describes the adjective as *focusing*, rather than *selective*, "An adjective with a focusing function.... is 'co-selected' with the noun, and, as such, is closely linked to the nominal choice rather than being a choice in it [*sic*] own right" (1993, p.194). Sinclair (1992) believes that delexicalisation and co-selection are strongly related. He forwards the examples of *physical assault, scientific assessment, full enquiry* and *general trend* as such cases, arguing that, "In all these cases if the adjective is removed there is no difficulty whatsoever in interpreting the meaning of the noun in the way it was intended. The adjective is not adding any distinct and clear unit of meaning, but is simply underlining part of the meaning of the noun" (1992, p.16). While Sinclair and Tognini-Bonelli use the term 'co-selection', this does not necessarily mean that they are talking of a psycholinguistically real 'unit'; however, the possibility that certain collocations are like words, in terms of their storage in the lexicon, is an interesting possibility. It is to this subject that we now turn.

4. Collocation representation: psycholinguistic perspectives

4.1. Background

The representation of multi-word items in the mental lexicon has become a subject of increased interest in recent years, and the connection between corpus data and the mental

lexicon has been explicitly made by Schmid (2000), who forwards the 'From-corpus-to-cognition principle': "Frequency in text instantiates entrenchment in the cognitive system" (p.39)[22]. In this section I briefly trace the history of the admission of the idiom into the lexicon, the admission of non-idiomatic items and then note how several authors see a place for the lexicalisation of collocations.

The first multi-word item entrant admitted into the lexicon was the idiom. Chomsky allowed for this, and so did others before him. Swinney & Cutler (1979, pp.523, 524) noted how idioms were 'problematic': they defied traditional syntactic and semantic analyses, were often not well formed syntactically, violated selectional and subcategorisational restrictions and were ambiguous. They forwarded the 'lexical representation hypothesis' to cope with these problems, in which idioms were essentially viewed as big words, and, like any other lexical item, were entered directly into the mental lexicon. Whether idioms as a class should be admitted entry into the lexicon has become something of a moot point. D'Arcais (1993) notes that 'dictionary entry theories', like that of Swinney & Cutler, are only one possible way of dealing with the 'problem' of idioms, the other two being either that idioms are, "reconstructed on the basis of one or more of its individual lexemes, which would be characterized by appropriate pointers to the idiomatic meaning" (1993, p.83), or that they are, "computed ...via processes of analogical inference" (ibid, p.83). Cacciari (1993) argues that more and more researchers are beginning to question the 'big word' theory of idioms, noting that this shift has been driven by the recognition that idioms vary in their semantic transparency, and the results of research which suggest that senses of the individual words in idioms are available and recognized (1993, p.49). Two kinds of idioms are (still) allowed entry into the lexicon. The first are those not allowing lexical or syntactic manipulation. Stock et al. (1993, p.230) forward *pig in a poke*, *by and large*, and *hither and thither* as such examples. The second, are idioms which are **totally** non-

[22] Note also the following comment by Stubbs & Barth (2003), "It is plausible that sequences which occur frequently in a corpus, across the language of many independent speakers, have a cognitive status, and that chains are surface evidence of psycholinguistic units which are exploited in producing and interpreting fluent language use" (p.81); see also Schmitt et al. (2004).

compositional: Fellbaum (1993, p.271) provides examples of *trip the light fantastic* and *kick the bucket*[23].

In spite of the challenge to 'idiom as large word' theories, the move to put more and more multi-word items into the lexicon as 'units' has continued unabated, a process sometimes termed 'lexicalisation'. If a series of words is lexicalised, they become, "stored and processed unanalyzed as if [they] were a simple lexical item" (Howarth 1998a, p.25). A seminal paper forwarding this view, and expanding the case for lexicalisation was a study by Pawley & Syder (1983). They argued that many clauses are memorized sequences which they defined as, "strings which the speaker....is capable of consciously assembling...but which on most occasions of use are recalled as wholes or as automatically chained strings" (1983, p.205). They distinguished these from lexicalized sentence stems which they defined as sequences in which the meaning of the expression is not clear from its form, and which behave syntactically as a minimal unit, "a conventional label for a conventional concept" (ibid, p.209). Though they included idioms in this class, they went on to argue that most of these stems are not idioms (ibid, p.211). Examples forwarded as being lexicalised sentence stems are: *What do you think, Think twice before you VP, P_i thinks the world of P_j* (ibid, p.213). They argue that a large part of the lexicon consists of these lexicalised sentence stems. It should be noted that this theory is not, as is the case with idioms, a semantically driven one, but rather, one driven by the desire to explain nativelike selection (i.e. the use of idiomatic language, as opposed to 'grammatically correct' language) and nativelike fluency (i.e. the speed at which language is processed in fluent speech).

With regards to collocations, several writers have made quite broad pronouncements on psycholinguistic representation. For example, Hoey (1991) believes that the ability to guess a missing word (the same word) from several sentences from concordance lines, suggests that collocation is "psycholinguistically real" going on to argue that, "each lexical item is stored...in the context of the sentence in which it was used" (1991, p.154).

[23] Although the traditional view is that idioms are non-compositional, this has been challenged, e.g. by Fellbaum (1993, p.271) and Gibbs (1993, p.62).

This view suggests that restricted, free, lexical and grammatical collocations are potentially stored holistically as a consequence of previous encounters. Backman (1978) seems to argue a similar point, i.e. that linguistic recurrence has a 'psycholinguistic counterpart' (1978, p.2). He believes that the ability to rank multi-word items relatively accurately according to their frequency constitutes evidence of this[24]. Sosa & MacFarlane (2002, p.227) also forward the view that collocations may be stored and accessed holistically. They argue that highly frequent grammatical collocations such as *kind of, lot of,* and *sort of* are stored holistically, as opposed to less frequent *of* constructions e.g. *sense of, piece of.* In such a theory, frequency is the key driving force behind lexicalisation[25] (see also Stemberger & MacWhinney 1986; Schönefeld 1999, p.142 on the connection between frequent repetition and lexicalisation), as opposed to semantics (e.g. idioms), or processing constraints (e.g. lexicalised sentence stems).

Other writers seem to suggest that restricted collocations in particular, are those likely to be lexicalised. Kjellmer believes that, "a large part of our mental lexicon consists of combinations of words that customarily co-occur" (1991, p.112). He suggests that some of these are 'fossilized' combinations (either right or left predictive, e.g. *artesian well* and *arms akimbo* respectively), and others are idioms. However, he also argues that an additional class would be, "sequences of words that co-occur more often than their individual frequencies would lead us to expect" (ibid, p.113). This looks like a statistical definition of collocation; however, it is unlikely that Kjellmer would include 'free combinations' in this class, as he goes on to clarify that, "One word will tend to co-occur with one or a few out of a great number of words that can co-occur with it" (ibid, p.113). This arbitrariness suggests that he seems to have in mind restricted collocations only as potential candidates for lexicalisation[26]. Howarth (1998a, p.42) also seems to suggest that restricted collocations have a privileged lexical representation, but that this is more likely to be the case for severely restricted collocations, rather than for less restricted collocations.

---

[24] This study is discussed further in chapter 4.
[25] This study is a 'Bybeean' one; see section 4.3 for more details on this theory of lexicalisation.
[26] Adjective-noun collocation examples that he forwards are: *classical music, close friend, civilian clothes* (1991, p.114).

4.2. Psycholinguistic explanations for data differences

We now turn our attention more specifically to psycholinguistic explanations for why collocational partners of some words may not be produced in a decontextualised setting in a test type situation[27]. In the discussion that follows I begin by outlining implicit and explicit learning theory and discuss such concepts as knowledge representation, processing and metaknowledge. Though implicit learning theorists have not engaged in the corpus-data elicited-data debate, I argue that their theories are very relevant to our discussion, and writers such as Sinclair and Nattinger & DeCarrico have adopted implicit learning theory concepts in their explanations for the mismatch between corpus data and introspective/elicited data.

Reber (1993), one of the key thinkers in the field of implicit learning[28], defines implicit learning as, "the acquisition of knowledge that takes place largely independently of conscious attempts to learn and largely in the absence of explicit knowledge about what was acquired" (p.5). What is interesting about this definition is the dissociation between what is 'known' and what can be explained, as it seems to mirror our ability to use language, and, at the same time, our inability to accurately describe it[29]. This connection seems especially valid as first language learning is, typically, viewed as implicit learning (see, e.g. N. Ellis 1993, p.290). In implicit learning theory a connection is typically made[30] between the representation of knowledge, the processing of that knowledge and metaknowledge, i.e. what is learned implicitly is processed automatically (i.e. procedurally) and, as a consequence, that knowledge is not available to consciousness (see Sun et al. 2001; Winter & Reber 1994). The alternative way of learning is explicit learning, i.e. learning through hypothesis testing and awareness and instruction. This, unlike implicit learning, manifests itself in declarative knowledge. A key factor forwarded for the difference between implicit and explicitly represented knowledge is the

[27] This discussion covers strictly psycholinguistic arguments, as opposed to the more general arguments noted in chapter 1, sections 4.1 and 4.2, though these two areas may be connected.
[28] It should be noted though that he is not unchallenged in his views, e.g. see Dulaney et al. (1984) and Berry (1996).
[29] This is the view of some corpus linguists.
[30] Though this is not always how writers view this: Rod Ellis (1993, 1994) and Bialystok (1994) forward slightly different views.

degree of analysis involved in the learning process, explicit learning being more consciously 'analysed' learning.

This dissociation sets up on principled grounds why it may be that intuition about language use differs from usage data: put simply, elicited data reflects explicit knowledge and corpus data is a record of implicitly learned language use. However, proceduralisation, as such, says nothing about dealing with different size units in processing – it is simply fast processing. Sinclair (1997) and Nattinger & DeCarrico (1992) have, however, argued that a connection exists between the *size* of the unit processed, the type of processing and metaknowledge.

Discussing the difference between intuitions and language behaviour, Sinclair has commented, "The difference is so marked and regular that it is likely to be systematic. Put starkly, it suggests that the main organizing procedures for composing utterances are **subliminal, and not available to conscious introspection**" (1997, p.29, emphasis mine; see also Rastall 1997, p.90 who makes a similar point). Sinclair suggests that automaticity of processing is a hindrance to introspection, and he is not alone in this belief. Nattinger & DeCarrico (1992), like Sinclair, appeal to the automatisation of processing as an explanation for the mismatch between ideas about typicality and corpus data. They argue that empirical research is needed to determine what the key lexical phrases and structures used in discourse are, because even native speaker interviews will fail to provide researchers with all the data required. They comment, "most lexical phrases are used so automatically that they are quite beyond conscious retrieval" (1992, p.175). However, they do not simply forward a proceduralisation theory. Both Sinclair and Nattinger & DeCarrico forward the view that some strings of language are stored and accessed holistically. Though not forwarding a theory to account for this, they seem to suggest that lexicalised units will be the units that are less open to conscious introspection. Sinclair, in his advocacy of the idiom principle argues that "...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 analyzable into segments" (1991a, p.110). Nattinger & DeCarrico similarly argue for the existence of "prefabricated

lexical chunks" in the lexicon, "readily accessible as completely or partially assembled units" (1992, p.8). How these 'semi-preconstructed phrases' or 'prefabricated lexical chunks' come into being is not entirely clear. While neither Sinclair nor Nattinger & DeCarrico forward theories connecting learning with representation of knowledge, processing and metaknowledge, their views seem to fit in well with implicit learning theory without too much difficulty. The key difference is that the size of unit processed seems to have implications for processing and metaknowledge in Sinclair's and Nattinger & DeCarrico's theories: their theories are not straightforward proceduralisation theories.

However, these explanations for the mismatch between elicited data and corpus data leave too many questions unanswered. Why is it that certain items are lexicalised in the first place? Are all holistically stored units unavailable to our introspection? How do these theories explain the idea that partners of very restricted collocations are available in elicited tests? To try and answer these questions we now turn to an examination of Wray's formulaic language model. The reason for this, is that Wray connects many of the points noted above into a comprehensive model of the lexicon.

4.3. Wray's formulaic language model

Before looking at Wray's model it is useful to explain how 'formulas' are typically viewed by other writers, because there is a fundamental difference about what this term means for Wray and for many other writers.

N. Ellis (2002b) argues that, "Formulas are lexical chunks that result from binding frequent collocations" (p.155). The factors required for binding to occur are frequency of recurrence and repetition (Bybee & Scheibman 1999, p.575; Boyland 2001, p.395). I shall term this the Bybeean approach to formulas, as Bybee has been a key advocate of this theory. Bybee and her followers believe that a lexicalised unit will be different from the sum of its parts in two ways. The first is that the items in the chunk may be reduced in form. For example, Bybee & Scheibman (1999) note that, phonologically, *don't* is more reduced in form in its most frequent usages. Secondly, the resultant unit seems to lose

connection with its constituent parts. For example, Sosa & MacFarlane (2002, noted above in section 4.1), argue that the combination *kind of* may become autonomous from the constituent units *kind* and *of* (2002, p.234). They believe this on the basis of findings in research in which they required respondants to monitor the occasions when they heard *of* from 24 utterances from the Switchboard Corpus of American English telephone conversations. The utterances contained ___ *of* collocations ranging in frequency, and the most frequent was *kind of*. Collocations were placed into different frequency ranges on the basis of their frequency of occurrence in the corpus, and in the test, respondants were asked to press a key whenever they heard the word *of* in the utterances. Sosa & McFarlane's finding was that the reaction time to the most frequent ___ *of* collocations was significantly higher than the reaction times to the less frequent ___*of* collocations. They examined whether it was the case that the response patterns could be explained by *of* being more reduced in form in the most frequent collocations, but this was not found to be the case, "the majority of the reduced forms occurred in the lower frequency collocations to which participants responded significantly faster" (2002, p.235). They argue that their results suggest that very frequent ___*of* collocations are chunked, stored as single processing units, and are autonomous from their constituent parts. It is important to note that this experiment was conducted with *grammatical* collocations. With the exception of the Nordquist (2004) study noted below, in section 4.4, I am not aware of any research conducted on *lexical* collocations which has substantiated the Bybeean view of lexicalisation[31].

Wray defines formulaic language as, "Words and word strings which appear to be processed without recourse to their lowest level of composition" (2002, p.4). Though she believes that lexical units can combine and become fused, for Wray, this is not 'prototypical' formulaic language (ibid, pp.189, 190). Wray views formulas as holistically stored multi-word items, which, rather than having been fused together, have been stored 'word-like' *from the beginning*: "The basic principle [in first language acquisition] is to operate with the largest possible unit" (ibid, p.138; see also Peters 1983,

[31] Bybee & Hopper (2001, p.9) provide a summary of items which have shown "evidence of autonomy characteristic of stored items", including: *I don't know, you and I, come on, over here, over there.* Note that these are *not* lexical collocations.

p.89; Widdowson 1989, p.131). From the formula there may be subsequent analysis, i.e. the language may be broken down (i.e. segmented), but this only happens when needs require it (Wray 2002, pp.122, 130); it is not a default operation. This is a very different way of explaining holistically stored units in the mind.

Wray argues for holistic (non-analysed) representation of a considerable amount of language (ibid, p.202), not just/only language that is frequent or irregular[32]. Indeed, she believes that some infrequent chains will be formulaic and some frequent ones will not be, rejecting a simple 'taxonomical approach' for identifying formulas, arguing that "there is no simple way of categorising formulaic sequences, either by form or function" (ibid, p.70).

An important consequence of Wray's position is that the mental lexicon contains multiple representations of the same word (ibid, pp.202, 203): the same word may be entered in the lexicon as a word, and also as a part of a formulaic string (whether that be fully fixed or a frame). For example, Wray believes that the items *watch*, *your* and *bag* would be stored as separate entries in the lexicon. It may be that these words are combined, if, for example, a child's robot is moving around inside a bag, and someone wishes to draw attention to that fact. In addition, the whole unit *watch your bag* (issued as a warning in a certain situation) might be stored as a unit (or a string '*watch your N*', ibid, pp.252, 253)[33]. While Wray has a role for morphemes in her model, unlike atomistic models of the lexicon, these are not the building blocks in all cases of language production: units bigger than the morpheme play a significant role in language production.

Wray suggests that holistically stored units, "can be analyzed, but only are on a piecemeal basis, and often not at all" (ibid, p.122). It is important to note that analysis is not a default operation, but rather a needs driven operation for Wray, who believes that certain words within expressions may never be analysed. She offers the example of *have*

[32] Wray (2002) does, however, note that irregularity is often a feature of formulas (pp.130, 131).
[33] Another example that Wray (2002) provides, illustrating the same point, is *take it slowly* (p.262). She distinguishes how the chain can be built up from morphemes or accessed holistically if meaning 'perform your action with care'.

in '*thank you for having me*' as an example. Its use is untypical, but this does not trouble the productive lexicon, as the performative function, plus its invariability, will tend to leave the unit unanalyzed. Wray believes that, being based on how we need to analyse our input, our lexicons differ from one another (ibid, p.268). However, there is, perhaps, a considerable amount of similarity too between people, in terms of what material is holistically stored, for as Wray notes, prevalent usage patterns in a community influence this (ibid, p.74). This is because for Wray one of the key functions of formulaic language is to get things done successfully in a society using society's conventionalized ways of executing those tasks (ibid, p.92). The work of Coulmas (1979, 1981), Pawley & Syder (1983), and Nattinger & DeCarrico (1992) is important in this regard, as these studies attempt to delineate what some of those conventionalised norms are.

An important part of Wray's argument is that the native speaker (NS) and non-native speaker (NNS) lexicon may differ quite radically, suggesting that the NNS, who tackles a language later on in life, may have a lexicon of more individual words (proportionately) than native speakers (2002, p.209). In particular, she believes that literacy has a profound effect on the size of units that are acquired, arguing that, "after literacy, the second language learner is increasingly likely to deliberately aim to acquire a lexicon of word-sized units. The relative balance of words to formulaic word strings will be quite different from those of a native speaker" (ibid, p.206). Wray also argues that a particular type of native speaker may have a different kind of mental lexicon compared to other native speakers. Wray argues that "highly literate people" may have analysed their language more, and as a result have more word items in the lexicon, than their less literate counterparts (ibid, p. 268).

Wray believes that collocations may be stored holistically, and she gives an example of how she believes a native speaker encounters and stores a collocation. The adjective-noun example she gives is *major catastrophe* (ibid, pp.206-209). She argues that this collocation would be both, "noticed and remembered as a sequence" (ibid, p.206), for the native speaker. For Wray this means that the individual components are not analysed and that the collocation is stored holistically with its associated meaning. Wray contrasts this

with the second language learner who, bringing an analytic (post literacy) approach to learning, immediately breaks the collocation down: the individual items are not stored together. The consequence of this, is that when a NNS wants to describe a major catastrophe, online assembly must take place, with a variety of success e.g. *major bad event, bad catastrophe*' (ibid, p.206)[34]. For the NS, this is not the case, because the search for a lexical realization of the complete meaning leads to the collocation as a unit. Having said this, it is important to note that Wray allows for collocations to be analyzed and broken down by the native speaker when necessary or desired (ibid, p.211) and, in addition, she sees it as a possibility that (for the NNS) the individual items can be joined together to form a pair (ibid, p.211), i.e. that fusion can occur – a position similar to that of Bybee and Ellis noted above, though Wray is dubious about the importance of frequency in driving this process.

Wray believes that segmentation of a formula will occur, "where the word occurs in a context of actual or potential paradigmatic variation" (2002, p.277). So, then, we might expect segmentation of a hypothetically holistically stored collocation *fast car*, because *fast* can be replaced by *slow*, *clean*, *expensive* etc., and the first language learner will hear these variations. However, according to Wray's theory, we would not expect segmentation in combinations such as *by and large*, as paradigmatic variation is not permitted for any of the words, and no variations are encountered. According to Wray, even if segmentation does occur, the original holistic unit can be retained if it has its own saliency.

In contrast to Cook (1998), who believes that intuition is, "our random and incomplete access to our experience of language" (p.59), Wray believes intuition is more principled, calling it, "a legitimate indicator of lexical organization" (2002, p.281), arguing that her model can explain the differences between intuitive knowledge and corpus data. She believes that 'patterns of knowledge' (i.e. intuition) cannot be equated with 'patterns of

[34] Empirical support for this position is found in Granger (1998), a study referred to in chapter 1, section 3.1. It was noted there, that while native speakers tended to agree on the most typical collocates of the stimulus words, non-native speakers did not, though presumably they had also been exposed to the same combinations.

use' (e.g. evidence available to us from corpora) because the former accesses only a subset of the latter (ibid, p.277). For Wray, intuition is not random it is simply incomplete, but incomplete in a principled way: the constituent parts of an unanalyzed unit are not as available to us as more analysed units: what is segmented is more analysed. Unlike corpus software, which can be set to comprehensively search *all* of the material in a corpus, lexical intuitions about words and their meanings and partners are not comprehensive. Segmented material, i.e. more analysed material is more available. As a result, Wray believes that, "the intuitive meaning of a word is likely to be concrete and discrete" (2002, p.277). If respondants were asked to use *by* in a sentence, the meaning that is given to it is likely to be its *next to* meaning: Wray would not expect respondants to produce a sentence containing the expression *by and large*.

In attempting to test Wray's theory there are two major problems. The first, is that it would be very difficult, if not impossible, to distinguish between proceduralised units (words assembled on line and not stored as 'larger' units i.e. words that are simply processed fast/automatically), fused sequences (which start as smaller units but become chunks over time, i.e. the Bybee theory) and prototypical 'formulaic language' which is holistically stored and has always been so (and may or may not have been analysed). 'Fusion' advocates have argued many of the same points for their version of formulaicity as Wray argues in her theory of formulaic language, for example: resource conserving (Bush 2001, p.268; Wray 1999, p.215; 2002, pp.16, 69); loss of meaning of individual elements of holistically stored units (Bush 2001, p.269; Wray 2002, p.200); and multiple representation (Bush 2001, p.277; Wray 1999, 2002, p.262). Apart from totally different presuppositions, there are perhaps two key differences between the fusion theory and Wray's theory: the role of frequency (dominant in the Bybeean approach) and the importance of paradigmatic variation in segmenting formulas (significant in Wray's model). An eye must be kept on these two factors in the research that follows.

Secondly, there is a testing problem. Wray is critical of experiments purportedly constructed to gain insights into language processing, but which, by their very nature, encourage analyticity. She believes, for example, that some material in the lexicon in her

model would not be activated in tests which encourage analysis, arguing that many (word focused) experimental tasks, "block out access to the larger lexical units in [the] lexicons" (2002, p.271). Wray argues that her model can only be tested if it does not sacrifice "genuine interaction" (ibid, p.280).

However, it is submitted that Wray's theory can be tested (indirectly) using explicit knowledge tests. The relationship between task type and the type of knowledge 'tapped' is well recognized in memory testing (see Toth 2000). In an explicit knowledge task the respondant has to access and analyse memory (a fairly effortful task), rather than speak spontaneously (a fairly effortless process) in a natural context. Rod Ellis (1993) comments, "implicit knowledge becomes manifest only in actual performance" (p.93), and Koriat (2000) believes that, "a critical condition for effective retrieval is the extent to which the processing that occurs during retrieval reinstates the processing that took place during encoding" (p.337). Explicit knowledge tests would then, seem to be inappropriate tests of Wray's theory. But is this the case? One could argue that in an explicit language test, more analysed language will be searched. As a consequence single words in the lexicon would be accessed, or the items in segmented formulas. Formulas which have not been analysed, or which have been analysed but still retain their holistic status may not be so accessible. Therefore, in an explicit knowledge task, it could be hypothesised, following Wray, that *only* non-formulaic language would be accessed. To conclude this section I look at existing evidence for and against the existence of holistically stored language from elicitation experiments.

4.4. Evidence for the existence of fused language or formulaic language

Having noted in chapter 1 that corpus linguists have done very little to compare elicited data and corpus data, we turn now to a study by Nordquist (2004), whose specific aim was to compare the two types of data. Her study is a very relevant one to our research focus. Nordquist, (approaching her study from a 'Bybeean' fusing perspective) required 25 students studying at an American university to provide three sentences (orally) for each of eighteen stimulus words. She reports the results obtained for three of these words

(*I, cause* and *small*), comparing the resulting data with her corpora: an American spoken corpus and the Switchboard corpus. The *cause* example has been noted in chapter 1, and so not elaborated on here[35]. Nordquist records that the most common verbs used with '*I*' in her corpora were *think, know* and *guess* (combined, these three verbs come to 18% and 29% of the data, respectively, in the 2 different corpora she used). However, the elicited data for '*I*' was quite different: these verbs represented only 3% of the responses. With regards to the stimulus *small*, the tendency of her respondants to use the adjective attributively was similar to the corpus data (70% and 77% respectively). She argues that the syntactic position alignment between the two sets of data is part of a general schema for the use of the word. However, the nouns which *small* described in the elicited data, were quite different from the typical noun collocates noted by Biber et al. (1998)[36], which she notes are often quantities (e.g. *amount, piece, sum* etc.). Twelve percent of the elicited responses contained quantity type nouns (e.g. *small amount*), but the remaining data were quite varied (e.g. *small car, small animal, small dog* and *small room*). With regards to this difference she comments, "The specific, high frequency *small*-NOUN dyads...are likely to have separate storage in the lexicon. If highly entrenched, these phrases will have lost connections to lone *small*....decreasing the likelihood of being accessed in the elicitation task" (2004, p.220). She argues a similar line for why the '*I*' responses were out of line with the corpus data, i.e. that the respondants accessed '*I*' and that the connection between the word '*I*' and entrenched phrases containing the word (e.g. *I know*) are weak. Borrowing Sinclair's term, she suggests that the respondants rely on 'open-principle' processing in an elicitation task, as opposed to 'idiom-principle' processing, which is used in typical language use. Nordquist suggests that a possible reason why the collocates provided for *cause* were similar to the corpus data (at least in prosody) is that these collocates are not entrenched, because there are no particularly frequent collocates of *cause*[37]. She concludes her study by stating that, "lexically-specific, highly entrenched units will not be reproduced in elicitation...because of their autonomous mental representations and the higher likelihood of open choice processing

[35] Nordquist found a strong correlation between the corpus data and the elicited data in terms of the dominant negative prosody (See chapter 1, section 3.3).
[36] It is not entirely clear why she refers to Biber et al.'s data and not her own corpora at this point.
[37] One might also add that the prosody of *cause* is far more dominant than the semantic preference for quantity type nouns with *small.*

in the context of elicitation" (2004, p.221). This research can be interpreted as supporting either the fusing position, or Wray's formulaic language position. However, while Nordquist's research is important, it does not exactly match our research interest in elicited data and corpus data for several reasons:

- The respondants in Nordquist's study are not actually being asked to produce typical or frequent examples of usage. One might hypothesize that it would be highly unlikely for a respondant to forward *small dog* as being the most frequent collocation in the English language containing *small*, on the grounds that frequency information is automatically encoded (see chapter 4) and this is such an infrequent collocation.

- Nordquist fails to inform us of how typical or frequent the collocates provided by the respondants were. Not all frequent collocates of *small* are quantity type collocates as she implies (e.g. the 10th most frequent noun collocate of *small* in the BNC is *children*). More information about the elicited responses and their frequencies is required.

- Unfortunately Nordquist only reports on three of her stimulus words and it is not clear how the other results turned out. What of the other data?

- With regards to the responses to '*I*', she makes little of the very frequent responses (with strong corpus attestation) *am*. She comments "*Am*, which occurs frequently in corpora was in fact the most common response in the elicitation task" (2004, p.215). Though she notes that *'m* was not as commonly produced as *am* and that there was much more data for the contracted *'m* in her corpus (2004, p.218), this argument sidelines the fact that a highly frequent dyad (*I am*) was accessed from one of its components in her experiment.

- Regarding her explanation for the responses, Nordquist argues that the dyads are entrenched. She fails to recognize that the unit entrenched may well be larger than the dyad. '*Small amount*' does not, typically, occur as a noun head. It is typically (76% of the

time according to the BNC), in the chain 'DET small amount of NP'. If frequency drives entrenchment then we would expect the chain, not the dyad to be entrenched.

Gilquin (2005a) adopted a similar experimental procedure to that of Nordquist, but her data are rather mixed in terms of the support for the idea that holistically stored material is not accessible in elicitation tests. Gilquin required 40 native speakers of American English to type the first sentence that they could think of when presented with 18 stimulus words on a computer screen: eleven polysemous verbs (including her items of interest *give* and *take*) and seven grammatical words. Although the stimuli were provided twice, only the first replies were analysed in the belief that this would better reflect cognitive salience. She compared her elicited data for *give* and *take* with five hundred examples of *give* and *take* from the Frown corpus and the Switchboard corpus. Gilquin made a semantic classification for her corpus data, and then classified the elicited data according to the same criteria. She notes that the 'hand' meaning of *give* was dominant in the elicited data (42.5%), and that this concurs with the idea forwarded in the literature that this meaning of *give* is the most prototypical. She notes that her corpus data had only 7.7% of this type of meaning. This finding is consistent with the idea that the concrete, discrete meaning of a word will be uppermost in the mind when asked to use a word in isolation (note the views of Stubbs and Sinclair in chapter 1, section 4.1 on this, and Wray's comments in section 4.3 in this chapter).

In another analysis of her results (in which she included all 80 of the elicited sentences from her data), she notes that there was actually a larger percentage of idiom type responses in her elicited data than in the corpus data (16% versus 5% respectively) for the *give* data (2005b). She notes that one of these idiom responses *give me a break* was given in 10% of the responses, and that it is also a very common idiom in the corpus. She notes that this finding - the production of an idiom from a component word - does not sit well alongside theories that forward that idioms are stored holistically, and as such, should not be so easily elicited by a constituent word. The fact though, that the idiom responses tended to *begin* with the stimulus word, might suggest that the provision of the first word may have helped the respondants access formulas beginning with that word.

The responses to *take* also challenge the Wray/Bybee accessibility theories. Prototypical *take*, i.e. the 'grab' meaning, had only a small representation in the corpus data (2.2%). However, for this word (unlike *give*), it was **not** the case that prototypical *take* was dominant in the elicited data. Gilquin (2005a) notes that the most common uses in the elicited data were the 'move' sense of *take* (e.g. *I will take you home*) and phrasal verb instances, (22.5% each respectively of the elicited data). This is a particularly difficult case for Wray's theory as Wray specifically argues that, "an intuitive definition of *take* will home in on its concrete meaning of 'grasp' or 'capture' because it is in this meaning that it is most segmentable. Its common occurrences as an abstract carrier verb, in for example *take part*, *take on* ....are much less visible to our intuition, because there will have been little if any drive to segment *take* out of these strings" (2002, p.277). This finding of Gilquin would also be problematic for the Bybeean position if it were the case that the phrasal verbs produced (e.g. *take place*) were particularly common in the language; however, she does not state which phrasal verbs were produced. These findings are very interesting, and the *take* example, in particular, does not fit in well with the comments made about discrete word meaning affecting intuitions.

These latter data, plus the ability to produce highly frequent collocates of words in frozen collocations (as reported in chapter 1, section 3.1) do challenge the Wray/Bybee theories. While then, there is some empirical evidence for the existence of either 'fused' language or formulaic language affecting our intuitions in the matter of collocations, the data are rather ambiguous in the support that they give for these theories. Gilquin (2005b, p.157) argues that her data suggest that only **some** clusters might be stored in the mind – though she does not provide a rationale for which ones. Clearly, more research should attempt to discover which ones and why. A tentative suggestion is that some holistically stored material may retain stronger links with the component parts than other holistically stored material. For example, idioms may be more or less analysed. Stock et al. (1993) distinguish non-analyzable and analyzable idioms and Gibbs argues that there is a strong relationship between how analyzable an idiom is, and how salient its individual parts are (1993, pp.62, 63). Perhaps some formulaic language may be more accessible and some less accessible in elicitation tasks. It is important to note that, at least in the context of

aphasia, Wray (2002, p.246) argues that *some* formulaic sequences are more easily accessed than others, noting in particular the absence of semi-fixed strings in nonfluent aphasia. In chapter 6, I reanalyze word association data, in an attempt to dig a little deeper into this area, and explore the various options open to us about why it may be that some material, which one would expect to be holistically stored, may be accessible in elicitation tests.

5. Summary

An important classification in the literature is to divide collocations into restricted and free. An attempt has been made in this chapter to make connections between classification and the psycholinguistic representation of collocations. Whether there is a psycholinguistic reality to restricted collocations (alone) is an interesting possibility, and this is explored further in chapter 5. In addition to the view that partners of words in frozen collocations may be produced in explicit test conditions, it has been noted that words with very specific semantic selectional restrictions might more easily elicit frequent partners. Further, if it is the case that the denotational meaning of a word affects lexical intuitions about typical collocates, then it may be that in three situations collocates within restricted collocations will not be easily elicited from one of the component words: when the stimulus word is supportive and delexicalised; when the stimulus word has a different meaning than usual in the combination; and when the stimulus word in the combination cannot be substituted with a synonym having the same key semantic feature. There is some evidence for the idea that holistically stored material may not be accessed by a stimulus component word in an elicitation test, but this is not very conclusive, and some experimental data does not sit comfortably alongside the views of Wray and Bybee.

Chapter 4 – Word Frequency Estimation and Frequency Estimation Theory

1. Introduction

'Frequency' is a key concept in this research, and the main focus of this thesis is the ability to recognize and produce frequent collocates of words. Up to this point, though mentioned many times, frequency encoding and access has not been discussed in any great detail. In this chapter I look at research examining the ability to provide accurate subjective frequency estimates (SFEs), particularly of words, and the theoretical framework that has been built up to explain frequency encoding and memory assessment strategies. There are two sections in this chapter. The first deals with word frequency estimation research and the second is concerned with theoretical explanations for the ability or inability to provide reliable frequency estimates.

2. Frequency estimation research

In his oft-cited paper of 1976, Richards outlines his eight vocabulary knowledge assumptions. Of particular interest to us here is the second assumption: "Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also 'know' the sorts of words most likely to be found associated with the word" (1976, p.79). This assumption proposes accurate knowledge about word frequency and collocation frequency. As such it stands in direct opposition to Hunston's view (chapter 1, section 3.2) that, "It is almost impossible to be conscious of the relative frequency of words, phrases and structures except in very general terms" (2002, p.21) and Stubbs' (1995a, pp.24, 25) comment that, "[native speakers] certainly cannot document collocations with any degree of thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations". It is important to note that Richards forwarded his view at a time when corpora were small; however, he did use corpora in his own research on word frequency estimation and he refers to other corpus-based work investigating this subject to substantiate the assumption noted above. As noted in chapter 1, a considerable amount of research has investigated subjective word

frequency estimates (though very little has been written about collocation frequency estimates) and in the first section of this chapter the relevant work is reviewed. The key issues covered include research into the ability to rank words according to their relative frequencies, the differences or similarities in correlations between SFEs and small or large corpora and frequency estimation abilities in other fields (linguistic and non-linguistic).

2.1. Word frequency

In the review that follows, rather than describe each piece of research separately, the key findings of the main studies on word frequency estimation carried out over the last four decades are presented in Table 4.1. The studies are listed chronologically, and provide the basic information about the experiments, the words ranked, the methodology employed and the correlations found. The discussion that follows the table highlights some of the more interesting issues that arise from these studies.

Table 4.1. Summary of research investigating word frequency estimation

	Tryk (1968)	**Shapiro (1969)**	**Carroll (1971)**
Subjects	50 University students (24 men / 26 women)	6th graders, 9th graders, College sophomores, Industrial chemists, Elementary teachers, Newspaper reporters (All x20)	Group a: 15 lexicographers Group b: 13 non-lexicographers (all college educated)
Words	100 nouns (evenly log sampled from very rare to very common)	Mixture (majority nouns) from Thorndike-Lorge (1944) / Kučera & Francis (1967)	As Shapiro (1969)
Methodology	Rank words in format: 'no.: time period' a) Public use (conversation) b) Private use (conversation)	a) Multiple Rank Order (MRO)* b) Subjective Magnitude Estimation (SME)* Half respondants: spoken language estimates, other half written language estimates	Subjective Magnitude estimation (Give no. to a word and then rank relative to other words)
Obj. standard/ Correlation	Thorndike-Lorge (1944) Public .775 (Av. Test / retest) Private .75 (Av. Test / retest)[1]	Thorndike-Lorge (1944) / Kučera & Francis (1967) MRO: .952-.975 SME: .920-.958[2]	Thorndike-Lorge (1944) / Kučera and Francis (1967) Group a) .970 Group b) .923[3]

* See discussion below for more on these terms

[1] Test / Retest correlation =.97; The Public-Private difference was not statistically significant.
[2] There was no significant difference between the groups
[3] There was a highly significant difference between Group a) and Group b); i.e. the lexicographers performed better at this task than the other subjects.

Table 4.1. (continued) Summary of research investigating word frequency estimation

	Richards (1974)	**Backman (1976)**	**Frey (1981)**
Subjects	1000 Canadian college students (50 subjects for each group of 250 cards)	40 students (2 groups: 1 group word frequency estimation, 1 group word familiarity estimation).	Group of 46 native US students, faculty, office employees, professionals, high school students
Words	4495 concrete nouns (tested in groups of 250 cards)	50 / 60 items from Shapiro	12 nouns
Methodology	Sort the nouns into 5 groups from 'very often seen, heard or used' to 'never…'	Magnitude estimation a) Frequency b) Familiarity	Rank 12 nouns according to general frequency (and specific text type frequency)
Obj. standard/ Correlation	.5750 Kučera & Francis (1967) .6051 Rinsland (1945) .2777 West (1953) .3753 Schonell et al. (1956) .3953 Howes (1966)	Thorndike-Lorge (1944) / Kučera & Francis (1967) Frequency: .917 - .943 Familiarity: .825 - .903	.94 Kučera & Francis (1967) .913 Thorndike-Lorge (1952) .952 Howes (1966)

Table 4.1. (continued) Summary of research investigating word frequency estimation

	Ringeling (1984)	**Arnaud (1990)**	**Schmitt & Dunham (1999)**
Subjects	5 Dutch (advanced speakers of English) 5 Native English speakers All Staff members of Dutch university	French respondants: 126 1st year students (studying English) English respondants: 87 American sophomores at French university.	5 groups native speaker 8 groups non-native speakers (Intermediate-Advanced)
Words	18 nouns from Word Frequency book (Carroll et al.1971)	30 words: nouns and adjectives	12 set of synonyms
Methodology	a) Public Estimate (i.e. language in general) b) Private Estimate (i.e. own linguistic use) Ranking method	Rank ordering	Rank items in synonym set (relative to one item already marked 1)
Obj. standard / Correlation	Word Frequency book (Carroll et al. 1971) Dutch a) .85 b) .68 English a) .82 b) .79	(French words) Juilland et al. (1970) / (English words) Carroll et al. (1971) Francophones: With French list, mean average .61 / With English list, mean average .70 Anglophones: With French list, mean average .63 / With English list, mean average .76	BNC / COBUILD Native speakers .530 Non-native speakers .577

Table 4.1. (continued) Summary of research investigating word frequency estimation

	Desrochers & Bergeron (2000)	**Balota et al. (2001)**	**Blair et al. (2002)**
Subjects	-	574 undergraduates 1590 'mixed' adults, all Native Speakers	33 Undergraduates American University
Words	1916 French nouns	2938 monosyllabic words Groups of words rated by 30 – 69 respondants.	a) Standard (250 mixed part of speech) b) Non-standard (132)
Methodology	Rank according to 7 scales of frequency	Rank according to 7 scales of frequency: 'never' to 'several times a day'	Rank 1-5 according to familiarity
Obj. standard Correlation /	Baudot (1992) Equal to or greater than .78	Kučera & Francis (1967) CELEX (Dutch center for lexical information 1995) Between .78 and .83	Kučera & Francis (1967) CELEX 4 search engines a) .40 (standard average against all objective lists) b) .61 (non-standard average against all objective lists)

It can be seen from Table 4.1 that word frequency estimates (generally) correlate well with established word counts/corpus counts. Of the 12 experiments listed, only 3 indicate that word frequency intuitions correlate only mildly with objective counts: Richards (1974), Schmitt & Dunham (1999), and Blair et al. (2002). If we take these results at face value, word frequency estimate ability appears weaker in the ability to accurately rank: concrete nouns (Richards 1974[4]), synonyms (Schmitt & Dunham 1999) and words which are, "relatively familiar, but have very different objective frequencies in the language" (Blair et al. 2002, p.290). In the discussion that follows some of the key issues arising from these research findings are examined, and possible reasons for the variety in the results are discussed.

2.1.1. Methodological issues

It is important to note that the research summarised in Table 4.1 elicits the frequency estimates in quite different ways. Arnaud (1990) distinguishes *absolute* methods of ranking words and *relative* methods. In the former, respondants indicate how often they use a word and this is termed the magnitiude estimation method. This is either elicited by asking the respondants to indicate how often they use a word, e.g. on a scale from 'very often' to 'never' (e.g. as used by Richards (1974), Desrochers & Bergeron (2000), Balota et al. (2001) and Blair et al. (2002)), or by requiring the respondants to provide a frequency estimation themselves, using the template 'Number::Time period' e.g. *5 times a day*, a method employed by Tryk (1968). The 'relative' method of testing word frequency estimation can also be divided into two categories: ranking words in sequential order from a random group (a method used by Shapiro (1969), Frey (1981), Ringeling (1984) and Arnaud (1990)); or ranking the items from an anchor word (e.g. as used by Schmitt & Dunham 1999). In the former case the randomly presented words are simply ranked by the subjects from the most frequent to the least frequent. In the latter methodology there is also a variation. One of the words might be given a specific number by the researcher(s) and the respondants are asked to rank the other words relative to the

[4] Richards (1976, p.79) specifically excludes concrete nouns from his second vocabulary assumption (mentioned in section 2), on the basis of his 1974 study.

frequency of the anchor word (a methodology adopted by Schmitt & Dunham, 1999 and Backman, 1976). For example, Schmitt & Dunham put the number 1 next to one of the words in a set of synonyms. They required the respondants to rank the other words by indicating how much more frequent or less frequent the other words were, relative to that word. For example, if respondants believed that another word was ten times more frequent it would be labelled 10, and if they thought it was 3 times less frequent it would be labelled 0.33. An alternative relative ranking technique from an anchor word is when the respondants *themselves* are required to make one of the words an anchor word, and give it an appropriate number (a methodology adopted by Carroll 1971). Of the four main techniques described above, the relative ranking method when there is an anchor may confuse numeracy skills and word frequency estimation abilities. For example, Aizawa et al. (2001) point out that in Schmitt & Dunham's experiment (which utilised this technique, with the anchor word provided by the researchers and given the number 1), the correct ranking (according to BNC and COBUILD frequency figures) for the following set of synonyms (in which *catastrophe* is provided as the anchor word, and given the number 1) would be as follows:

Disaster	6.85
Tragedy	4.17
Catastrophe	1
Calamity	0.29
Cataclysm	0.06

Clearly, to mark these words according to their relative frequencies with the above numbers involves some quite difficult arithmetic skills and Aizawa et al. are extremely critical of Schmitt & Dunham's research, because of this design weakness. Schmitt & Dunham failed to follow the advice of Carroll (1971, p.723), who suggested in his instructions to his respondants that the anchor word be given a *large* number to avoid the problems of dealing with fractions.

Typically the words used in the older experiments were chosen from a published list of word frequencies compiled from a small corpus (e.g. Kučera & Francis (1967) based on the 1 million word Brown corpus, and Carroll et al. (1971) based on the 5 million word American Heritage Intermediate Corpus)). The words in the tests were chosen according to their frequencies with equal steps in range of frequencies between the items to be ranked. Little attention has been paid to the sensitivity of word frequency estimation, i.e. at what point frequency estimates fail to be reliable. Carroll (1971, p.726) believes that differences in frequency of one order of magnitude, i.e. x10 raw frequency are clearly distinguishable. On the basis of his study, Frey (1981) extends this to x3 raw frequency estimation recognition for, "people with good stylistic sensitivity" (p.405).

2.1.2. Different corpora different results?

As Schmitt & Dunham (1999) note, much of the research conducted into subjective word frequency estimation hails from the mid 1960s to the 1980s. Researchers working in that period utilised reference works based on only small corpora, as noted above. Those conducting the studies were well aware of the problems of trying to compare subjective frequency estimates with such corpora. With regards to size, Carroll (1971, p.728) notes that the corpora are small compared to an individual's linguistic experience, and Ringeling (1984, p.61) makes a similar point. On the issue of sampling error, Carroll (1971, p.728) notes that word counts may misrepresent low frequency words in particular, and Frey also makes the same point (1981, p.401); indeed Ringeling (1984) extends this 'skew' argument, commenting that, "it is likely that objective counts misrepresent frequency in some cases rather unpredictably, and not, as is generally recognized, with respect to extremely infrequent words only" (p.68). In addition, the idea that a corpus ages and fails to keep up to date with either the common words in the language or the changes in the usages of words is commonly noted (e.g. Tryk, 1968, p.171; Blair et al. 2002, p.286; Balota et al. 2001, p.639) and forwarded as a reason why the corpus evidence and the subjective estimates may differ.

Researchers utilizing computational methods to investigate word frequency would probably concur with such views: the criticisms quite rightly recognize the possibility of sample size and sample bias error in corpus creation and use, as discussed in chapter 2, section 3. What is contentious, however, is the interpretation of divergence between the corpus data and the frequency estimates.

It seems somewhat ironic that Tryk, Ringeling and Carroll make the above comments, for their studies are amongst those which indicate a **high** correlation between subjective frequency estimates and the corpora/word lists used. Their comments are made, in part at least, to explain why there is not a *perfect* correlation between the subjective frequency estimates and the corpus data. These word frequency researchers approach the issue of frequency estimation from a completely different set of presuppositions than some of the corpus linguists mentioned in chapter 1. Rather than trusting the objective data they trust intuitions - implicitly. So, when the two differ (even in a small way), some of the above noted researchers have held the intuitions to be more reliable, and, for the reasons noted above, have been critical of the objective counts. For example, Ringeling (1984) states that his research, "demonstrates that objective counts are not necessarily better reflections of linguistic reality than subjective estimations. In fact, the reverse may be true" (p.59). Carroll (1971) claims that, "It can be argued that the subjective estimates are more valid, on the grounds that objective frequency counts ....are subject to biases of various kinds in sampling.... and that human observers are better able to discount such biases" (p.728), and Frey (1981) believes that, "our best assumption is to view the subjective estimates as a more accurate reflection of the "actual" frequency distribution in the current language usage than any of the frequency counts available so far" (p.401). There are various methodological explanations forwarded to explain why there is not a perfect correlation between the respondants SFEs and the word lists/corpora. For example, Tryk (1968, p.175) suggests that variations in rank order can be explained by respondants being asked to provide *oral* estimates of frequency, whereas the corpora / word lists are based on *written* data, and Richards (1974, p.78) argues a similar point about why the concrete nouns in his study were not well ranked. In addition, it has been suggested that

instructions have not always made it clear whether the words are to be ranked according to *personal* usage or *general* use (Ringeling 1984, p.62).

The key theoretical basis for the belief in the ability to accurately rank the relative frequencies of words (and developed in greater detail in section 3 below) is provided by Tryk, who sums up clearly the suppositions underlying the frequency estimate research of his era:

> This study was generated by the assumption that individuals are able to give valid and reliable reports reflecting the degrees to which they have processed given words. That is, it was assumed that people carry with them a kind of subjective 'yardstick' of word frequency enabling them to measure the magnitude of words on a dimension of word frequency, much as they give quantitative estimations of perceived intensity, length, duration, and numerosity in psychophysics (1968, p.170).

The alternative position to take, when faced with differences between corpora and frequency estimates, is to trust the corpus data and see the weaknesses in the subjective estimates. Such an attitude is adopted by Schmitt & Dunham (1999) and Blair et al. (2002) in their research. As noted previously, the results from both of these experiments indicate a weaker correlation between frequency estimates and the corpus/corpora data, than the other studies reported[5]. The increased size in the corpora that these researchers used, plus the strong correlations between different 'objective' corpora (BNC and COBUILD for Schmitt & Dunham, and search engines and older word lists for Blair et al., see chapter 2, section 7.2.1) give added support to these writers' trust in the reliability of corpus data rather than intuitions. It should be noted, however, that an increase in corpus size does not *necessarily* lead to a decrease in the degree of correlation with subjective frequency estimates. Balota et al. (2001) used CELEX (Dutch Center for Lexical Information 1995) in their study, an 18 million word corpus, and found a high correlation between the SFEs and this corpus.

[5] Schmitt & Dunham's study may, however, be quite flawed in its methodology, see section 2.1.1 in this chapter.

Schmitt & Dunham (1999, p.393) have questioned the validity of the results obtained from the early studies into SFE ability indicating high correlations, partly on the grounds that the objective base against which the subjective estimates were measured was too small in these studies[6]. Is it the case though, that the results from these earlier studies can be dismissed so easily?

The clearest evidence that can be forwarded to either substantiate or challenge Schmitt & Dunham's view would be to substitute the small corpus data, against which the older subjective estimates were measured, for large modern corpus or internet search engine data on the relative frequencies of the words. Using the newly established ranks the correlations between the subjective estimates can be recalculated. Fortunately, this can be done with the Ringeling (1984) data, for his research provides not only the words used in the experiment, but also the ranks of the words according to the objective base and the rankings provided by the 10 respondants. Ringeling required 5 native speakers of English and 5 advanced non-native speakers (Dutch university staff) to rank 24 nouns, from a range of frequencies from Carroll et al. (1971)[7], using a relative ranking technique i.e. ranking the randomly provided words.

The following procedure was followed in recalculating the correlations from Ringeling (1984) against a new objective base. Word searches were conducted in Altavista (23/08/05) for the 24 words used in Ringeling's study. These were ranked and compared with the Word Frequency Book (WFB) ranks (i.e. the ranks used as the objective standard in the Ringeling 1981 study). The resulting correlation between the two (objective) ranks[8] is .931. This is a highly significant correlation, and this might be considered surprising, given the age of the Word Frequency Book, published in 1971. This in itself is an interesting finding[9]. The 10 respondants' ranks were then re-ranked

[6] Other methodological weaknesses that they believe to be in these studies are that the words vary widely in their frequencies (e.g. *the* at one end of the spectrum and *echidna* at the other).

[7] These words, from highest to lowest frequency were: *time, man, end, top, class, game, language, teacher, science, desk, student, blanket, literature, turtle, judgement, grammar, boot, starvation, ingredient, contradiction, imagery, fissure, benefactor, gusset.*

[8] This was the case when the first letter 'e' in *judgement*, was retained in the Altavista search.

[9] This should be seen as additional support against the argument in chapter 2, section 5 that corpus results from larger and newer corpora will be different in what they reveal about the relative frequencies of words.

against the Altavista data, to investigate whether the ranking against a large corpus would reveal significantly different correlations. The results are given in Table 4.2 below.

Table 4.2. A reanalysis of the Ringeling word frequency estimation data (condition general) comparing correlations of SFEs with WFB and Altavista search engine results data (D-Dutch subjects: E-English subjects)

	D1	D2	D3	D4	D5	E1	E2	E3	E4	E5
Ringeling study (WFB)	.88	.90	.83	.87	.75	.85	.89	.84	.78	.74
Ringeling replication (Altavista)	.896	.896	.755	.844	.691	.875	.869	.783	.767	.686
Ringeling WFB mean average	.846					.82				
Altavista mean average	.816					.796				

It can be seen that the subjective frequency estimation correlations are a little lower against the search engine rankings (.816, .796) than they were against the WFB (1971) data (.846, .82), but the results still indicate a significant correlation between the SFEs and the corpus. On the basis of this finding then, it seems that the criticisms leveled by Tryk, Carroll and Ringeling against their own small corpora are not as valid as they assumed. Neither, it should be added, is the challenge of Schmitt & Dunham so valid in the light of the data presented above. The fact that the objective data used in the early subjective frequency estimate studies were from small corpora does not seem to have significantly affected the validity of the findings from those earlier studies. As a result, it

can be said with greater confidence that word frequency estimates are, on the whole, quite accurate.

Excursus – Second language learners and word frequency estimates

With regard to research into subjective frequency estimates conducted with non-native speakers of English, it has been suggested that tests may actually reveal *nothing* about a subject's appreciation of the relative frequencies of words in the second language when there are a lot of common cognates between the mother tongue and the second language being learned (Arnaud 1990, Aizawa et al.1991, p.81). Subjects may be able to 'accurately' rank words purely on the basis of the cognate frequency in the mother tongue. Partly for this reason, Arnaud (1990) has questioned the view that word frequency estimation can be used as a tool in testing second language proficiency, though some researchers believe that non-native speakers can develop quite accurate sensitivities of word frequencies in a second language. Schmitt & Dunham believe that their research shows that educated non-native speakers may be able to better judge the relative frequencies of words than uneducated non-native speakers (1999, p.402) and Ringeling's study (discussed above) in particular, also indicates such an ability.

2.1.3. Priming effects

Can differences between SFEs and corpora data be explained because of priming effects? A study by Toth & Daniels (2002) is relevant to this question. In their study, they investigated whether priming occurred in judgements of normative word frequency, using two sets of respondants tested in different conditions. In the 'full attention' task, subjects were asked to read out words from a computer screen (nouns and adjectives from Kučera and Francis' data) and remember them for a later task. After this a normed frequency judgement task was given, requiring subjects to indicate whether words (presented on a computer screen) were high or low frequency in language. The subjects were informed that some of the words in the frequency task had been in the first task, but that this should not affect the task in hand. The second set of respondants participated in a divided

attention task (a reading and aural digit attention task, where the focus was on the listening), before being presented with the words on the computer screen. The findings were that for the first set of respondants prior presentation of an item resulted in that item being (norm ranked) *higher* than was the case for the second set of respondants. In their discussion, Toth & Daniels note that there is a wealth of evidence supporting good frequency judgements, including word frequency judgements, but their findings (in the full attention task) were that, "a single prior experience [of a word] can bias the apparent normative frequency of words" (2002, p.846).

So then, it seems that there is some warrant here for the idea that recency of exposure to words in SFE tasks may influence frequency judgements. However, it must be noted that this 'contamination' was the finding in the full attention task only (not the divided attention task), and particularly when the judgement about the word's frequency was made quickly.

2.2. Collocation frequency

It will be recalled that part of Richards' (1976) second vocabulary assumption was: "For many words we also "know" the sorts of words most likely to be found associated with the word" (1976, p.79). Very little has been done to statistically verify this part of Richards' assumption – the estimation of collocation frequency. Stubbs (2002b, p.215) notes the lack of a 'phrase frequency list', as compared to word frequency lists and this, perhaps, has hindered interest in phrase frequency estimates in the past. However, there has been some research on this subject. Backman (1978) conducted research into the ranking of multi-word items, more specifically three word chains (in Swedish)[10]. Backman used the magnitude estimation technique (see section 2.1.1 for an explanation of this term), requiring 15 respondants to rank various three word Swedish combinations against an anchor. He found a correlation of .56 between the subjective and objective data. On the basis of this (fairly mild) correlation he suggests that collocations, "can be

[10] There are a mixture of collocation types in these multi-word items, e.g. *it may be, at heart, to devote oneself to, in the course of time, a great deal, of various kinds.* (It should be noted that these are English translations of the collocations used by Backman.)

supposed to have psychological counterparts" (1978, p.2). There is then, some prior research which suggests the sort of ability (i.e. collocation frequency estimation ability) questioned by Stubbs and Hunston, as mentioned in section 2 of this chapter.

2.3. Miscellaneous

As noted above a considerable amount of research indicates that subjective word frequency estimates are (on the whole) accurate, i.e. they are in line with the objective corpus data. In other areas of language too, subjective frequency estimation has also been demonstrated to broadly concur with objective measures. Attneave (1953) required 90 airmen trainees to estimate the frequency of the letters in the alphabet in 1000 words of text. He explained to them that if each letter were given equal weighting that it would be present 38 times (a useful note) and that the totals that the respondants provided should add up to 1000. The median judged frequency correlation between the actual frequencies of the words and the respondants' estimates was .79. In the area of event frequency, the findings are also positive with regard to subjective frequency estimates. Accurate estimation of matters as diverse as the number of restaurants in different fast-food chains (noted by Jonides & Jones 1992), and the ability to guess the age of people with certain names accurately[11] have also been observed (Cosmides & Tooby 1996). There are, however, a small number of studies indicating that frequency estimates may be weak both in areas of language (e.g. Tversky & Kahneman 1973) and in general event frequencies (e.g. Lichtenstein et al. 1978). These studies are discussed in more detail below.

3. The coding of and access to frequency information

The goal of this section is to explain, in simple terms, the explanations forwarded for the ability to judge frequency accurately, and also to explain why it may be that such an ability fails at times. Insights here could help explain the (claimed) mismatch between lexical intuitions and corpus data in frequency estimation activities.

[11] These showed an appreciation of generational name differences.

There are two distinct factors that bear upon the matter of frequency estimation. They are the representation of frequency in memory and the ability to access this representation. If representation of frequency is in some way 'privileged' and not easily confused, **and** if access is unhindered, then it follows that frequency estimates will (always) be good. If, on the other hand, either representation or access is 'compromised' in some sense, then it may be that estimates are more 'hit and miss'. The high correlation between word frequency estimates and objective data has spawned the theoretical interest in this specific subject, the coding of frequency information and access to that information. In the discussion below, there follows an overview of the key ideas in this area.

3.1. Coding of frequency information

3.1.1. Indirect and direct coding

There are two basic theories about the coding of frequency information: indirect coding theories and direct coding theories. Indirect coding theory posits that it is not frequency per se which is coded, but the traces of an event which are recorded. The repetition of the event leads to a trace being multiplied or simply strengthened over time (or a combination of both, as argued, for example, by Howell (1973)). In this model, frequency information is different from 'normal' propositionally encoded information (e.g. that Jack's birthday is in March). One of the reasons why some researchers doubt that frequency is coded directly (i.e. like 'normal' propositional information) is that if this were so, the encoding would be optional, i.e. one could choose to ignore frequency information. Part of the distinctiveness of frequency information, Hintzman (1978, p.548) argues, lies in its being 'obligatory', i.e. the coding of frequency is an automatic process.

3.1.2 The automatic encoding of frequency

As noted above, some researchers believe that frequency information is coded automatically. This particular theory is typically attributed to Hasher, Zacks and colleagues, although they note that it was Posner & Snyder who termed some cognitive

processes 'automatic' (Hasher & Chromiak 1977, p.173). Hasher & Zacks (1979, p.358) believe that the encoding of frequency is one of these processes. There are several reasons why they hold this view. The first is that frequency estimation seems immune to attention: the instruction to attend to frequency information in psychology experiments seems to have no effect on resulting performance (Hasher & Zacks 1984, p.1374; Zacks et al. 1982, p.106). Hasher & Zacks state:

> We assume that for an automatically encoded attribute to enter long-term memory, the person must be attending to the input in question. For such encoding it does not matter whether attention is incidentally or intentionally guided. If an individual is attending to an input, the encoding of frequency ...requires little or no specific further attentional processing (1979, p.359).

Frequency tagging appears to persist when attention is following a conversation (Hasher & Chromiak 1977, p.173), and is not hindered by other demands on attention (Zacks et al. 1982). These kinds of attributes in memory tasks are, what Hasher & Zacks (1984, p.1380) term, 'atypical'. The fact that age, ability, education and mood (Hasher & Zacks 1979), intoxication (Hasher & Chromiak 1977, p.179) and learning disability (Zacks et al. 1982, p.115) seem to make no difference to frequency estimation ability supports the view that frequency coding seems automatic. Hasher & Zacks (1979, p.369) believe that their theory of automatic frequency encoding is supported by some of the studies referred to earlier, e.g. Shapiro (1969), on the ability to rank words according to their relative frequencies, and Attneave (1953), on the ability to estimate letter frequency.

3.2. Access to frequency information

Encoding is important, but representation must not be equated with accessibility. Brown (1995, p.1540) believes that a variety of strategies can be employed in *accessing* frequency information. Broadly speaking there are two different approaches: enumeration and non-enumeration strategies. In the former, prior events are retrieved and counted. Such a strategy is not of particular interest to us here, as it is generally recognized that in normative judgements of word frequency, the numbers involved would be so high as to

disallow the use of such a strategy (see Toth & Daniels 2002, p.848). Non-enumeration strategies have been divided into *direct retrieval* strategies and *memory assessment* strategies. Once again, as with enumeration strategies, it has been questioned whether subjects could employ a direct retrieval strategy about normative frequencies of words in the language, as the numbers are so large. It is memory assessment strategies in particular that have received far more attention in the literature investigating normative frequency estimates, and heuristic strategies in particular.

Of particular interest to us here is Tversky & Kahneman's (1982, p.18) view that an availability heuristic may be employed in making frequency judgements, which, they argue, may occasionally lead to errors. While the evidence reported up to this point has been favourable regarding the accuracy of subjective frequency estimates, some data do not fit in well with this view. The names of Tversky & Kahneman have become synonymous with the use of heuristics in judgements of uncertainty. Heuristics have been defined as, "strategies that simplify complex tasks and get the job done well enough – they don't optimize they do 'satisfice'" (Cosmides and Tooby 1996, p.11). Tversky & Kahneman (1982) forwarded the idea that three heuristics are employed in judgements under uncertainty: availability, representativeness and anchoring, and adjustment. Of particular interest to us is the availability heuristic. Tversky & Kahneman explain how this operates in the following way:

> A person could estimate the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences by assessing the ease with which the relevant mental operation of retrieval, construction, or association can be carried out. A person is said to employ the availability heuristic whenever he estimates frequency or probability by the ease with which instances or associations could be brought to mind (1973, p.208)[12].

They go on to note that availability, while positively related to frequency (i.e. what is more frequent is more available) is also affected by other factors, (e.g. salience) and such factors may affect how frequent an event appears to be (1973, pp.207, 208; 1982, p.11).

[12] Note also N. Ellis (2002a, p.317): "We have no conscious access to the frequencies represented in our language processing systems, so we have to generate some exemplars in order to scrutinize them". The ease of generating those exemplars is the distinctive contribution of Tversky & Kahneman's availability theory.

This is an extremely important insight and sets up the possibility (on principled grounds) why it might be that subjective estimates and objective data (of whatever kind) can differ: the key factor affecting the quality of the judgement is the availability of relevant instances. The most relevant of their experiments to our research interest is one in which Tversky & Kahneman (1973) asked respondants to judge whether the relative frequencies of 5 different consonants were greater in first or third letter position of non three-letter words (the consonants being r, k, n, l, v). The respondants were also asked to give a ratio indicating how much more frequent the tendency was for a word to begin with, or have as its third letter, these different letters. Of the 152 subjects, 105 judged the first position to be more probable by a median average of 2:1. All of the letters are more common in third, not first position. Tversky & Kahneman explain the result by arguing that the 'first letter' examples of words are more available than their third letter position counterparts, and that this explains why respondants' frequency estimates were biased. As a result their estimates were biased in a way that the third letter instances were underestimated. Tversky & Kahneman also note research which showed that editors believed *'re'* to be more common at the beginning of words than at the end of words in their own writing. This also was wrong, and Tversky & Kahneman suggest the same reason for this error: availability (1973, p.212) – words ending with *'re'* seem not to have been so available in the searches upon which the estimates were based. More recent research also supports this view. Wänke et al. (1995) found that their German respondants reported it more difficult to produce German words with *'t'* in third position, rather than in first position, even though the frequency of *'t'* in third position of German words is nearly four times that of *'t'* in first position.

As noted in passing above, one way in which information may *appear* to be more frequent than it really is, is because of its saliency. Tversky & Kahneman (1982, p.11) note, for example, that seeing a house on fire, (as opposed to reading about a house burning down) is likely to affect ones ideas about how common or rare such an event is, and Taylor (1982) explains this concept in the following way: "Salience biases refer to the fact that colorful, dynamic or other distinctive stimuli disproportionately engage attention and accordingly disproportionately affect judgement" (p.192). An experiment

supporting this view is that of Lichtenstein et al. (1978) who researched frequency estimates on the causes of death (another experiment where frequency estimates were found to be poor). Lichtenstein et al. suggest that more memorable or dramatic incidents may appear to be more frequent than they actually are, because these kinds of memories have such a salient representation in memory (1978, p.552). On the basis of the above research, it would seem that accessibility or availability may be a key factor in affecting frequency estimation, and as such, it may be a key factor in explaining, post-hoc, the elicited-data corpus-data mismatches in the matter of frequent collocations. More particularly, it may be that salient meanings or uses of a word, or salient combinations will be considered *more frequent* than corpus data suggests. Further, it may be that words or combinations which are embedded in collocational frameworks are not so evident in frequency searches, e.g. the collocation *similar fashion*, typically occurring in *in a similar fashion* may not be so accessible as cases where *similar* is in a combination which can function as a noun head, (e.g. *similar ideas*). Finally, if it is the case that items fuse (or if a chunk has never been broken down), and if the resulting holistically stored item does indeed have weaker connections to the component words, then searches for the typical partners of a component word, which occur in the larger chain, may not be so successful.

4. Summary

As Halliday & Hasan (1976) note (appropriately, in the light of our discussion) "We have a very clear idea of the relative frequency of words in our own language" (p.290). In this chapter I have described theoretical principles forwarded to explain such abilities in word frequency estimation tests. However, in some specific cases it has been noted that frequency estimates may fail and a theory has to be able to account for this. The availability heuristic theory is one which, if considered in conjunction with either the Bybee theory or the Wray theory as noted in chapter 3, section 4.3, has the potential to explain why it may be that intuitions about frequent collocations may on the one hand be accurate *at times* (see chapter 1, section 3.1), and, on the other hand, sometimes be wrong. The key difference in the two cases, it is suggested, is the availability of the information required to make the judgement (accurately). Some key information *may* be

more hidden than other information in efforts to recall instances and some other information may be *more* accessible than it is frequent, perhaps due to its saliency. In the next chapter I report on an experiment testing the ability to rank collocations in which hypotheses formed on the discussions in this chapter and chapter 3, are tested.

Chapter 5 – Ranking Restricted and Free Collocations

1. Introduction

Some studies looking at the phenomenon of collocation have utilised statistical measurements to indicate the strength of attraction between the 2 words (e.g. z, MI scores[1]) rather than raw frequency of co-occurrence data. The formulas for calculating the z-score and MI score are provided in appendix 1. Generally, the value of raw co-occurrence data has been superseded by the use of these statistical measures in the study of collocation. Before conducting research into the ability to rank collocations it is important to understand how these different statistical tests work, and how the information they provide may be different from raw frequency of co-occurrence data. In section 2 this subject is discussed, and following on from a comparison of the collocates returned by different statistical tests, it is argued that raw frequency co-occurrence data is the best measure against which to compare lexical intuitions about collocations. In the second part of the chapter, the first experiment of the thesis is reported, which was designed to examine the ability of respondants to rank collocations (2 'restricted' sets and 2 'free' sets) against an objective standard.

2. Statistical measures of collocation strength and raw frequency co-occurrence data

The idea that raw frequency data and strength association data are two different things is forwarded by Lewis (1997). He suggests that teachers need to differentiate the two: "frequency alone is only a poor guide to the strength, and corresponding pedagogic usefulness. Teachers need to be aware of both strength and frequency when directing learners' attention to collocations" (p.30). In what follows, the results from different analyses of the same corpus are discussed.

The simplest analysis of collocations has been with raw frequency co-occurrence data. A notable proponent of this methodology is Stubbs (1995a) whose research has utilized raw

[1] I note these scores in particular, as they are the statistical tests available with the BNC world edition.

co-occurrence data in investigating collocations. While not discounting the value of statistical tests of association strength such as the z-score and MI score, Stubbs points out that the foundation upon which they depend, a calculation based upon observed frequency (O) and expected frequency (E), is fundamentally flawed when looking at words in corpora. Stubbs argues that it is wrong to believe that we can use 'expected' in a normal statistical way, when talking about text. He comments: "standard statistical procedures assume proper random samples in which values are independent observations, but since textual data are never in this form, this calls into question whether such statistics can reasonably be used on language data" (1995a, p.31). Barnbrook (1996) makes a similar point: "Even if the specific word being used as the node had no collocational effect on the words around it, the grammar of the language would constrain the types of words in different ways depending on the grammatical properties of the node word" (pp.92, 93). Stubbs believes that raw frequency co-occurrence should not be sidelined, arguing that results from statistical tests of association strength can 'hide', or misrepresent the reality (1995a, pp.37, 40). He adds: "for lexical words, use of raw frequency of joint occurrence as a statistic is unlikely to lead to any significant collocates being missed" (ibid, p.38).

Those who have questioned the value of raw frequency co-occurrence data do so because such figures are insensitive to the overall frequency of the individual words in the corpus. For example, Stefanowitsch & Gries (2003, p.216) argue that raw frequency analyses will just turn up function words as the most frequent collocates of a node word, primarily because they are so frequent in the language anyway. This is not a problem for the experiments reported on in this thesis as the raw frequency data can be trimmed of non-eligible partners: if respondants are asked to provide noun collocates for adjectives then their responses can be compared with noun data from the corpus (i.e. grammar words can be excluded).

However, there is a potentially bigger problem here. The above noted 'insensitivity' suggests, implicitly at least, that raw frequency co-occurrence data may well simply return higher frequency words as frequent collocates than lower frequency words, i.e.

high frequency lexical words will push their way to the top of a collocation frequency list, simply on the basis of their frequency in language – not because of their strength of association with the word per se. If so, collocation raw frequency rankings will simply mirror the individual word frequencies of the collocate of the node word. For example, because *way* is more frequent than *thing*, and the adjective *great* can qualify both, then it might be that the higher frequency of the word *way* in the corpus will mean that it is more likely to occur with *great*, and by sheer force of numbers (rather than attraction strength), it will push its way higher up the frequency collocation rankings. This is important to investigate, as if this is the case, then collocation frequency estimates, will, potentially, be confounded with word frequency estimates. An argument against this would be that words are frequent because of their occurrences in certain (specific) chains or collocations, ie. they are not *indiscriminate* in their lexical associations. Clearly, though, the matter should be investigated.

In Table 5.1 below, the collocates which are the most significant for three different analyses of *great* in the BNC are provided. The most frequent noun collocates of *great* in the BNC are listed in the first unshaded column, along with the number of instances in the BNC in the second column. In the third column, the collocates with the highest MI scores are listed. In the fourth column the MI score for each collocate is provided (it should be noted that the MI score is the same for all the words) and the number of instances of the collocation in the BNC is indicated in the fifth column. In the sixth column the most significant z-score collocates are listed, with the actual scores provided in column 7. The number of instances of these collocations in the BNC is provided in column 8.

Table 5.1. The most significant collocates of *great* in the BNC according to different statistical measures (+1 right window collocate search).

Raw Frequency co-occurrence		MI			Z score		
Collocate	BNC instances	Collocate	MI	BNC instances	Collocate	z-score	BNC Instances
Deal	2673	Daybog	41.1	7	Deal	1048.6	2673
Majority	389	Engeham	41.1	6	Hural	244.4	44
Success	370	Brington	41.1	5	Casterton	244.3	32
Interest	322	Dixter	41.1	5	Witcombe	213.6	20
Importance	317	Voltigeur	41.1	4	Majority	199.1	389
Care	290	Wonder-rabbi	41.1	3	Chesterford	193.2	16
Difficulty	260	Instauration	41.1	2	Fun	171.2	238
War	257	Destriers	41.1	2	Difficulty	167.3	260
Fun	238	Whirlo	41.1	2	Enchanter	165.6	18
Man	189	Dalangs	41.1	2	Importance	164.1	317

Clearly the different measures used to investigate collocation strength / frequency, turn up quite different collocates at the top of their lists. As can be seen, both the MI and z-score calculations return a large number of place names (e.g. *Great Daybog, Great Hural*). It is also clear that the highest MI scores are actually for very infrequent collocations. It is an acknowledged fact that MI can be high because of the infrequency of the collocate item – a problem noted by Church & Hanks (1989) and Stefanowitsch & Gries (2003, p.217). It is somewhat ironic, therefore, that Kennedy (2003) in his study on amplifier adverbs comments, "some of the collocations that contain the strongest bonds as measured by the MI score ....are in fact infrequent" (p.483). He would be more accurate in saying that '**because** certain items are infrequent they have high MI scores'. What Kennedy makes as a passing comment is in fact a fundamental fact of using the MI measure in the first place: it is no co-incidence as he seems to imply.

Returning to the table, this 'problem' with the MI score can be explained quite easily. For example, the fact that *instauration* is itself rare (occurring twice), makes its combination with *great* (on both occasions) very significant for the MI score – indeed all of the top 10 MI collocates only occur with the adjective *great* in front of them in the BNC. The MI score falls very gradually, as soon as the exclusivity of the relationship weakens. For example, *Winglebury* occurs three times in the corpus, and on two occasions it is directly preceded by *great* – the resulting MI score is 40.8. *Totham* occurs 5 times, and on two occasions *great* is immediately to the left, the MI score 40.3 and so on. Church & Hanks (1989) note that the MI score calculation becomes "unstable when the counts are very small" (p.77) stating that there must be at least 5 occurrences of a word before the MI calculation can be conducted with any reliability.

With regards to the high ranking z-score collocates, several observations can be made. Firstly, it should be noted that the collocate with the highest z-score is actually the most frequent collocate according to raw frequency of co-occurrence data. In addition, there are a number of other collocates with high z-scores which are listed in the first column (ie. *majority, fun, difficulty* and *importance*). However, using the z-score calculation also returns quite infrequent collocates as having high z-scores (e.g. *Hural, Casterton, Witcombe*). Because it is recognized that MI, and to a lesser extent the z-score, inflates the importance of these low occurrence items, it makes sense to trim the findings. The question is at what point? In Table 5.2 below, I exclude place names and draw a cut off point at 50 attested BNC instances for MI and z score collocates to be recorded, to see whether the resulting sets of collocates returned are more similar to the raw co-occurrence data.

Table 5.2. Raw frequency, MI and z-score data for collocates of *great* with at least 50 instances in the BNC, excluding place names

Raw Frequency co-occurrence		MI			Z score		
Collocate	BNC instances	collocate	MI	BNC instances	Collocate	z-score	BNC Instances
Deal	2673	Deal	39.3	2673	Deal	1048.6	2673
Majority	389	Lengths	38	77	Majority	199.1	389
Success	370	Fun	37.7	238	Fun	171.2	238
Interest	322	Difficulty	37.5	260	Difficulty	167.3	260
Importance	317	Majority	37.4	389	Importance	164.1	317
Care	290	Pleasure	37.3	187	Success	162.5	370
Difficulty	260	Pity	37.3	68	Strides	146.4	51
War	257	Importance	37.2	317	Pleasure	134.0	187
Fun	238	Bulk	37.0	61	Admirer	123.6	37
Man	189	Expectations	36.9	92	Lengths	109.3	77

The trimming results in six words becoming common to all the lists: *deal, fun, difficulty, majority, importance* and *pleasure*. Only one additional word is common to the z-score and MI lists, and not in the raw frequency column (*lengths*). The Raw frequency, MI and z-scores have the same collocate at the top of their ranks, and it can be noted that the raw frequency and z-score have the same collocates in positions 1 and 2. The unique occurrences in the table above are noted below in Table 5.3.

Table 5.3. Unique instances of collocates among the 10 highest ranked collocates according to different calculations

Raw	man, care, importance
MI	pity, bulk, expectations
z-score	strides, admirer

Man, care and *importance* have higher individual raw frequencies of occurrence than *pity, bulk, expectations, strides* and *admirer* – hence they are picked up with the raw frequency co-occurrence count. The less frequent words are picked up by the MI and z score because of the way that they calculate the strength of association. All of these measures, no doubt, tell us useful information, and the idea that a multiplicity of measures gives us a more complete picture about a word's collocates has been argued by Cantos & Sánchez, (2001, p.202) and Barnbrook (1996, p.101). Though, as noted above, Stubbs (1995a) is critical of these statistical tests he argues that the tests utilising O/E give, "a (rough) indication of the strength of the association between two words" (p.33).

Regarding the possibility hinted at by Stefanowitsch & Gries, that more frequent words will become frequent collocates solely by virtue of their raw occurrence frequency, it should be noted that only one of the most common nouns of English (according to the BNC – *time, year, people, way, man*) is present in the above collocate listings (*great man* 189 instances), though all the other highly frequent nouns are semantically possible collocates of *great*, and are attested in the BNC: *great time* (113 hits), *great year* (12 hits), *great people* (21 hits) and *great way* (48 hits). This suggests that raw frequency co-occurrence figures, while not picking up the more infrequent noun collocates with a strong attraction to the node word, will not necessarily just turn up the most common words in English. This adds support to the idea that words are frequent because of their occurrences in certain preferred phrases, collocation frameworks etc. (for more on this see chapter 7, section 2.7). However, what of the words that are frequent collocates? Is it the case that there is a strong correlation between the raw frequency co-occurrence rankings and the relative ranks of the nouns (alone)? In Table 5.4 below, the nouns are ranked firstly according to their relative frequencies as they occur in the collocation *great NOUN*, and then, in the second column as they occur in the BNC in their individual noun frequency counts.

Table 5.4. A comparison of the relative frequencies of the most frequent noun collocates of *great* and the frequencies of the nouns (alone) in the BNC.

Collocate (with *great*) and BNC instances (ranked most frequent to least frequent)	Nouns from column 1 ranked according to the noun instances in the BNC (from most frequent to least frequent)
Deal (2673)	Man (57699)
Majority (389)	Interest (26459)
Success (370)	Care (14991)
Interest (322)	Success (13239)
Importance (317)	Majority (9831)
Care (290)	Importance (9573)
Difficulty (260)	Deal (7305)
Fun (238)	Difficulty (6229)
Man (189)	Pleasure (4897)
Pleasure (187)	Fun (2039)

The Spearman Rho correlation score for these ranks is .556 (two-tailed not significant at p=0.05). As can be seen there is a mild correlation but it gives us no warrant to believe that raw collocate frequency data for a particular node word simply mirrors raw frequency noun data for the collocates, and as such, this finding suggests that a comparison of lexical intuitions with raw collocation data is a valid research focus. In addition to the importance of this finding in affecting this decision, for other reasons too, this focus seems appropriate:

1. The focus on raw frequency is methodologically necessary as it is much easier to explain a task to respondants in terms of frequency of co-occurrence than strength of association.

2. It is theoretically desirable: it enables us to draw on the findings of frequency theory research as described in chapter 4.

3. It enables us to compare the research with word frequency research, as the latter is based on raw frequency data.

4. It enables us to use search engine data as a secondary check on the BNC data.

Clearly, though, it must be recognized that strength of association or the frequency of the collocate of the node, rather than its co-occurrence frequency might affect collocation frequency estimates and references are made to such issues throughout the chapters that follow.

3. Experiment 1

3.1. Rationale for experiment and hypotheses

Conducting this experiment enables us to investigate whether Stubbs and Hunston are correct in their view that native speakers of English cannot document the frequency of collocations. While there is some evidence that the ability to accurately rank collocations has already been established (Backman 1978), that research did not investigate qualitative issues, and many of the collocations which Backman asked his respondants to rank were *grammatical* collocations. In this study I focus on *lexical* collocations alone. The following hypotheses are made:

1. Because of their 'unit-like status', restricted collocations will be ranked more accurately than their 'free' counterparts. In chapter 3, section 4.1 it was noted that some writers (e.g. Kjellmer and Howarth) believe that restricted collocations are more likely to be lexicalised. In a ranking task of the nature proposed below, one would assume that, if the frequency of the free collocations has to be computed 'online', as opposed to the chunk being recalled whole, then there will be a greater likelihood of error in ranking the frequency of the free collocations. It is more likely that an item in the free collocation, rather than the collocation itself will be ranked. For the restricted set of collocations this will not be the case. The assumption is made that the subjects already possess knowledge

of the frequency of a restricted collocation because it acts as a single unit and thus has its own frequency representation. It is hypothesized that, because of their (hypothesized) privileged psycholinguistic status, restricted collocations will be ranked more accurately than a set of free collocations.

2. There will be greater accuracy in ranking collocates which are complete units (e.g. *good luck*) rather than those which typically occur in larger chains of language (e.g. *good mind* which typically occurs in *I've a good mind to [verb]*). These latter types of collocation will be *under*-ranked as the collocate subcomponents of the larger chain within which they occur may be less noticed, i.e. we hypothesize an availability restriction for items in these collocation frames in accordance with the views of Wray and Bybee and the use of an availability heuristic in the making of the judgements.

3. Collocations which are more 'salient' are more likely to be over-ranked, i.e. considered to be more frequent than they are in the objective standard. How exactly a collocation can be more salient than another is, at this stage, not particularly clear. As Bley-Vroman (2002) notes, "The mechanisms that are hidden behind the word "salient" remain largely mysterious" (p.213). However, it may be, for example, that collocations containing concrete, as opposed to abstract nouns will be over-ranked. It has been argued that concrete nouns may have image and verbal codes making them more distinct (Hamilton and Rajaram 2001, p.113; Jessen et al. 2000, p.104). If this is the case, then collocations containing concrete nouns may be over-ranked.

3.2. Experiment design

3.2.1. Choice of adjectives and collocations

To enable a comparison between a restricted set of collocations and a free set (and in order to minimize the effect of having a different node word in the sets – i.e. the word that remains constant), the decision was made to create two sets of collocations in which the same adjective is used, but in one set the collocations are classified as restricted and

in the other free. As noted in chapter 3, section 2.3, there is no generally accepted definition of free and restricted collocations and writers may well classify collocations differently. There are, however, several well-known collocation dictionaries. One of these, written by Benson et al. (1986) claims only to contain restricted collocations: "The combinatory dictionary does not include free lexical combinations. Free lexical combinations are those in which the two elements do not repeatedly co-occur; the elements are not bound specifically to each other; they occur with other lexical items freely" (1986, p.xxiv). While one may not always agree that the items included in their dictionary are restricted collocations[2], it was decided, on balance to use this resource in designing the sets of collocations to be ranked. If a collocation was in the dictionary it was deemed 'restricted' and if it was not found in this dictionary it was deemed to be 'free'.

In generating a set of collocations to be ranked it was felt best to keep one word constant in all the collocations. The decision to keep the adjective constant in the collocation (rather than keeping the noun constant) was made on the grounds that keeping the initial word in the collocation the same should ease the availability of instances when the searches for the collocations are made. Finding the adjective entries in the Benson et al. (1986) dictionary was a long and tiresome task, as entries are not arranged by adjective entry, but rather by noun entry. This meant that the whole dictionary had to be read and reread to find the adjective collocations that Benson et al. deem to be restricted for hundreds of nouns.

3.2.2. Differences in frequency between test items

For a robust experiment, within the bounds of practicality, the differences in frequencies between the collocations in the different sets should be identical, as should the frequency spacing between the items in the lists. Typically, word frequency estimation experiments have used words whose relative frequencies have been around twice as frequent as the

[2] See for example Cowie (1998), who argues that their dictionary actually includes free collocations (1998, pp.226, 227) though, of course, according to his definition of the term.

next closest item in the set to be ranked. For example, the average difference between the words used in Frey's (1981) set, in terms of the relative raw frequency differences between the two closest items in the set is between x2 and x4. In Ringeling's research the average difference between the proximate words is x1.7 times[3]. To enable us to make some sort of comparison with these experiments[4], it made sense to aim for similar figures in terms of the differences in the relative frequencies of the proximate items in the lists.

Unlike words, collocations do not have such high frequency representation in a corpus. For example, though there are 14171 instances of *news* as a noun in the BNC, there are only 637 instances of *bad news*, and this is a fairly frequent lexical collocation, relatively speaking. This means that we would only be able to provide 8 or 9 collocations to rank, if the criterion for differences in relative frequency was x2-3 between the items in the list to be ranked. This can be done, but what is problematic is putting any confidence in the lower ranks of the set, when the data is so sparse: if there is only 1 instance of a collocation in the corpus, does this really mean that it is less frequent (in the language as a whole) than a collocation which has three instances? Because of this problem and because the results from research indicate that internet search engines have high correlations with more traditional corpora (see chapter 2, sections 7.2.1, 7.2.2), and they provide much more data, it was decided to use the Altavista search engine as a secondary check on the BNC figures, particularly when the frequencies were lower – indeed to defer to these findings when comparing them with the lower frequency collocation instances in the BNC. In effect, this limits our focus to strictly adjacent collocating items in the BNC, because a collocation window cannot be set in an Internet search, and it makes sense to adopt the same search methodology in both corpora.

3.2.3. Collocation frequency and raw noun frequency

As noted in section 2, there is a possibility that the relative frequencies of collocations in which one word remains constant, may reflect the frequencies of the variable items in the

[3] This is a calculation based on the numbers he provides.
[4] These studies in particular are important as the methodology of the research reported here and that of Ringeling, Frey and Arnaud are very similar.

different collocations. For example, below are some of the collocates of *young* taken from the BNC (at intervals of x10 frequency differences approximately), where numbers in brackets indicate instances in the BNC.

Young people (3613)
Young girls (272)
Young baby (27)
Young pupil (3)

If we run a raw noun frequency check, as done below (numbers after the words are noun search instances in the BNC), it can be seen that the ranks of the nouns alone mirror the ranks of the collocations

People (121774)
Girls (9081)
Baby (8604)
Pupil (2307)

Such 'mirroring' may not always occur (see Table 5.4); however, it is an essential part of the design process to ensure that the correlation between the collocation set and the raw noun set be non-significant. This means that it must not be possible to predict the order of the collocation set at a statistically significant level by using the frequency rankings of the nouns alone. Below is a summary of the criteria determining the choice of the collocations to be ranked:

1. A free and restricted set of collocations for the same adjective must be provided, with the same number of collocations in each set.

2 The frequency difference between each collocation should be at least x2 approximately, to enable comparisons with the comparable work on word frequency estimate research.

3. It must not be possible to rank the collocations 'successfully' (i.e. in a statistically significant way), with recourse only to the frequency ranks of the nouns.

4. BNC and Altavista must agree on the ranks. On the lower frequency items, where BNC data is scarce, the Altavista hits determine the ranks.

5. The adjectives must be very frequent, otherwise there will not be enough data in the BNC.

Although this sounds quite simple, it was, in practice extremely difficult to construct and was very time consuming. After much searching through dozens of possible candidates, a free set and a restricted set for the adjectives *personal* and *bad* were constructed. Ideally one would wish to test more; however, in practice, adherence to the above criteria disallowed the inclusion of other sets of collocations for testing.

3.2.4. The sets for testing

The collocation sets used are listed in Tables 5.5 to 5.8 below. In each table the collocations are ranked from highest to lowest frequency. In addition, the correlation is noted between the relative frequencies of the nouns in the set (i.e. their relative ranking according to their noun search instances in the BNC) and the relative frequencies of their collocate ranks in the collocations from both BNC and (where possible) Altavista.

Table 5.5. Set 1 – Free collocations of *Personal*

Collocation	BNC instances	Altavista Exact phrase hits 23/03/04
Personal experience	262	486,932
Personal problems	88	94,863
Personal letters	28	34,123
Personal number	14	12,497
Personal disaster	7	3,480
Personal corruption	3	837
Personal initials	2	367
Personal minute	2	98

Correlation of 'noun alone' relative frequencies with collocation frequencies:

- BNC = .57143 (0.05 one tailed significance level is .643), not significant.

Altavista – Not possible because *minute* is also an adjective.

Table 5.6. Set 2 –Restricted collocations of *Bad*

Collocation	BNC instances	Altavista Exact phrase hits 23/03/04
Bad news	637	738,240
Bad luck	266	221,966
Bad habit	51	78,733
Bad form	16	24,174
Bad headache	11	8,994
Bad egg	2	3,468
Bad sport	1	1,337
Bad miscalculation	0	110

Correlation of 'noun alone' relative frequencies with collocation frequencies:

BNC = .4762 (0.05 one tailed significance level is .643), not significant.

Altavista = Not possible because *form* is also a verb.

Table 5.7. Set 3 – Restricted collocations of *Personal*

Collocation	BNC instances	Altavista Exact phrase hits 17/03/04
Personal computer	654	1,067,818
Personal life	171	318,254
Personal belongings	58	109,998
Personal liberty	25	29,025
Personal quality	8	10,670
Personal impression	3	3,924
Personal sorrow	2	1,225
Personal setback	2	371

Correlation of 'noun alone' relative frequencies with collocation frequencies:

BNC = .61905 (0.05 one tailed significance is .643), not significant.

Altavista = .54762 (0.05 one tailed significance is .643), not significant.

Table 5.8. Set 4 – Free collocations of *Bad*

Collocation	BNC instances	Altavista Exact phrase hits 23/3/04
Bad idea	119	420,759
Bad reputation	46	79,982
Bad decision	22	26,990
Bad terms	15	7,538
Bad injury	7	2,170
Bad figure	2	888
Bad danger	2	106
Bad revelation	0	31

Correlation of 'noun alone' relative frequencies with collocation frequencies:
BNC = .5595 (0.05 one tailed significance is .643), not significant.
Altavista = Not possible because *terms* is also a verb.

3.3. Methodology

Several frequency estimation techniques have been employed in testing word frequency estimation as noted in chapter 4, section 2.1.1. The technique chosen here, mainly on the grounds of its simplicity, is a relative ranking from a random group of words, i.e. a methodology adopted by Ringeling (1984), Arnaud (1990) and Frey (1981). In such a test no anchor word is provided and the respondants are simply asked to rank the random listing of collocations in the set, from the most frequent to the least frequent, according to the actual rankings in the English language as shown by the BNC and Altavista search findings.

Brief background information about the BNC and Altavista was provided on the test paper, before the items for ranking were presented[5]. Four test papers were designed, testing the same words, but these were presented in different orders. The sequences of the collocations in the four sets were ordered randomly. Two of the four sets were placed on the front of a sheet of A4 paper, and two on the back. The order of the sets (as opposed to the order of the collocations within the sets) was constrained in that the restricted and free sets of the same adjective were never on the same side of the page. Secondly, the order of the four sets was varied[6]. An introductory explanation of the task and instructions on how to complete it preceded the sets. Respondants were asked to rank the items according to their frequency in the English language as a whole as evident in the BNC and Altavista (i.e. not according to their own personal usage of the items). The subjects tested were also asked to indicate how long it took them to complete the task.

3.4. Subjects

The first set of subjects comprised 16 male native speaker teachers of English for academic purposes at King Fahd University of Petroleum and Minerals, Saudi Arabia. They were either asked to do the task personally, or the task was left in their pigeonholes for them to collect and do in their free time. The task was not supervised.

The second set of subjects were undergraduate students at Cardiff University, enrolled in an 'Introduction to Language' course[7]. In addition to asking the students to record how long the task took them, they were also asked to indicate their sex and whether they were native speakers of English or not. Those indicating that they were non-native speakers were excluded from the analysis that followed[8]. Incomplete scripts or scripts which contained ranking errors (e.g. the same number twice) were also excluded. There were a

---

[5] See appendix 2 for the test paper.

[6] From left to right and front to back page: version 1: *personal* free, *bad* restricted, *personal* restricted, *bad* free; version 2: *personal* restricted, *bad* free, *personal* free, *bad* restricted; version 3: *bad* free, *personal* restricted, *bad* restricted, *personal* free; version 4: *bad* restricted, *personal* free, *bad* free, *personal* restricted.

[7] I would like to thank Prof. Wray for running this experiment for me.

[8] It was not considered appropriate to analyse the results of the non-native speakers, as some of the more infrequent nouns in the collocations may not have been known to them.

total of 115 papers which were included in the analyses: 87 of these were completed by females.

3.5. Results and analyses

Average times to complete the tasks were three minutes for the undergraduates and 6-7 minutes for the lecturers. Five quantitative analyses were conducted on the data and are reported below.

Analysis 1- Correct identification of the most frequent collocate

Table 5.9. Number and percentage of respondants correctly identifying the most frequent collocation for each of the four sets

	Number of respondants correctly identifying the number 1 rank word
Set 1 (*Personal* free)	67/131 (51.15%)
Set 2 (*Bad* restricted)	53/131 (40.46%)
Set 3 (*Personal* restricted)	27/131 (20.61%)
Set 4 (*Bad* free)	99/131 (75.57%)

Table 5.9 shows how many of the respondants were able to identify the most frequent collocation from the sets. The 4 sets each contained 8 collocations. Therefore, according to chance alone, we would expect a particular response to have around 12/13% of the votes. Chi-squared (χ^2) goodness of fit analyses were conducted to see whether the ability to choose the most frequent collocate was significant. In using the chi-squared test we must be sure that the expected number for each cell be no less than 5. This condition was satisfied. Items in cells are independent, and the actual numbers obtained are used.

Table 5.10. χ^2 figures for the choice of the most frequent collocate

	Set 1 – *Personal* free	Set 2 – *Bad* restricted	Set 3 – *Personal* restricted	Set 4 – *Bad* free
χ^2	178.87*.	93.57*	7.87*	476.28*

* ($p<0.01$ significance level = 6.64, df =1)

As can be seen, the ability to identify the number 1 rank varied widely. Using a chi-squared goodness of fit analysis, chi-squared is significant in all cases. Yates' correction for continuity was applied to Set 3, as this is only just significant[9]. The resulting χ^2 = 7.15, which is still significant. What this means is that the set of observed frequencies do not correspond to the expected frequencies. There is a significant difference between the responses and a normal distribution. However, contrary to hypothesis 1, respondants correctly identified the most frequent collocation for the free sets more accurately than for the restricted sets - see table 5.11 below.

Table 5.11. Comparison of number of respondants who correctly identified the most frequent collocates from the restricted sets and free sets.

Sets	Respondants choosing the most frequent collocation
Free	166/262 (63.36%)
Restricted	80/262 (30.53%)

[9] Yates' correction for continuity should be applied when df =1, but it was only calculated for the *personal* restricted set, as this 'correction statistic' has just a small impact on the resulting χ^2 figure. The other χ^2 figures are very high and, as a result, it was not considered worthwhile to recalculate the figures.

Analysis 2 - Spearman Rho correlation score: respondants ranks and the objective ranks

In order to see whether the individual subjects' rankings correlated with the objective rankings of the words in the BNC/Altavista, Spearman Rho correlations were conducted for each respondant's ordering of the items. This resulted in 131 x 4 (524) calculations being made. Upon completion of these calculations, the mean average correlation for each set of words ranked was obtained, together with the standard deviation. Table 5.12 below, indicates the correlation results for the 4 sets of collocations.

Table 5.12. Spearman Rho correlation figures for the 4 sets of collocations

	Set 1 (*Personal* free)	Set 2 (*Bad* restricted)	Set 3 (*Personal* restricted)	Set 4 (*Bad* free)
Total group (N=131)	0.778588 * sd = 0.148	0.760580* sd= 0.137	0.527649 sd= 0.251	0.845969 ** sd= 0.110

Average of 4 sets = .728197

Average of sds = 0.139

* Significant at $p<0.05$

** Significant at $p< 0.01$

For three of the four sets, the respondants rankings, when compared against the BNC / Altavista rankings, were statistically significant (once at $p<0.01$ the other two at $p<0.05$). The significant correlation figures for the *personal* free, *bad* restricted and *bad* free sets indicate that the respondants' ideas about the relative frequencies of the collocations in these sets concur (significantly) with the objective base. However, the '*personal* restricted' set was not ranked in a statistically significant way. While there was a mild positive correlation between the respondants' ranks and the objective data, this was not statistically significant.

Analysis 3 - Kendall's coefficient of concordance for the respondants

The results from analysis 2 indicate that for three of the four sets, there was a significant correlation between the BNC/Altavista data and the subjects' responses. What though of the similarity or differences between the different subjects' responses? Did respondant A have similar lexical intuitions to respondant B? The only reason why this should be investigated is to see whether there is a significant coefficient of concordance for the *personal* restricted set. We can assume, given the significant correlations between the objective ranks and the subjective estimates for the other three sets that the respondants' concurred in their judgement; however, such an assumption can not be made for the *personal* restricted set. Kendall's coefficient of concordance (W) is a non-parametric test of correlation. Using it we can check the similarity/difference between the 131 respondants. In mathematical terms, "The coefficient of concordance is a ratio of the variance of the sums of the ranks for the subjects....divided by the maximum possible value that can be computed for the variance of the sums of the ranks" (Sheskin 2000, p.897). It is an ideal way of comparing the ranks of the subjects, rather than comparing their ranks with the BNC/Altavista data (Analysis 2). The table below shows the Kendall coefficient for the 4 different sets of collocations.

Table 5.13. Kendall's coefficient of concordance (W) correlations

	Set 1 (*Personal* free)	Set 2 (*Bad* restricted)	Set 3 (*Personal* restricted)	Set 4 (*Bad* free)
Total (N=131)	.729*	.704*	.609*	.803*

*Significant at $p<0.001$

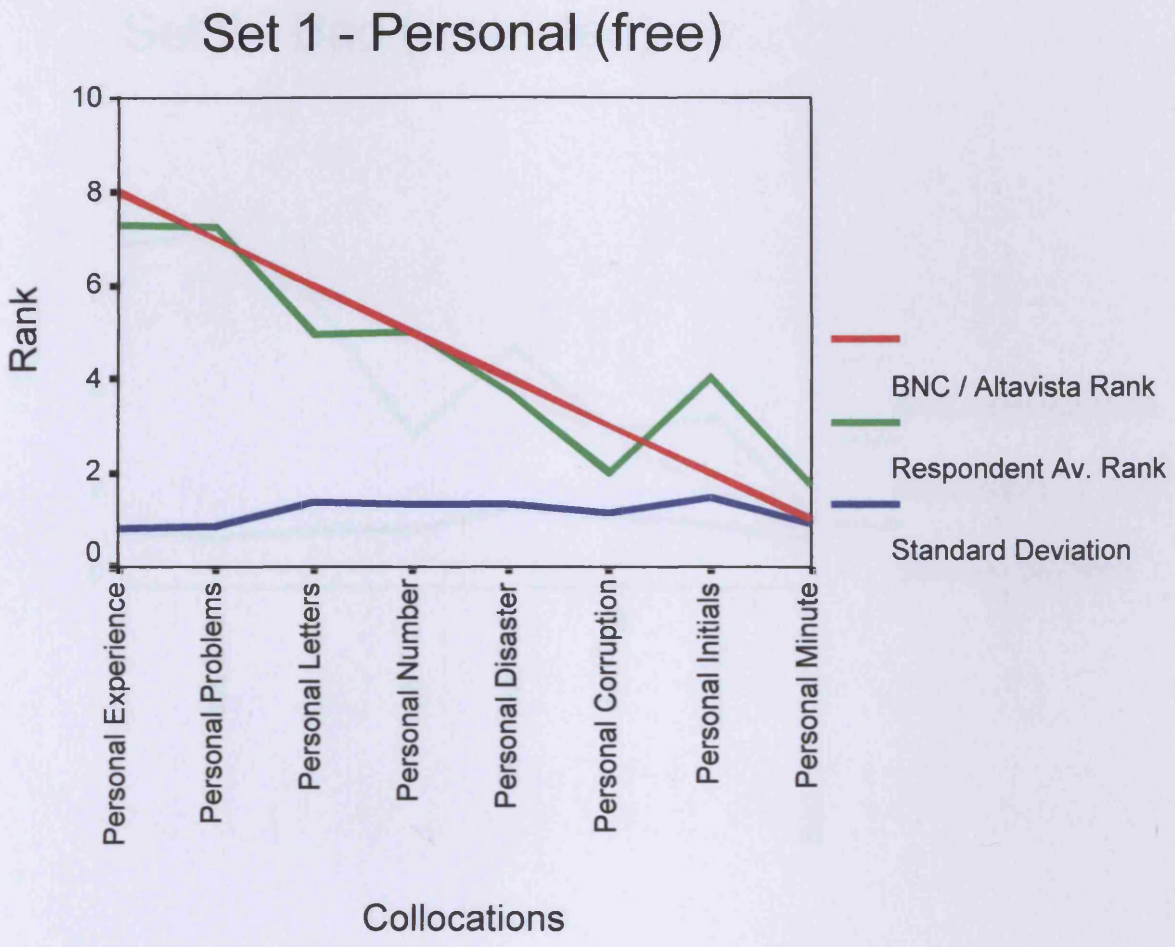
As would be expected, (given the results in analysis 2) W for sets 1,2 and 4 is significant. However, importantly, while the Spearman Rho correlation (Analysis 2) was not statistically significant for the '*Personal* restricted' set in which the respondant ranks were compared with the BNC/Altavista data, if we simply compare the subjects' rankings

(as a group) Kendall's coefficient of concordance is statistically significant. This means that when considered as a group there was significant agreement between the subjects about the ordering of the items, even though that ordering did not concur with the ranking from the objective data.

Analysis 4: Mean average ranks of the collocations according to the respondants and standard deviations for each word ranked

This analysis enables us to see if there were any particular collocations within the sets which the respondants ranked well, too high or too low. This is a key interest in the research (and enables us to test hypotheses 2 and 3 stated earlier). The data is presented in graph form to enable the reader to best appreciate how the items ordered fared against the BNC/Altavista rankings. Brief observations are made on these findings, and developed in more detail in section 6, below.

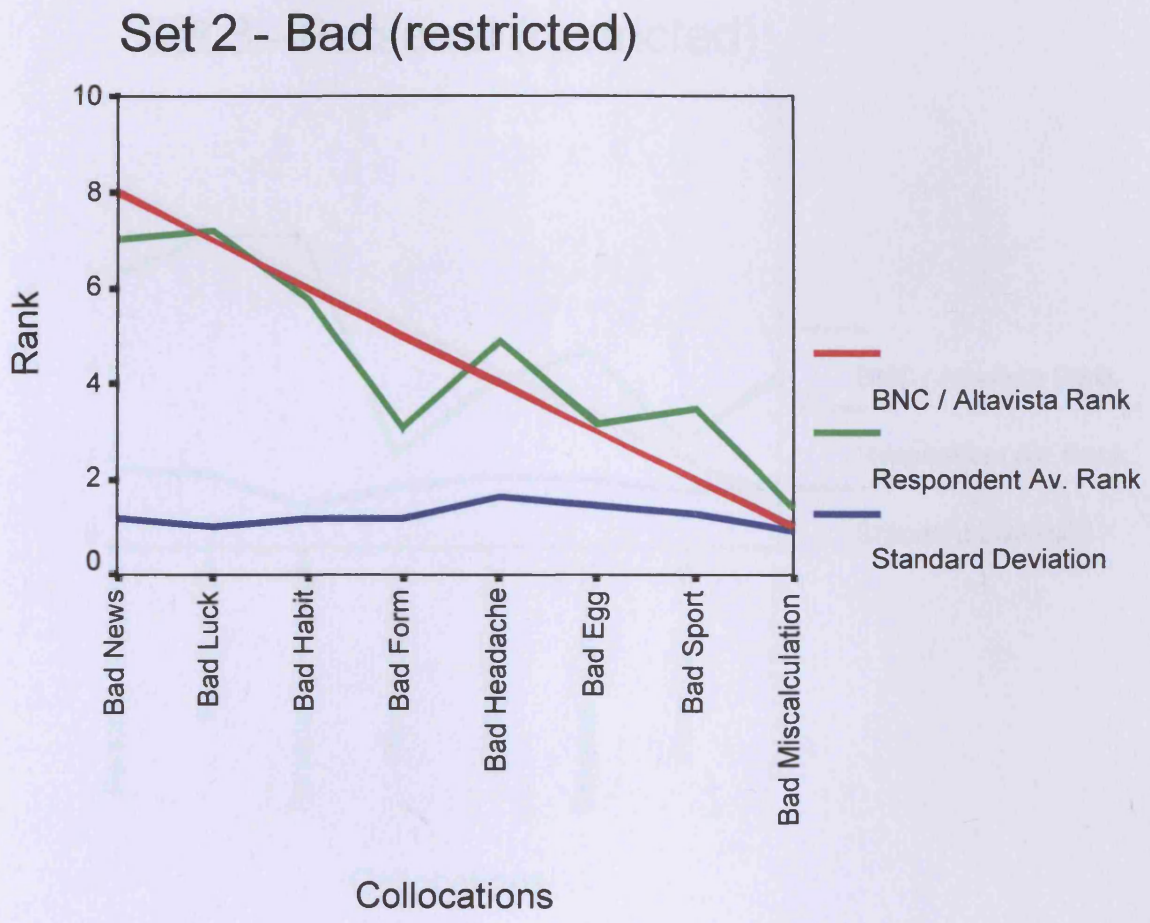
Figure 5.1. Mean average rank for each collocation, BNC / Altavista rank and standard deviations for Set 1



Mean average standard deviation = 1.161

The only collocation which was ranked noticeably differently from the BNC/Altavista rank was *personal initials*. The sd. of the responses for this word was also the highest (1.470), indicating that there were more differences between the respondants about how to rank this word than the other words. *Personal problems* and *personal number* were ranked the most accurately.

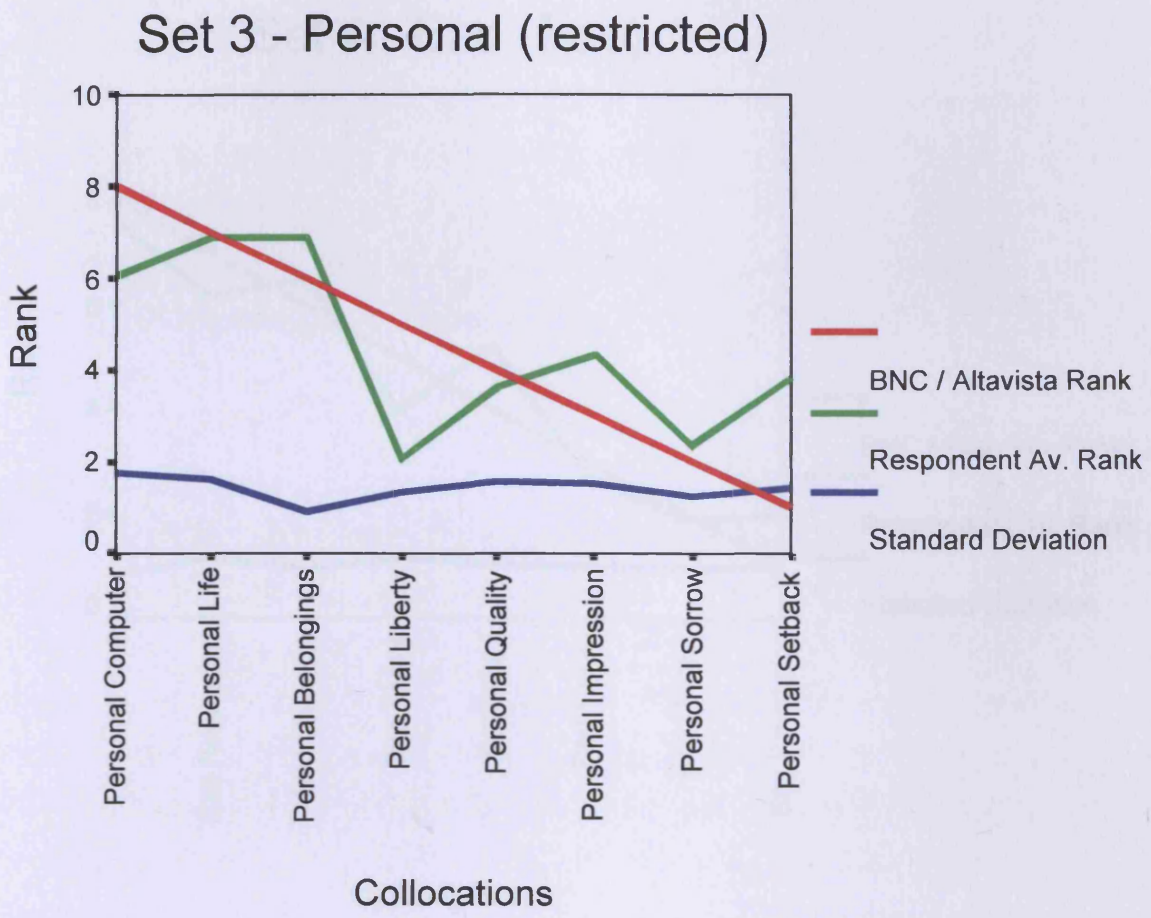
Figure 5.2. Mean average rank for each collocation, BNC / Altavista rank and standard deviations for Set 2



Mean average standard deviation = 1.233

The ranks provided for *bad form* were lower than for the corpus data, and the average rank for *bad sport* was higher than BNC/Altavista. The greatest variety in responses (i.e. the highest sd) was to *bad headache* (sd. = 1.617). *Bad luck, bad habit, bad egg* and *bad miscalculation* were the most accurately ranked collocations.

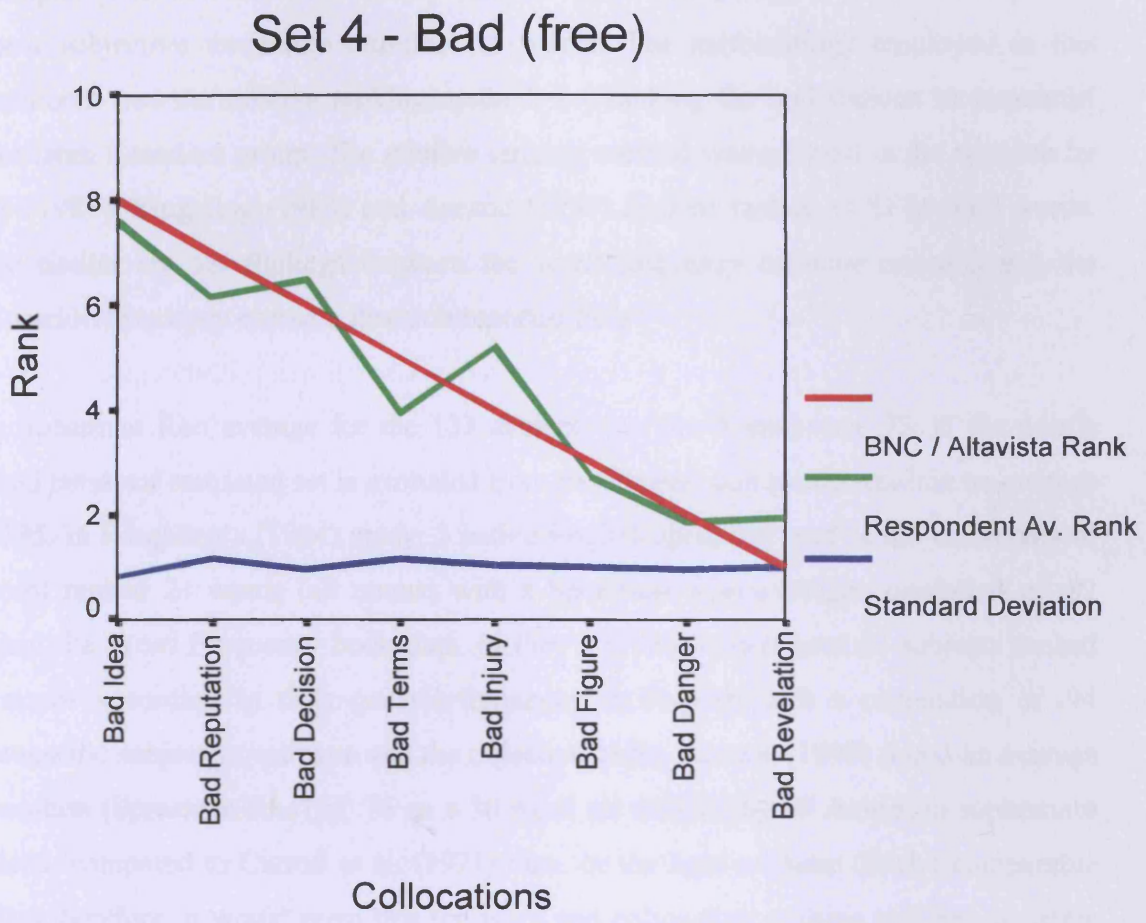
Figure 5.3. Mean average rank for each collocation, BNC / Altavista rank and standard deviations for Set 3



Mean average standard deviation = 1.412

Clearly, the mean average ranks for the collocations in this set show a considerable difference from the BNC/Altavista ranks. The collocations *personal computer* and *personal liberty* were ranked too low and *personal setback* and *personal impression* were ranked too high. The rankings closest to the corpus ranks were for *personal life* and *personal sorrow*.

Figure 5.4. Mean average rank for each collocation, BNC / Altavista rank and standard deviations for Set 4



Mean average standard deviation = 1.017

The standard deviations were, on the whole, lower for this set than for the other set, indicating that there was more agreement between the subjects on how to rank these collocations, compared to the collocations in the other sets. *Bad terms* and *bad reputation* were ranked a little low, and *bad injury* and *bad revelation* a little high. The rest of the words were ranked quite accurately.

Analysis 5 - Comparison of the collocation estimation ranks with comparable word frequency ranks research

In chapter 4, section 2.1.1, it was noted that different methodologies have been employed to test subjective frequency estimates of words. The methodology employed in this experiment was the relative ranking method, i.e. ranking the collocations in sequential order from a random group. The relative ranking method was adopted in the research by Frey (1981), Ringeling (1984) and Arnaud (1990) in their testing of SFEs with words. How similar are the findings between the word frequency estimate research and the collocation frequency estimate research reported here?

The Spearman Rho average for the 131 subjects (on the 4 sets) was .73. If the poorly ranked *personal* restricted set is excluded from the figures, this would result in an average of .795. In Ringeling's (1984) study, 5 native English speaking staff of the University of Utrecht ranked 24 words (all nouns) with a Spearman Rho average correlation of .82 against the Word Frequency book data. In Frey's (1981) experiment 46 subjects ranked 12 nouns according to their general frequency in English with a correlation of .94 between the subjective estimate and the objective order. Arnaud (1990) found an average correlation (Spearman Rho) of .76 on a 30 word set ordered by 87 American sophomore students compared to Carroll et al. (1971) data. In the light of these (fairly) comparable studies therefore, it would seem that the word and collocation ranking abilities of native speakers are similar. This finding provides empirical support to Richards' view that, "For many words we also 'know' the sorts of words most likely to be found associated with the word" (1976, p.79).

3.6. Discussion: testing the hypotheses

Hypothesis 1 stated that the collocations in the restricted sets would be more accurately ranked than the collocations in the free sets. This was not the case - indeed the opposite situation arose. Restrictedness (as defined by Benson et al.), in and of itself, does not seem to be a particularly important factor in affecting the ability to rank the words. As

such, the notion that restrictedness implies a more psycholinguistically real representation in the mind does not seem justified according to the results from this experiment. As noted in chapter 3, definitions of restricted collocation vary and it may be that different writers would have classified the collocations differently. But it may be that the factor of restrictedness is not really so important in affecting our lexical intuitions of collocation frequency. As suggested below, there are some plausible explanations for why some of the collocations were not well ranked, and these have nothing to do with the fact that these collocations were classified as restricted. It may be that only the more frequent collocations are lexicalised, and this may be independent of whether the collocation is restricted or free; however, this explanation fails to account for the inaccurate rankings of some of the most frequent collocations (noted in analysis 1 above). The results require us to look elsewhere for what might have been affecting the SFEs, and for why some combinations were ranked either too high or too low (i.e. we move on to examine hypotheses 2 and 3).

Is it the case that some of the inaccurate ranks can be explained with reference to the use of an availability heuristic in the frequency judgement? As noted in chapter 4 the availability heuristic hypothesis is compatible with both *under-* and *over*-ranking. An item may be over-ranked because it is salient, and under-ranked because it is not so available when exemplars are generated. Collocations for which the average rank was ± 1 whole rank different from the corpus ranks are noted below in Table 5.14 (*bad news* is also noted as it was .99 under ranked).

Table 5.14. Average ranks and correct ranks for collocations over- or under-rated ± 1 rank. (Rank 8 is the highest).

Over-ranked items			**Under-ranked items**		
Correct rank	Collocation	Mean Average rank	Correct rank	Collocation	Mean Average rank
4	Bad injury	5.21	5	Bad terms	3.96
2	Personal initials	4.03	6	Personal letters	4.97
2	Bad sport	3.49	3	Personal corruption	2
3	Personal impression	4.33	8	Bad news	7.01
1	Personal setback	3.82	8	Personal computer	6.04
			5	Personal liberty	2.05
			5	Bad form	3.08

3.6.1. Evidence of embedding effects

Is there any evidence that any of the collocations were under-ranked because of their being 'embedded' in a longer chain of language? Just as the letter 'k' might be more difficult to notice in third, rather than first position of words when generating exemplars, is it possible that this principle of availability might also operate in estimating the frequency of an adjective-noun collocation typically embedded in a larger chain? In such cases, we would anticipate an average lower ranking from subjects than is found in the BNC/Altavista data.

Bad terms was under-ranked. It is always preceded by *on* in its 15 instances in the BNC (though sometimes separated from *on* by another word e.g. *continuously, particularly*), and usually with *with* coming after the collocation: there are 11 instances of *on ...bad terms with* in the BNC, i.e. the 'bare' collocation *bad terms* occurs in this chain 73% of the time of its occurrences in the BNC. We might hypothesize then, that respondants failed to generate enough exemplars of this collocation as they did not recognize that the collocation *bad terms* typically occurs in the longer chain. *On bad terms with NP* may be stored and accessed as a unit, and as a result the subcomponents may not be so analysed or so accessible. Support for this is that the meaning of *terms* in the combination is not a typical use, and the only noticeable variation in the phrase is the substitution of *bad* for *good*, indeed the latter is more common according to the BNC[10]. The fact that the chain does not *begin* with the first word of the bare collocation (*bad*) may also make the adjective-noun collocation more 'hidden'. This finding is consistent with search restrictions resulting from the use of an availability heuristic in the estimation of the frequency of the collocation as discussed in chapter 4, section 3.2. Wray's hypothesis, in particular, would be able to account for the under-ranking as the chain *on bad terms with* is not very frequent and frequency would seem to be a necessary requirement for the Bybee fusion theory to be supported. There are no other occasions where availability appears to be a plausible explanation for any other of the other 6 under-ranked collocations.

3.6.2. Evidence of saliency effects

Were any 'salient' collocations over-ranked? The third hypothesis is that this would be the case. What makes a particular combination more salient than another? It is possible that encountering the collocation *bad injury* may have stimulated memories of a particularly bad injury for the respondant and that as a result this collocation was over-ranked. Similarly, it may seem, because of their consequences perhaps, that *personal setbacks* appear more frequent than they are. Are *injuries* prototypically *bad* and are *setbacks* prototypically *personal*? More work needs to be conducted into identifying what

[10] There are 54 instances of *on ....good terms* with in the BNC.

is, or is not salient for respondants, and this matter is discussed further in chapter 7. Importantly, it does not seem that collocations containing concrete nouns were generally over-ranked. For example, *personal impression* was over-ranked and *personal letters* was under-ranked.

3.6.3. Miscellaneous explanations

Various possible explanations can be forwarded to explain the poor ranks, and these will be noted briefly below. However, it would be unwise to attempt to try and unravel all the complexities of what might be going on in people's minds in doing these tasks – particularly because of the effect that one 'poor' ranking can have on another item to be ranked. This type of task is not particularly conducive to us probing into what is going on in people's minds: a productive task would give us a much better idea of this, as discussed in chapter 7.

The collocation which was the most under-ranked was *personal liberty*, and, partly as a consequence of this, the *personal* restricted set (set 3) of collocations was poorly ranked overall. It seems quite possible that the reason for this under-ranking was a US/UK national difference in the frequency of the word *liberty*. Hofland & Johansson (1982) indicate that there is a statistically significant difference in the frequency of the word *liberty* in LOB and Brown. While there are 16 instances in the British LOB, there are 46 in the American Brown, and this difference is statistically significant according to the Chi-square test at $p<0.001$. Though of course the BNC is a British corpus, Altavista searches material from around the world, and there is a good chance that it accessed many thousands of American web pages in the search. It would be interesting to see if a large American cohort would have ranked the collocation *personal liberty* higher than the British subjects[11].

It seems strange that *personal computer* was not recognized as the most frequent collocation in the *personal* restricted set (set 3). There are perhaps two reasons why it

[11] It should be noted, however, that the BNC's 25 instances are well spread in 21 documents and so the ranking does not seem skewed because of distribution problems.

fared so badly: firstly, it may be that '*PC*' is more common in daily usage, and secondly it may be that people just use the word *computer* without the *personal*, which is somewhat redundant in daily usage among non-IT specialists. One collocation that may be suffering from an age change is *personal letters*. With the increased use of email, SMS etc. personal letter writing is, no doubt, becoming rarer: this might explain why this collocation was under-ranked. One of the collocations under-ranked, *bad form*, has various meanings. It can mean, for example, that one's behaviour is not ethical: '*From one's earliest years one is taught that the showing of emotion publicly is bad form*'[12], or it can mean that someone's performance (not ethical behaviour) is not what it should be: '*Buckner's bad form continues*'[13]. It may be that the respondants considered only one of these meanings and, as a result, under-ranked the collocation.

A key interest is whether the raw frequency of the nouns had an effect on the rankings. It does not appear so. For only one of the poorly ranked collocations above, is this a possible explanation, and that is the over-ranked collocation *bad sport*. In that set (*bad* restricted) *sport* as a noun, when compared to the frequencies of the other nouns in the set, and as attested by the BNC, had a much higher rank compared to the relative frequency of the collocation *bad sport*, i.e. *bad sport* is not a very frequent collocation, but *sport* (as a noun) is frequent. However, for all the other combinations, there is either no evidence for collocations with frequent nouns in them being over-ranked, or else there is evidence to the contrary. For example, although *luc*k is four times less frequent than *news* in the BNC, and nearly 20 times less frequent in Altavista, it was not the case that *bad news* was the overwhelmingly more popular choice for the highest rank in the *bad* restricted set – the respondants were quite evenly split on whether *bad news* or *bad luck* was the most frequent collocation in this set. Other evidence against the possibility that respondants ranked according to the frequency of the noun is the under-ranking of *bad terms* and *bad form* where *terms* and *form* are very frequent nouns in their respective sets. Conversely, there were high rankings of *personal setback* and *personal initials*, but *setback* and *initials* are the most *in*frequent nouns in their respective sets.

[12] Text EA8 BNC
[13] This sentence is from the Internet. There are no BNC references in which *bad form* has this type of meaning.

4. Summary

Respondants can, generally, rank collocations quite accurately according to their frequencies in English according to BNC and Altavista search engine data according to the results from this experiment. Contrary to our first hypothesis, the 'restricted sets' were not as well ranked as the 'free sets'. The results from this test do not support the idea that restricted collocations as a class have 'privileged' representation in the mind, and, as noted above, it may well be that factors other than restrictedness affected the results. There is marginal evidence supporting the hypothesis that the use of an availability heuristic in ranking the items may have resulted in the under-ranking of the collocation *bad terms*. The effect of saliency is purely conjectural, and more research must be conducted into this factor before more definitive statements about its role in affecting frequency judgements can be made. If, rather than being asked to rank items, respondants are actually asked to produce high frequency collocates, this may help us to investigate more confidently the effects of frequency, saliency, hiddenness and availability on the subjects' judgements.

Chapter 6 - Insights from Word Association Studies

1. Introduction

In this chapter we move our focus of attention from recognition/ranking abilities to a review of the relevant literature about *productive* knowledge of collocates of words. As noted in chapter 1, section 3.1, corpus linguists have produced little evidence to substantiate their claims about the poor knowledge of the typical/frequent collocates of words. One source of relevant data that does exist is word association data, but I am not aware of any reference being made to these data by corpus linguists. This is somewhat surprising as it has been observed by researchers using word association tests that collocation type responses are fairly common (second only to co-ordinates according to Jenkins, cited in Schmitt 1998a, p.28). Even though respondants in these tests are not being asked to produce typical collocates, the fact remains that collocates are sometimes produced as responses and arguably the production of such items is indicative of strong connections between the words. Reviewing this research should enable us to investigate the claims made by corpus linguists (as noted in chapter 1), and probe further the representation issues discussed in chapters 3 and 4 (concerning inter alia, saliency, frequency, availability effects on the responses). This chapter is split into two main sections. The first is a review of the relevant word association research, covering methodological issues and an analysis of response types. Although most of the discussion is concerned with native speaker responses, some comments are also made about non-native speakers. The reason for this wider focus, (not directly relevant to the claims of corpus linguists in chapter 1), is that, as noted in chapter 3, section 4.3, Wray (2002) proposes that the mental lexicon of the native speaker and non-native speaker differs with regards to the representation of collocations and multi-word phrases. As this knowledge will be tested later (see chapters 7 and 8), theoretical issues are addressed here. The second section of this chapter is a detailed analysis of adjective-noun word association data from Moss & Older (1996). I attempt to categorise the responses, note the role of factors such as frequency and restrictedness, and make connections between this data and the relevant theories described in chapters 3 and 4.

2. Review of word association testing

2.1. Background

Word association testing has a long history and Aitchison (1994, pp.23, 24) traces it back over a hundred years. Researchers have been attracted to association tests because they believe that they indicate something about how the mind works and our thinking processes (e.g. Cramer 1968, p.2). There are a variety of word association tests, the most well-known and common being the free association test. These tests came to prominence in the work of Freud and psychoanalysis (according to Rozin et al. 2002, p.421), and the attraction of such a methodology was its unconventionality in probing the innermost mind in an 'unmonitored' way. Word association tests are still alive and well in psychology; for example, responses to the word *food* have recently been discussed with reference to nationality, gender and age (Rozin et al. 2002).

In addition to psychologists, both linguists and psycholinguists have shown an interest in word association tests and the resultant data. Rather than viewing word association responses as idiosyncratic and particular to the individual, the data is considered indicative of connections between words in the mind, which show a degree of commonality between different respondants. Groot (1989) terms word association responses, "relatively pure indicators of the way human knowledge is mentally represented" (p.824). There are many word association norm lists available: in the UK, the most well known are probably the Edinburgh Associative Thesaurus (EAT)[1], and the Birkbeck word association norms (Moss & Older 1996). In the United States, research by Nelson et al. (1998) has resulted in the massive University of South Florida (USF) word association, rhyme and word fragment norms being publicly available on the web[2]. This work is considerably larger than the earlier California Norms (CN) of Postman (1970). These various data sources enable us to see the different responses to a stimulus word and, *inter alia*, enable us to note the 'primary response' – the most stereotypical. Below,

[1] http://www.eat.rl.ac.uk.
[2] USF Free association norms are available at http://w3.usf.edu

in Table 6.1 an example is provided: the free responses to the stimulus word *trouble* are noted from these different data sources.

Table 6.1. A comparison of word association data from Moss & Older, EAT, USF and CN norms. Number of respondants and the number of different responses also noted (NDR=No. of different responses).

	Moss & Older N= 41-50 NDR: 29	EAT N=97 NDR: 62	USF N=146 NDR: -	CN N=1000, NDR: 290
Trouble	Maker (15.6%) Double (6.7%) In (6.7%) * (4.4%) Northern (4.4%) Strife (4.4%) With (4.4%) Problem (3.3%)	Double (8%) Maker (6%) Strife (6%) Shooters (4%) Worry (4%) Anger (3%) Bad (3%) Shooter (3%) Bother (2%) Bubble (2%) Mind (2%)	Bad (12.3 %) Problem (8.9%) Shooting (4.8%) Help (4.1%) Police (4.1%) Danger (2.7%) Fight (2.7%) Big (2.1%) Double (2.1%) Easy (2.1%)	Problem/s (13.2%) Bad (11%) Worry/ies (5.7%) Help (3.5%) Police (2.2%) Pain (2.1%) Fight (1.8%) Difficulty/ies (1.7%) Danger (1.5%) Wrong (1.2%)

* This symbol indicates that the responses provided were not clear in the data collected by Moss & Older.

As can be seen, some of the responses are synonyms e.g. *problem*, others seem to be a left syntagm response, e.g. *double*[3], and others a right syntagm response, e.g. *maker, shooters, shooting*. Some combinations are adjective-noun combinations, e.g. *double trouble*, others noun-noun, e.g. *trouble shooting*, and still others preposition-noun e.g. *in*

[3] This may even be a phonological association.

trouble. While there are differences between the sets, and some references are made to these differences in section 2.4 below, these initial observations suggest that a careful investigation of existing word association norms may help us investigate which collocating connections from a word are more available, and/or more salient to respondants in these kinds of tasks.

If we wish to use word associations to track collocations, there is a need to be aware of the findings of previous research, and the impact of different methodological approaches in the actual testing. The following section discusses the methodological issues.

2.2. Methodological issues

As mentioned above there are a number of different types of word association test. They may differ, for example, in the way in which the stimulus word is presented and how the response is recorded (e.g. aural/oral/written). Kruse et al. (1987, p.143) note two additional key variations. The first is when there are restrictions on the type of response allowed. When there are, these have been variously termed 'controlled' (Cramer 1968, p.24), 'restricted' (Riegel & Zivian 1972) or 'bound' (Groot 1989, p.824). The second difference concerns how many responses are allowed (single or multiple responses). These two issues are discussed in more detail below.

2.2.1. Controlled association and free association

Free association tests encourage the subject to report the first word(s) that come(s) to mind when a word or stimulus is presented to him/her. Controlled tests, on the other hand, restrict (in various ways) the type of response. Comparatively little research has been conducted using this latter type of test. Meara (1980, p.238) gives examples of restrictions such as the requirement to provide a co-ordinate of the stimulus word, or a word describing the stimulus. An example of a controlled association test is that of Riegel & Zivian (1972) who required their subjects to give a large number of 'restricted' responses, including superordinates, synonyms, preceding words etc. to the stimulus

words they provided. Cramer (1968, p.24) speaks in more general terms of semantic restrictions and conceptual restrictions on the type of response allowed and includes within the 'controlled' category of testing cases where the tester actually provides various alternative responses – a kind of multiple-choice test. The word associates test of Read (1993, p.359) in which non-native speakers of English have to identify related words to the stimulus words from paradigmatic, syntagmatic, analytic and non-associates would be classified as a controlled association test according to the criteria of Cramer.

2.2.2. Single and multiple responses

A single response word association task, as the name suggests, allows only one response to the stimulus; multiple responses allow a number of responses to the same stimulus word. Both methods have been widely used, and Groot (1989, p.824) argues that the multiple test is appropriate when examining the *number* of associates of a stimulus word, and the single response method is appropriate when investigating *association strength.* The possibility of 'chaining' responses, i.e., moving away from the initial stimulus and providing associates to associates (rather than associates to the initial stimulus word) has been noted by some researchers, e.g. Nelson & Schreiber (1992, p.240); Nelson & McEvoy (2000, p.510); Nelson et al. (2000, p.891), and is a reason forwarded for the unreliability of the multiple response test. A good example of a chaining response would be the (hypothetical) chain '*cat-dog-bone*' (from Nelson & McEvoy 2000, p.510). Kruse et al. (1987, p.147) attempted to limit this possibility by removing responses from the computer screen after they had been typed in. However, such an approach seems a little simplistic, for of course, there is no guarantee that the trace has been removed from the mind or its influence diminished.

Some researchers though, have been critical of the single word response methodology. For example, Schmitt (1998b) argues that allowing only a single response may lead to idiosyncratic responses being produced. In his research he favoured giving respondants the chance to provide multiple responses, arguing that this is fairer (1998b, p.391). Like Groot, Schmitt believes that multiple responses give a better idea about the network of

associations of a word (ibid, p.391). Research by Nelson et al. (2000, p.891) has recently confirmed the view of Groot noted above, that the first response in a free association task indicates stronger associations than subsequent responses. They note that in some of their own research, which allowed multiple responses, "when the primary associate was not produced on the first opportunity, it tended not to be produced on the second" (2000, p.887). (It should be remembered that the primary response is not provided by all respondants, it is simply the most common response.)

A final methodological issue to note is the time allowed between the presentation of the stimulus word and the response. Clark (1970) argues that the speed of response in word association tasks is critical: if a person is given a long time, he argues that the response is more likely to be idiosyncratic. If, on the other hand, the response is given quickly, he argues that the association is more obviously connected to the stimulus and more indicative of the actual associations in the mind (1970, pp.272, 273).

2.3. Types of stimuli and response classification

2.3.1. Stimulus type and its effect on response type

As Church & Hanks (1989, p.78) note, most word association research has used noun stimuli; however, there is some research on other word classes too (see below). Aitchison (1994, p.102) notes that word class is typically retained in word association responses, noting that nouns, in particular, tend to elicit noun responses (80%), with adjectives and verbs showing a less exclusive tendency (50%) to elicit their own word class partners. When a particular word elicits another word from the same word class, these responses are typically called 'paradigmatic' responses[4] and they are seen to be the norm in word association tasks with adult native speakers.

[4] This is only so if defined exclusively with regards to their word class, as noted below.

In addition to word class having an effect on the response type, the role of the frequency of the stimulus word has also been shown, at times, to be significant in affecting both the type of word typically elicited, and also the heterogeneity of the responses.

Regarding the first of these points, Cramer (1968), on reviewing the relevant research up to 1968, notes that the tendency to produce syntagmatic responses is inversely related to the frequency of adjective stimulus words[5] (1968, p.63 presumably based on Deese 1962, p.81). More recent experimental evidence for this comes from Söderman (1993). She used 64 stimulus words in a free word association task, 60 of which were adjectives. Half of these were frequent, and the other half infrequent. She found that there was a reduction in the number of paradigmatic responses for the infrequent set compared to the frequent set of words, both for native speakers and non-native speakers.

Turning now to the second point noted above (the number of different responses that a stimulus word produces, not from the same respondant but from the cohort of respondants), Postman (1970, p.241) noted that in discrete association tests for noun stimuli, higher frequency words tended to have more homogeneity in response type. This difference, however, was not found with a study of concrete nouns in a study by Hirsh & Tree (2001, p.6) and is not supported by Nelson & McEvoy's research (2000, p.517). These different results suggest that it is unwise to make broad judgements on this matter as frequency is only one of several factors that might have an effect on the response type.

As noted previously, it has been argued that an important variable affecting how useful responses are in indicating actual mental connections between words, is how much time is given to respondants. It has also been found that the timing allowed has an effect on the word class of the response. Cramer has noted that adjectives as a class tend to elicit adjectives as responses more often under time constraints (Cramer 1968, p.68). In Table 6.2 below, some key findings from Cramer concerning the effects of different word class stimuli on responses are noted. These cover: response commonality (i.e. how similar the

[5] She also notes that high frequency adjectives tend to elicit antonym responses (e.g. *good-bad*), lower frequency adjectives produce synonym responses, and unfamiliar adjectives produce syntagms (1968, p.69).

responses are to associations from the other subjects); the tendency for grammatical part of speech to be retained (i.e. paradigmatic responses to be more typical); and (in the third column) details about which word forms are more likely to elicit syntagmatic responses[6].

Table 6.2. The role of word class in affecting response types in free association tests (adapted from Cramer 1968, p.7)

Response commonality	Most typically eliciting paradigmatic responses	Most typically eliciting syntagmatic responses
Adjectives > nouns >verbs	Count nouns > adjectives > intransitive Verbs > adverbs > mass nouns	Adverbs > adjectives > verbs > nouns

Several other factors believed to affect the type of response provided in a word association test have also been noted. The first of these is that frequent stimulus words tend to generate frequent words as associations. Johnson (1956) explains this by suggesting that, "the availability of a word for free association depends in part on its general familiarity or response-strength regardless of any associations with specific stimulus-words" (p.126). Nelson & McEvoy (2000) also argue that frequent words tend to be produced in word association tasks and believe that this is due to the fact that frequent words have the advantage of being repeated often, and are therefore more recent in the minds of respondants. As a consequence, the frequent words have a greater chance of accessibility and retrievability (2000, p.509). Finally, on the issue of types of responses, Nelson & Schreiber (1992) note research suggesting that concrete words tend to be given as associates to concrete stimuli and abstract words to abstract stimuli (1992, pp.248, 249).

[6] For more on this classification see section 2.3.2

2.3.2. Response types: paradigmatic, syntagmatic, clang

One way of classifying the kind of associates that a particular word has, and a taxonomy widely used in the word association literature, is the paradigmatic, syntagmatic and clang[7] distinction[8]. Although these terms have been widely used, there are definitional problems – particularly with regards to paradigms and syntagms[9]. Deese (1962), who is typically regarded as the founding father of linguistic interest in word association testing, defined the paradigm/syntagm differences as follows: "Paradigmatic associates are words which can occupy the same position in an utterance as the stimulus (generally they are members of the same word class). Syntagmatic associates are words which occupy other, generally contiguous positions in an utterance; they are members of different form classes" (1962, p.79). This definition is rather vague on the status of same word class responses which co-occur with the stimulus word. For example, Wolter (2001) notes that the response '*salt*' to '*pepper*', could be classified as paradigmatic or syntagmatic depending on the definition of these terms (2001, p.50). Stubbs (2002b, p.226) has noted a large variety of 'paradigmatic type' associations which occur in phrases (i.e. occur syntagmatically). He gives examples of: antonyms (e.g. *alive/dead*), co-hyponyms (e.g. *bowls/plates*), hyponym and superordinate (e.g. *buses/transport*), terms for member and group (e.g. *aunt/family*), and approximate synonyms (e.g. *ashamed/embarrassed*). Moss & Older (1996) have suggested that co-occurrence patterns may actually explain why it is that certain paradigmatic type responses are provided in word association tasks. They note, for example, that *goat* and *cow* (both farm animals) are not strongly associated in word association tasks; on the other hand, *cats* and *dogs* (both pets) are, and they suggest that part of the reason why this is so, is that they are often mentioned together and also occur in certain phrases (e.g. *raining cats and dogs*) (1996, p.2).

---

[7] 'Clang' here is referred to nonsense responses, phonologically or orthographically induced but not semantically related to the stimulus word.

[8] It should, however, be noted that Read (2000) and Greidanus & Nienhuis (2001) add 'analytic'. Read (2000) defines this as, "one aspect, or component, of the target word and is likely to form part of its dictionary meaning" (p.181). Some of the examples he forwards are *electron-tiny*, and *export-overseas*. Greidanus & Nienhuis adopt Read's classification.

[9] Meara (1980) suggests that the classificational problems are so serious as to be "unworkable in practice" (p.239).

A study which clearly illustrates the consequences of defining responses without regard to word class is that of Hirsh & Tree (2001), who conducted a (noun stimulus) word association test with two groups of subjects (young adults and older adults). They required their respondants to write down their first responses to 90, mainly concrete, nouns. Rather than following the 'word class' approach to the categorization of response types into paradigmatic and syntagmatic responses, they adopted Bandera et al's (1991) classification which formulates a 'hierarchical-categorical' and 'propositional-relational' distinction of responses types. The former admits same word class responses (e.g. co-ordinates, subordinates, metonyms etc.), i.e. it is similar to the typical paradigm classification; the latter admits words from different word classes but also, importantly, includes responses from the same word class as the stimulus word which are, "generally... associated with the stimulus word in a particular context or in a common phrase or expression" (Bandera et al. 1991, p.293). What is important to note here is that clearly contiguous same class form words are denied 'paradigmatic' status. In analyzing their data, Hirsh & Tree note that the number of 'propositional-relational' responses (a class which is similar to syntagmatic, with the above noted difference) in their study was 70% for the younger group and 60% for the older group. This is very different from the typically understood tendency for adult native speakers to produce paradigm responses to nouns in word association tasks[10]. They explain this difference in the following way:

> The reversal of the typical predominance of paradigmatic-type over syntagmatic-type responses is due to classification without regard to form class. When form class is the decisive factor for classification, the dominance of paradigmatic responses re-emerges: 14% syntagmatic responses for the older cohort and 13% for the young cohort. (2001, pp.6, 7)

A considerable number of the propositional-relational responses were deemed by Hirsh & Tree to be 'phrasal collocations'. Those which were common dominant responses for both groups of respondants were: *cloak-dagger, daisy-chain, drain-pipe, gas-fire, peanut-butter, sly-fox, tummy-ache.* These responses are discussed in more detail later on in this section.

[10] See the comments in section 2.3.1 above and Table 6.2.

Research such as the above suggests that the importance of syntagmatic connections in affecting free association responses may have been seriously underestimated in previous work. While a considerable amount of analysis has gone into examining paradigmatic responses, and classifying them, comparatively little has been said about the syntagmatic responses produced in word association tests (which is our interest because of the adjective-noun collocation focus). Another reason why there are gaps in this field is that syntagmatic responses are often viewed as a developmental step on the way to more 'mature' paradigmatic responses (see section 2.4.3 on the typical argument that there is a developmental progression from clang to syntagm to paradigm responses in native speakers).

Clark (1970) is one of the few writers who has tried to make sense of the different syntagmatic responses in word association tests and he does so from a generative grammar perspective. He argues that two rules deal with the bulk of the syntagmatic responses. The first is what he terms 'the selectional feature realization rule'[11]. For example, the word *young* has selectional restrictions, i.e., it is used to describe animate things that are not adult. This being so, Clark argues that syntagmatic responses to this word in free word association tests (e.g. *boy, child* etc.), simply 'realize' the above noted features. Of course, many adjectives (particularly frequent ones) do not have such narrow selectional restrictions as the *young* example provided above[12]. They occur in attributive position before a wide range of nouns and so how important this 'rule' is, in helping us to analyse syntagmatic associations is not very clear. Clark also argues (following Chomsky) that nouns do not have selectional features, and that this explains why they do not typically elicit syntagmatic responses in word association. As noted above, the idea that nouns do not elicit syntagmatic responses (defined without regard to word class) is actually questionable. However, in the Hirsh & Tree data at least, it should be noted that the syntagm often bears no semantic relationship to the noun, but is, rather, a phrasal partner.

[11] This rule was alluded to in chapter 3, section 2.4.
[12] As noted in chapter 3, section 3: "very common adjectives typically designate a range of meanings" (Biber et al.1999, p.509).

The second rule that Clark forwards to explain syntagmatic responses is 'The idiom-completion rule'. Clark explains how this rule works in the following way: "Find an idiom of which the stimulus is a part and produce the next main word" (1970, p.282). It is unclear what exactly Clark means by the term 'idiom' in the quote above as the examples he forwards (i.e. *cottage cheese, white house, so what, ham eggs, stove pipe, justice peace, how now, whistle stop*) are quite a hotchpotch of combinations (including idioms, restricted collocations, sayings etc.). This theory though, together with the combinations that Clark provides and the Hirsh & Tree 'phrasal collocations' data noted earlier appear 'problematic' in the light of our discussion of Wray's formulaic language theory in chapter 3, section 4.3.

In chapter 3 it was noted how Wray argues that segmentation of holistically stored material is facilitated when there is paradigmatic variation in the formula. Segmentation makes the constituent words of the formula more accessible to our intuitions. A word is 'hidden', on the other hand, if it occurs in a longer string from which it is not easily loosened by paradigmatic variation. For instance, *large* in *by and large* is not in paradigmatic variation with anything else *(*by and small; *by and huge*) and so is likely to be hidden from view when a native speaker interrogates his/her intuition about the frequency, meaning and normal collocates of *large.* The problem is that this theory and the Clark idiom completion rule theory seem to clash. Clark seems to suggest that words in restricted collocations and idioms are quite available and accessible to us, in providing syntagmatic responses to word stimuli. So, how does Wray's theory fit in with this view? At face value it does not seem to very well: the dominant responses from the Hirsh & Tree data, the production of collocates in frozen collocations (Greenbaum 1988), and the Gilquin data noted in chapter 3, section 4.4 suggest that the individual words inside some strong collocations and idioms (in which there is strong invariability) are more available than the Wray theory would suggest.

How can these responses be interpreted? Firstly, it does seem that frequency of co-occurrence is an important factor affecting the responses. Five of the seven dominant collocate partners produced from the Hirsh & Tree phrasal collocation data are the most

frequent noun collocates for the stimulus words in the BNC[13]. But what if there were a choice for the respondants? What if *cloak* had two frequent noun collocates, one of which was an idiom collocate and the other a non-idiom? If the denotational meaning of the word drives the association, one would assume that the non-idiom transparent partner would be produced, because in the resulting combination the stimulus word has its dictionary meaning. However, if there are no non-idiom collocates, then material that we might consider to be formulaic, or stored holistically may be more readily accessed. An example supporting this view is the primary noun response to *hold* in Moss & Older, EAT and USF. The primary response is *hand*/s: it is not '*horses*'[14] (indeed, there are no instances of *horses* as an association in any of the databases)[15]. Further, '*hand/s*' is the most frequent noun collocate of *hold* (in a ±5 window search) in the BNC. It is possible that because there is a common transparent associate in usage, the idiom (*hold your horses*) is not searched.

However, for some words, there may not be a frequent transparent collocate which a respondant can provide. In the case of '*cloak-dagger*' for example, *cloak* has no other typical collocates. A similar explanation can be forwarded for the *take* data from Gilquin (2005a and b), described in chapter 3, section 4.4. There it was noted that in the data provided by her respondants, a number of sentences were produced in which the stimulus words were used idiomatically. Indeed, it was found to be the case that the idiom uses were actually more frequent in the respondant data than in the corpus data. A ±5 collocation window search of *take* in the BNC reveals that some of the most frequent collocates are: *place*, *off* and *over*, i.e. the resulting combinations are idiomatic phrasal verbs.

---

[13] Excluding the names *Yates* and *Chainsaw* from the *daisy* data. For the cases of *tummy* and *sly* it is not the case that *ache* or *fox* are the most common responses respectively. While there are a number of *tummy ache* instances in the corpus, more common examples are *tummy muscles, tummy toning* etc. However, these examples are not as well distributed as the *tummy ache* examples, and so the response *ache* to *tummy* does concur with *modified* corpus data on collocation frequency. For *sly*, it is not the case that *fox* is a particularly common collocate – more typical are *smile*, *look* and *grin*.

[14] The percentage of respondants who provided *hand* + *hands* is: in EAT is 9%, Moss & Older 23.3% and USF 4.7%.

[15] This example would be better if *horses* were a more common collocate of *hold*, but the example still illustrates the point.

It is submitted that the provision of associates which are idiom partner words would only be problematic for Wray's theory if there were frequent transparent options available to the respondants and these were overlooked. The partners not occurring in idioms, it would be supposed, are more likely to be analysed and therefore be more accessible. It is possible then, that the availability of choices may be a crucial factor in affecting the responses. If there is a choice of frequent responses (between idiom and non-idiom), we posit that the choice will be the non-idiom choice; however, if there are no such choices then in this kind of task the idiom partner may be produced. The theoretical justification for this is that the stimulus word, when occurring in the idiom, will not have its prototypical meaning, and, as such, should be less obviously related to its denotational meaning - the meaning which we suppose is uppermost in respondants' minds when engaging in word association tests.

As suggested in chapter 3, section 4.1, the individual words in idioms may be more or less analysed. Word association data suggests that some words in idioms seem to be more readily associated with stimulus words than other words in idioms. A good example of this is the frequent idiom response *milk* to *spill* in the USF, EAT and Moss & Older data. *Beans* is a less common response in the word association data[16] than *milk*, and this suggests (perhaps because liquids are more typically spilled) that *spill* in *don't cry over spilt milk*, is more prototypical and accessible than *spill* in *spill the beans*. In the same way, it may also be the case that some words in other types of holistically stored formulas are more accessible than others, though it is not clear at present which ones.

Clark's idiom completion rule is not very clear; however, it has raised an interesting issue - the availability of collocates within different types of lexical combinations. The idea that the ability to provide frequent collocates might be affected by the types of choices open to the respondants may be an important factor in helping us to understand and interpret word association data.

[16] It should be noted that there are around 10 times more *spill the beans* instances in the BNC than *cry over spilt milk* instances.

2.4. Native speaker and non-native speaker responses

2.4.1. Native speakers

Read (1993, p.358) and others (e.g. Rozin et al. 2002, p.421) have noted that native speakers show stable response patterns to word stimuli when considered *as a group*[17]. The validity of this view is borne out when comparing word associate responses from different time periods. In discussing word norm lists from American students from 1927, 1952 and 1960, Jenkins & Palermo (1965, pp.304, 305) have noted that the most common response (the primary response) was fairly stable over time. Though some writers have glossed over possible differences between individuals, treating native speakers as a homogenous group (e.g. Meara 1980, p.234), there is evidence that certain factors do play a role in influencing the responses. Firstly, age makes a difference – at least when dealing with concrete nouns. Hirsh & Tree (2001), found that younger adults produced a greater heterogeneity of responses than older adults and there was less agreement within this group than within the older group. Secondly, and as intimated above, (see Table 6.1), there is evidence to suggest that association norms differ between varieties of English. Nelson et al. (1998, p.6) comment that a comparison between a set of British norms and the Florida (US) ones revealed "substantial differences". They believe that this is due to regional/national differences in language use. They note, for example, that students in Florida respond to the word *apple* most typically with *red* and *orange*, as opposed to the UK norms of *tree* and *pie*. Two other factors influencing responses noted by Rozenzweig (1964) are occupation and education. Rozenzweig compared word association responses from French construction workers and French students, and found a considerable number of differences between the two groups. For example, the primary responses of the two groups were the same for only 39 of the 98 stimulus words. He also found that the workmen gave many more syntagmatic, rather than paradigmatic responses compared to the student group. His findings were that, "adult members of the same 'language

[17] As Wolter (2001, p.47) notes, even a native speaker's lexicon is unstable in the sense that new words are added, old ones lost etc.

community' may have verbal habits that differ systematically according to social groupings within the community" (1964, p.68).

2.4.2. Non-native speakers

Read (1993, p.358) notes that language learners, unlike native speakers, show more diverse response patterns to word association stimuli, which are often phonologically confused. Such a belief is based, in part, on Meara's (1984) study. In Meara's study of associations with English learners of French, he notes a considerable number of cases where either phonological or orthographic confusion resulted in rather bizarre responses. An example of this was the response *conducteur* (conductor) to the stimulus *béton* (concrete), which Meara (1984, p.233) believes can be explained by arguing that the respondant confused the stimulus word *béton* (concrete) with *bâton* (stick). Meara suggests that the L2 learner has a very different mental lexicon compared to the native speaker (1984, pp.233, 234), noting elsewhere (1980, p.238) that L2 learners provide less homogenous responses as a group compared to native speakers and that this appears odd as there is a smaller group of words in the L2 lexicon for the word to be associated with.

2.4.3. Developmental issues

The belief in a 'clang-syntagm-paradigm' developmental shift in native speakers is fairly widely held[18]. It has been noted by many writers (e.g. Carter 1987, p.158; Söderman 1993, p.157; Meara 1980, p.235) that young children often produce 'clang' responses in word association tests where the word produced has some kind of non-semantic relationship to the stimulus word (usually phonological in nature, e.g. the response *far* as an associate of *car*). As well as producing 'clang' responses children also provide syntagmatic responses. The 'final step' in the developmental shift is the tendency to provide paradigm responses, which some writers believe occurs around the age of 7 (e.g. Carter 1987, p.158; Söderman 1993, p.157). Clark (1970, p.285) believes that this change does not happen to all classes of words at the same time, noting research which suggests

[18] See Wolter (2001, p.43) who notes the relevant research supporting this view.

that nouns, then adjectives, then verbs, and then adverbs increasingly elicit paradigmatic responses. Söderman (1993) reports on research which suggests that the development may be one which happens to different words (rather than classes of words) at different times.

Wolter (2001) is one of the very few researchers who questions the clang-syntagm-paradigm developmental shift, and his view deserves consideration. He argues that earlier research upon which the developmental theory was built may have failed to appreciate that the increase in paradigmatic responses is not due to fewer syntagmatic responses, but rather to decreases in clang and nonsensical responses (2001, p.62). Indeed, he prefers to view the 'shift' as one from semantically meaningless responses to semantically meaningful ones (which can be either syntagmatic or paradigmatic). He also believes that the tendency for native speakers to produce more paradigm responses than non-native speakers can be explained by the relatively greater percentage of paradigms present in the native speaker lexicon from which associates can be drawn (2001, pp.64, 65).

What of L2 learners? Meara (1980, p.239) argues that there is no evidence for the clang – syntagm-paradigm development. However, a study conducted by Söderman challenges this view. Söderman (1993) compared the types of response of four different levels of Finnish learners of English (7th graders, 'gymnasium' pupils, 1st year university students and advanced students), to 100 fairly frequent words of different word classes. She found that there was a significant proficiency effect in the production of paradigms (i.e. the advanced learners produced more paradigmatic responses than the other groups of learners in a single response free association task)[19]. She did not find a difference between the advanced learners and native speakers in terms of the percentage of syntagmatic and paradigmatic responses in another test, investigating responses to frequent and infrequent adjectives. These findings of Söderman should be handled carefully: while there may be a proficiency effect in the production of paradigms, this does not mean that the same words or connections will be in the NSS lexicon as in the NS lexicon. It is important to remember that the problems of non-native (advanced) speakers

[19] It should be noted that she uses the word 'paradigm' for same word class response (1993, p.157).

with collocations are well documented (e.g. Bahns and Eldaw 1993; Farghal and Obeidat 1995; Nesselhauf 2003; Herbst 1996), and it would be wrong to assume that the *network* of syntagmatic relations is equally well developed in the NNS lexicon, even if NNSs appear 'native-like' in the tendency to produce paradigm responses[20]. Indeed, it may be the case that paradigmatic knowledge is developmentally prior to syntagmatic knowledge for many language learners in terms of its breadth and coverage[21]. Greidanus and Nienhuis (2001, p.574), for example, found that Dutch advanced learners of French were more able to *identify* paradigmatic and analytic associations for stimulus words than syntagmatic ones, which suggests syntagmatic deficiencies. Indeed, it is useful to refer back to the comments of Wray (2002) at this point (chapter 3, section 4.3) and her belief that post-literate second language learners, because of their approach to learning, are more likely to be unaware of exactly what type of combinations are more native-like or acceptable.

The findings from the above review are used to help design the research reported in chapters 7 and 8.

3. Adjective-noun collocations in word association tests

3.1. Introduction and research questions

Our particular interest in word association studies is adjective-noun collocations, and as noted earlier, in free association tasks adjectives do (sometimes) elicit noun associates (particularly so when the adjectives are infrequent); however, nouns rarely elicit adjective responses[22]. Previous word association research has not given much attention to the syntagmatic responses to word association stimuli. As noted above, this might be a result of the pre-occupation with paradigmatic responses, plus the consequence of viewing the

[20] It is important at this point to stress that single word association tests are not supposed to indicate anything about networks of associates – see section 2.2.2.
[21] I say 'many' here, as classroom learners, one would expect, would be more analytical in their approach to language learning than 'natural' learners.
[22] Though, as was noted in section 2.3.2, many of the noun–noun responses may actually be syntagmatic type responses.

syntagmatic responses as not fully developed responses. Can existing data be analysed for the light it can shed on the connections between words, in particular adjective-noun collocations? Fortunately, it can and such an analysis can help identify what types of collocate responses are typically produced. More specifically, an analysis of the existing data should enable us to answer the following questions:

1. Is it the case that *only* infrequent adjectives elicit dominant noun responses?[23]
2. Is the primary collocation partner provided by the respondants a frequent partner according to corpus data?
3. Are the resulting collocations *free* or *restricted*?
4. Are any of the resulting combinations idioms?
5. Is it possible that some frequent collocates are not produced when it is the case that the stimulus word is delexicalised or used in a non-typical sense when it combines with the frequent collocate?
6. Is there any evidence that certain partners are *not* produced because they are hidden in some way?
7. What other factors (in addition to frequency of co-occurrence) play a role in influencing the responses?

All of these questions, with the exception of the first, are related to the theoretical issues of representation and accessibility as discussed in earlier chapters.

3.2. Analysis of Moss & Older (1996) data

An analysis of existing data from word association tests is not as straightforward as it may seem. The most common response to an adjective - the primary response - may or may not be a noun in a free association test. This is a consequence of using the free association methodology, and as noted in section 2.2.1, the vast majority of word association testing has adopted this methodology, rather than the controlled testing model, which could specify that a particular response type be provided. For example, in

[23] As noted in section 2.3.1, there is a tendency for less frequent adjectives to elicit nouns, but how significant this tendency is should be investigated.

the Moss & Older (1996) free association data the primary response to the highly frequent adjective *bad* is *good* (with 40.5% of the responses). There are also noun responses (e.g. *boy, breath, girl, taste, weather*), but these are less well attested: only 7.1% of the respondants provided the word *boy* as an associate to the word *bad*, and this was the most common noun response. It would be questionable indeed to reanalyze existing word association data and make much of responses for which there is so little evidence[24]. However, when a larger percentage of respondants give the same collocating noun to an adjective stimulus, then we have greater warrant for positing that the association between the stimulus word and the collocate is a strong one among the respondants as a group. Focusing on these responses is justifiable and would enable us to note different categories of responses and provide us with enough data to answer the questions provided above. In what follows, the methodology adopted in analyzing some of the existing adjective stimuli data is described.

3.2.1 Adjective stimuli and responses

3.2.1.1 Analysing the data

The adjective stimulus data (446 words) from Moss & Older (1996), together with the responses were analyzed. When the same noun was produced by at least 20% of the respondants this was recorded. The reason for choosing this figure as a cut-off point was somewhat arbitrary, but justifiable in that it indicates that, for the group, this was a fairly common response indicative of strong connections between the stimulus word and the response. Of course, as a result of adopting this requirement interesting collocate responses were sometimes missed. This was particularly so when several *different* collocating nouns were produced by a number of respondants. For example, *sharp* elicited *knife* (17.8%) and *edge* (15.6%) and neither of these responses has enough attestation to be recorded according to the criteria adopted in this analysis. However, such

[24] The same issue arises if we wish to check other claims made in chapter 1, e.g. about the collocates of *cause, a torrent of, set in, regime* etc.

a scenario rarely occurred, and the 20% attestation requirement provided a reasonable amount of data to be examined.

The goal of conducting this analysis is to investigate and answer the 7 questions noted above. However, there were some issues that needed to be clarified before proceeding with the analysis. Firstly, the stimulus word adjectives which elicited a 20% noun response needed to be classified according to their frequencies in order to enable us to answer question 1. Unfortunately, the terms 'high' and 'low' frequency have been used in different ways and in this investigation it is desirable to carefully classify the stimulus words, for the help that such information can provide in designing further research. In what follows the relevant research on word frequency classification is reviewed, and the criteria for establishing different frequency ranges is described.

With regards to high frequency words, at one extreme Segalowitz & Lane (2000, p.380) term occurrences >883 per million as high (and over 10,000 as superhigh), and on the other extreme is Söderman (1993, p.164) who classifies instances >50 per million as high frequency. However, the majority of researchers (e.g. MacAndrew & Harley (2000), May & Tryk (1970, p.300), Rugg et al. (1995)) seem happy to use a figure of 100 instances per million words, or thereabouts, to classify a word as high frequency. A few writers define this class at a slightly higher level: for example, Griffin & Bock (1998) and Alario et al. (2002, p.305) define frequent words as those with between 150-200 occurrences per million words. The unusually high figures forwarded by Segalowitz & Lane are best understood in the light of their focus on both content and function (grammar) words in their research. Function words are very common in corpus data, and they alone have instances of over 10,000 words per million in the Kučera and Francis (1967) list. The same list also indicates that there are only around 1000 words with frequencies over 100 instances in the 1 million word Brown corpus. Turning now to definitions of low frequency, most writers give a figure of around 10 instances or under per million (e.g. MacAndrew & Harley (2000)<9; Söderman (1993)<10; May & Tryk (1970)<3; Rugg et al. (1995)<7; Alario et al. (2002) mean average 13). Griffin & Bock (1998) provide a mean average for low frequency words which is a little higher: 15 or 28 words per

million (on two different corpora). As with the high frequency categorisation, Segalowitz & Lane (2000) are also high (relatively speaking) when it comes to low frequency categorization (0-122 words per million).

Following on from the general observations above, the following classifications were made for the following reanalysis of word association data:

High frequency words: >100 / million = >10,000 instances in the BNC[25]
Low frequency words: <10 / million = <1000 instances in the BNC
Intermediate frequency words: 1000 to 10,000 instances in the BNC

With regards to question 2, Benson et al. (1986) was used to determine whether the resulting collocation was restricted or free. To determine whether the word association response was a frequent collocate, the stimulus words were searched in the BNC (in a ±5 window search) and the most frequent noun collocates were recorded to enable a comparison between the dominant response and the corpus data.

3.2.1.2 Results

Table 6.3 contains the results of the analysis. In the first column the adjectives which elicited a particular noun response from at least 20% of the respondants in the Moss & Older data are listed. The same column also includes (in brackets) the number of instances of the adjective in the BNC. In the second column, using the frequency criteria described above, the adjective is classified as (H)igh, (I)ntermediate or (L)ow frequency. The third column contains the primary response plus the percentage of respondants who provided it. Occasionally, there were two responses which were collocating nouns of the stimulus word with over 20% of the subjects providing each of them. If this was the case both responses are recorded. In the fourth column, the resulting combination from the Moss & Older data is classified as restricted or free according to its presence in Benson et al. (1986). If it is restricted this is recorded with the letter "R". In the fifth column BNC

[25] The frequency per million figures are multiplied by 100 as the BNC is a corpus of 100 million words.

raw frequency data for the most frequent noun collocates of the different adjective stimuli words are recorded. Occasionally, the output from the BNC was modified on the grounds that the data had very poor attestation from the sources in the BNC[26] – the instances where this occurred are noted under the table. The collocates are listed starting with the most frequent collocation. The number of collocates provided in the cell differs from word to word. For example, if the most frequent collocate in the BNC is the primary noun response in the Moss & Older data, then only the most frequent collocate from the BNC is noted. If, on the other hand, the primary response is the fifth most frequent collocate in the BNC, then the 5 most frequent collocates from the BNC are noted as this enables a comparison between the most frequent collocates in the BNC and the respondant data. However, when the primary noun response is not very frequent, I do not list all of the collocates in the BNC to enable a comparison, as this would make the table considerably longer than it is at present. In such cases, the BNC attestation for the dominant response is provided along with a few of the most frequent noun collocates according to the BNC. When this is so, the number of instances in the BNC is noted (in brackets after the collocate) but the collocate does not have a number in front of it.

[26] The theoretical justification for adopting this principle was discussed in chapter 2, section 2.3.

Table 6.3. Analysis of Moss & Older data (Adjective stimuli)

Stimulus word / BNC frequency (adj POS search)	Frequency of stimulus word (L,I,H)	Primary noun response / Percentage response	Benson et al. (R)	Most frequent collocating noun/s in BNC (±5 window search) with no. of instances of the collocation in brackets.
Blond (386)	L	Hair (29.2)	R	1.Hair (146) - Girl (10)
		Girl (20.8)	-	
Blue (7713)	I	Sky (24)	R	1. Eyes (959) 2. Sky (295)
Blunt (446)	L	Instrument (25.6)	R	1. Instrument (41) - Knife (7)
		Knife (23.3)	-	
Blatant (333)	L	Lie(s) (22)	R	1. Attempt (13) 2. Lie / Example (10) - Lies (3)
Brisk (474)	L	Walk (29.2)	R	1. Walk (39)
Bright (5161)	I	Light (21.5)	R	1. Eyes (349) 2. Red (318) 3. Light (261)
Broad (4773)	I	Bean(s) (25)	R	1. Range (184) - Bean/s (49)
Candid (149)	L	Camera (50)	R	1. Camera (19)

Stimulus word	Frequency	Noun associate	Restricted	BNC most frequent collocates
Casual (1724)	I	Clothes (21.4)	-	1. Workers (98) 2. Labour (58) 3. Basis (43) 4. Clothes (38)
Classical (3206)	I	Music (78.6)	R	1. Music (177)
Confidential (1101)	I	Secret (45.2)	-	1. Information (271) - Secrets (5) - Secret (7)
Cosmetic (322)	L	Surgery (21.4)	R	1. Surgery (55)
Curly (406)	L	Hair (49.5)	R	1. Hair (209)
Dark (5807)	I	Night (20)	R	1. Eyes (959) 2. Sky (295) 3. Night (149)
Drunken (599)	L	Stupor (38.1)	R	1. Driving (33) 2. Stupor (25)
Early (23626)	H	Morning (21.4)	R	1. Days (1113) 2. Century (1016) 3. Years (967) 4. Morning (803)
Elastic (251)	L	Band (54.8)	-	1. Band (31)
English (7389)	I	Language (28.6)	-	1. Law (409) 2. Literature (291) 3. Language (264)
Express (711)	L	Train (21.4)	R	1. Provision (50) 2. Train (36)
Fertile (598)	L	Soil (21.4)	R	1. Soil (56)
Formal (6331)	I	Dress (23)	R	1. Education (205) - Dress (21)

Stimulus word	Frequency	Noun associate	Restricted	BNC most frequent collocates
Fragrant (272)	L	Smell (28)	-	1. Flowers (19) - Smell (5) - Smells (2)
Front (3432)	I	Row (22.2)	-	1. Door (1669) - Row (25)
Gallant (252)	L	Knight (22)	-	1. Attempt (8) *Gallant Knight* not in BNC
Grand (4356)	I	Piano (29.2)	R	1. Slam (208) - Piano (47)
Green (6437)	I	Grass (35.4)	-	1. Party (287) - Grass (79)
Heavy (8985)	I	Weight (23.8)	-	1. Rain (294) - Weight (57)
Hollow (361)	L	Tree (24.4)	-	1. Ring (17) 2. Tree(s) (15) - Tree (7)
Humble (750)	L	Pie (32)	-	1. Beginnings (32) 2. Origins (28) 3. Pie (25)
Hypnotic (170)	L	Trance (23.8)	-	1. State (16) - Trance (2)
Initial (2289)	I	Letter (20)	-	1. Training (125) - Letter (5) - Letters (6)

Stimulus word	Frequency	Noun associate	Restricted	BNC most frequent collocates
Juvenile (346)	L	Delinquent (42.9)	R	1. Crime (49) 2. Delinquency (28) 3. Offenders (25) - Delinquents (12) - Delinquent (7)
Loud (1060)	I	Noise (28.9)	R	1.Voice (125) 2. Noise (68)
Manic (216)	L	Depressive (28.3)	*Depressive* not in Benson et al.	1. Depression (21) 2. Depressive (16) §
Merry (510)	L	Christmas (33.3)	R	1. Christmas (72)
Odd (4312)	L	Couple (20)	-	1. Way (80) - Couple (24)
Old (52275)	H	Man (20)	R	1. Man (2603) *
Outer (2374)	I	Space (22.2)	R	1. Hebrides (181) 2. Space (155)
Parallel (873)	L	Lines (40.5)	R	1. Systems (54) 2. Lines (33)
Polar (660)	L	Bear (58.3)	R	1. Regions (65) 2. Bear (48) 3. Bears (42)
Precious (1577)	I	Stone (27.1)	-	1.Stones (102) - Stone (10)
Prehistoric (374)	L	Dinosaur (32.6)	*Dinosaur* not in Benson et al.	1. Times (47) - Dinosaurs (4) - Dinosaur (0)

Stimulus word	Frequency	Noun associate	Restricted	BNC most frequent collocates
Raw (2383)	I	Meat (26.7)	R	1. Materials (638) 2. Material (296) 3. Data (88) 4. Meat (64)
Remote (2815)	I	Control (44)	R	1. Control (216)
Round (2347)	I	Ball (31)	-	1. Table (224) - Ball (12)
Senior (8059)	I	Citizen (26.2)	R	1. Management (496) - Citizens (158) - Citizen (27)
Serial (386)	L	Killer (46.7)	R	1. Killer (38)
Shallow (1363)	I	Water / waters (37.8)	-	1. Water (157) 2. Waters (51)
Silent (3493)	I	Night (34.9)	-	1. Moment (141) - Night (31)
Spare (1841)	I	Tyre (27.1)	R	1. Time (374) - Tyre (14)
Spicy (207)	L	Curry (20.4) Food (20.4)	- R	1. Food (18) - Curry (1)
Stormy (361)	L	Weather (45.2)	-	1. Night (29) 2. Weather (23)
Straight (2382)	I	Line (20.4)	R	1. Line (398)
Tame (264)	L	Lion (45.8)	-	1. Elephant / Animals (7) - Lion (1) * *
Tartan (260)	L	Kilt (26.2)	*Kilt* not in Benson et al.	1. Army (8) 2. Rug / Cap / Scarf (7) - Kilt (4) - Kilts (2)

Stimulus word	Frequency	Noun associate	Restricted	BNC most frequent collocates
Tepid (81)	L	Water (28.9)	R	1. Water (22)
Tidy (700)	L	Room (22.9)	-	1. Room (26)
Upper (5251)	I	Hand (23.8)	-	1.Class (248) - Hand (144)
Wavy (139)	L	Hair (31.2)	R	1. Hair (51)

§ *Street Preachers* excluded, as this band is post Moss & Older association data.

* *ref* (3624) excluded because of its presence in just 4 texts in the BNC, typically occurring in *Old Ref No*:

** *Valley* (13), *Canal* (9) excluded because of their presence in *Tame Valley Canal*

Before discussing these data, it should be noted that the inclusion of a few of the above responses in the list is rather subjective, and tabulating them as adjective-noun responses can be challenged. Though Moss & Older classify *initial* as an adjective stimulus, it is possible that the respondants treated it as a noun, in which case the response *letter* is a noun-noun response. A slightly different case is the response *secret* to *confidential*. Is this a synonym type response or is it a syntagmatic one? This cannot be known for sure, though I have listed is as an adjective-noun response.

3.2.1.3 Discussion

Moss & Older (1996) classify 446 stimulus words as adjectives. For 59 of these words – the words listed above in the first column (i.e. 13% of the adjective data) – at least one fifth of the respondants provided the same noun response. As a result these associates and the resulting collocations can be investigated for what they indicate about strong links between adjectives and nouns, on the grounds that they are less likely to be idiosyncratic, because of the degree of commonality among the group. In what follows, the 7 questions provided in section 3.1 are listed and answered in turn.

1. Is it the case that *only* infrequent adjectives elicit dominant noun responses?

Clearly, the vast majority of the stimulus words which had dominant noun responses are low frequency (31/59) or intermediate frequency words (26/59)[27]. Only 2 of the adjectives listed above are high frequency according to the definition of this term: *early* and *old*[28]. As a consequence of this, the data above fail to inform us whether high frequency adjectives do have strong noun associates. This needs to be investigated, as it is typically argued that frequent words are more typically delexicalised (Sinclair 1991a, p.101).

2. Is the primary collocation partner provided by the respondants a frequent partner according to corpus data?

For 37 of the 62 associates provided[29], the associate is found among the 5 most frequent noun collocates of the stimulus word according to raw frequency collocation data from the BNC. This means that the association data and BNC frequency data are roughly comparable in these cases[30]. For 16 of the 37 frequent responses, the dominant associate is the most frequent collocate according to the BNC. Though frequency of co-occurrence seems to be an important issue in affecting the responses, clearly, it is not the *only* issue affecting the responses, as discussed below under question 7.

3. Are the resulting collocations *free* or *restricted*?

Of the 62 associates provided, 34 of them are restricted according to Benson et al. (1986). It was noted in chapter 1, that Fox believes that frequent collocates of words will only be

[27] Note that 2 of these (*loud* and *confidential*) are close to being categorized as "L".
[28] It should be pointed out that some Moss & Older data indicate that infrequent adjectives did, at times, elicit primary dominant adjective responses (e.g. *exquisite-beautiful* (16.7%); *flagrant-obvious* (11.9%), *facetious-obnoxious* (7.1%), *inanimate-dead* (30.4%). The tendency for infrequent adjectives to elicit nouns should be viewed as a tendency – not an absolute.
[29] Note that for three of the adjectives 2 noun responses were provided which fulfilled the 20% attestation requirement.
[30] It should be remembered that the respondants are not being asked to produce frequent collocates: the word association data being analysed is from *free* association data.

provided when the stimulus word is the restricted word in severely restricted collocations. There is a connection between frequency and restrictedness in the responses: 28 of the 37 responses which are frequent collocates of the stimulus words make restricted collocations. The nine collocates not classified as restricted but which were frequent are: *casual clothes, elastic band, English language*[31], *hollow tree, humble pie, precious stone/s*[32], *shallow water/s, stormy weather, tidy room.* The six collocates which were classified as restricted but are not frequent are noted in Table 6.4 below. It would seem then, that there is a considerable overlap between restrictedness and frequency in line with Fox's view. However, it is important to remember the discussion on restricted collocations in chapter 3, section 2.3 at this point. It was noted that Fox's frozen restricted collocation was only one of four types of restricted collocation according to Mel'čuk. It should be noted that a number of the responses which were frequent and restricted are *not* frozen collocations, e.g. *loud noise, old man, blue sky, bright light, early morning, straight line.* This is an important finding. While it is the case that some of the collocations from the Moss & Older data are frozen collocations (e.g. *serial killer, drunken stupour, wavy hair*), it is not *only* these collocations that are produced by the respondants: some of the resulting frequent collocations are free and others are less restricted collocations.

4. Are any of the resulting combinations idioms?

Two of the dominant noun responses make the combination an idiom: *humble pie*, and *upper hand*[33]. The production of *pie* to *humble*, rather than the slightly more frequent *beginnings* and the production of *hand* to *upper*, rather than *class* (the most frequent collocate according to the BNC), seems to challenge the view forwarded earlier (in section 2.3.2), that idiom partners of stimulus words are less visible in syntagm searches. However, it may simply be that the adjective words in these idioms are more transparent than we might assume. Interestingly, there is a case when the most frequent collocate of

[31] *English language* can be viewed as a syntagm response, *or* an analytic response.
[32] It should be noted that it is the plural *stones* which is frequent in the BNC data.
[33] It is perhaps questionable to term *upper hand* an idiom. It is not listed in the Collins Cobuild Dictionary of Idioms (1995), yet the meaning of *upper hand* is not clear from its component parts.

one of the stimulus words results in the formation of an idiom, but the respondants did not provide the idiom partner – *hollow ring*. The dominant response to *hollow* from the respondants was *tree*. This suggests that *hollow* in *hollow ring* is not so directly connected to its lone word meaning: it is less 'prototypical' in this combination, and the meaning of the idiom is quite opaque. We can contrast this case with the dominant responses to *humble* and *upper*. A person is humbled if they have to eat humble pie, (i.e. there is a clear connection between the meaning of *humble* in the idiom and its effect on the person who 'eats' it) and *upper* does have its prototypical meaning of 'higher position' in *upper hand.*

5. Is it possible that some frequent collocates are not produced when it is the case that the stimulus word is delexicalised or used in a non-typical sense when it combines with the frequent collocate?

There is some support for the idea that where the meaning of the adjective in a very frequent collocation is untypical, that the frequent collocate is not produced, and that the response is affected by the denotational meaning of the word. For example, the dominant response to *senior* was *citizen.* The prototypical meaning of the word is connected to age – at least in the minds of the respondants, but in the most frequent collocation, (*senior management*) it means 'high position'. Similarly, the primary noun response to *casual* was *clothes*, in which *casual* has the meaning 'not formal', whereas when in combination with the most frequent collocate *labour*, it has the meaning 'employed for a short period of time'. In the same way, the respondants did not produce the most frequent collocate of *heavy* i.e. *rain.* The non-production of this item is also consistent with the view that the denotational meaning of a word affects ideas about its meaning and uses – *heavy* does not have its 'weighty' meaning in this collocation.

6. Is there any evidence that certain partners are not produced because they are 'hidden' in some way?

If we look at cases where the most frequent collocate was *not* produced there are some interesting tendencies. Several of the most frequent collocates not provided as dominant responses are typically 'embedded' in larger chunks of language when they collocate with the stimulus word, according to corpus data (e.g. ***a broad range** of NP, **in the early days**, **in an odd way**, **in / from prehistoric times**)*. The fact that the nouns in these chains were not dominant responses is consistent with an availability heuristic being employed in searches for collocates. It may be that the chains are holistically stored and the fact that the 'bare' collocation usually exists within the chain[34], and is often not the first word of the chain, results in searches overlooking the collocates in such frames.

7. What other factors (in addition to frequency of co-occurrence) play a role in influencing the responses?

As noted earlier, some of the dominant responses were not frequent collocates, and there are various reasons why this is so. In Table 6.4 the (infrequent) collocate responses (of which there are 25), are grouped into 8 different categories:

Category 1 – The adjective is *the* salient feature of the noun or the noun is the adjective stereotypically.
Category 2 – The adjective is the opposite of the stereotypical feature of the noun
Category 3 – The resulting noun is a compound noun, i.e. it has a specific meaning with the adjective and a separate dictionary entry.
Category 4 – Idiom
Category 5 – Restricted collocation. The adjective qualifies the noun from a small number of possibilities from the same semantic field (though it is not among the most frequent of those restricted collocations in the BNC).

[34] For example, there are 159 occurrences of *broad range* in the BNC, and 119 instances of *a broad range of NP*, i.e. in 75% of its occurrences *broad range* occurs within this chain.

Category 6 – The adjective is a 'polar' quality of the noun

Category 7 – Quotation

Category 8 – Miscellaneous (not easy to classify)

Table 6.4. Categorising infrequent noun collocates produced by respondants in the Moss & Older data (R = restricted according to Benson et al. 1986)

Category 1	Category 2	Category 3	Category 4
-Green grass -Spicy curry -Round ball -Confidential secret* -Tartan kilt -Gallant knight -Prehistoric dinosaur -Hypnotic trance	-Blunt knife (anti *sharp*) -Tame lion (anti *wild*) -Odd couple (anti *suited / normal*)	-Broad beans (R) -Spare tyre (R) -Precious stone -Senior citizen (R) -Grand piano (R) -Front row§	-Upper hand
Category 5	**Category 6**	**Category 7**	**Category 8**
-Juvenile delinquent (R) -Formal dress (R)	-Heavy weight¶ -Fragrant smell -Blond girl	-Silent night	-Initial letter*

* As noted earlier there are interpretation problems with these collocations, though I have included and categorized them here.

§ This can be technical in a rugby sense, though is not necessarily so.

¶ The association *weight* to *heavy* might be a response resulting in a compound noun, and would therefore be placed in category 3.

The above categorisation suggests that frequency of co-occurrence may not be the only factor affecting the collocation type responses, and this is hardly surprising: it should not

be forgotten that the respondants are not actually being asked to provide the most frequent collocate. In Categories 1, 2 and 6 in particular, it seems that a key salient semantic feature of the stimulus word is identified, and the prototypical noun having that adjective quality is produced – grass is green, balls are round, knives are sharp/blunt etc. As such then, these responses may not really be driven by a syntagmatic co-occurrence effect at all. It is also interesting to note the compound noun responses (category 3). In these cases the 'larger words' are 'completed' by the response, and there clearly is a syntagmatic drive in the response, but not one which is driven by frequency (at least according to the BNC data), or a salient quality of the adjective, rather it seems to be driven by the desire for completeness. Wray (2002, p.74) argues that compound nouns are stored as single lexical items with their meaning. Interestingly, though, it would seem that the components of these combinations are quite accessible according to the data provided here, and it should also be noted that many of the frequent collocations are also of this type e.g. *polar bear, express train.*

Having examined the noun response data in some detail, it makes sense to see whether any additional information can be gleaned by looking at the adjective responses to nouns in the Moss & Older data.

3.2.2. Noun stimuli and responses

3.2.2.1. Analysing the data

It was noted in the review preceding the analysis of the Moss & Older adjective-noun data, that nouns typically elicit noun associates in free word association[35]. However, it may be that some nouns elicit dominant adjective responses and this possibility should be investigated. A check of the noun data from Moss & Older was conducted, and as in the former analysis, when a particular stimulus elicited the same collocate response from at

[35] It was also noted that it is contested how these responses should be classified, i.e. as paradigmatic or syntagmatic.

least 20% of the respondants this was recorded. There are very few instances as Table 6.5, below, shows. The format of this table is the same as Table 6.4.

3.2.2.2. Results

Table 6.5. Analysis of Moss & Older data (noun stimuli)

Stimulus word / BNC frequency (**Noun** POS search)	Frequency of stimulus word (L,I,H)	Primary adjective response in Moss & Older data and %age response	Benson et al. (R)	Most frequent collocating adjectives in BNC (±5 window search) with no. of instances of the collocation in brackets.
Age (19659)	H	Old (38.1)	Old Age R	1. Old (1405)
Bus (4694)	I	Red (20.4)	-	1. Local (123) - Red (13)
Canyon (252)*	L	Grand (58.1)	*Canyon* not in Benson et al.	1. Grand (94)
Custard (267)	L	Yellow (21.4)	*Custard* not in Benson et al.	1. Lumpy (13) *Yellow Custard* not in BNC
Decker (130)*	L	Double (57.8)	*Decker* not in Benson et al.	1. Black (60) 2. Double (33)
Grass (3907)	I	Green (52.1)	-	1. Long (169) 2. Green (141)
Honey (1038)	I	Sweet (20)	*As sweet as honey* R	1. Clear (15) 2. Golden (12) 3. Sweet (11)

Stimulus word	Frequency	Adjective associate	Restricted	BNC most frequent collocates
Lagoon (200)*	L	Blue (23.9)	*Lagoon* not in Benson et al.	1. Blue (12)
Malice (318)	L	Aforethought (21.4)	R	1. Aforethought (9)
Oven (1316)	I	Hot (26.2)	-	1. Hot (35)
Raven (33)	L	Black (47.9)	*Raven* not in Benson et al.	1. Black (5)
Razor (388)	L	Sharp (26.1)	R	1. Sharp (9)
Ruler (884)	L	Straight (22.2)	*Ruler* not in Benson et al.¶	1. Great (21) - Straight (6)
Snow (2647)	I	White (48.1)	-	1. White (69)
Sunflower (45)	L	Yellow (25)	*Sunflower* not in Benson et al.	1. Warm (2) *Yellow Sunflower* not in BNC
Tradition (4993)	I	Old (22)	R	1. Long (192) 2. Old (91)
World (57470)	H	Wide (20)	-	1. Second (1841) - Wide (270)

* Including proper noun instances

¶ In the 'measuring instrument' meaning of the word.

3.2.2.3. Discussion

Of the 17 nouns listed above, 9 are low frequency, 6 are Intermediate frequency[36] and 2 are high frequency. There are 5 dominant responses which fall outside the 5 most frequent collocates of the stimulus words in the BNC, or indeed have no instances in the BNC at all:, *yellow custard, straight ruler*[37], *yellow sunflower, red bus*[38] and *wide world.* In the first three of these collocations the adjective describes a/the key quality of the noun, i.e. these responses are comparable to the category 1 class for the adjective-noun responses noted above in Table 6.4. It may be wrong to view the responses as syntagmatic phrasal responses at all, rather it may be that an adjective is provided which has a/the key defining feature of noun. This may also be the case for a number of the more frequent combinations (e.g. *sweet honey, white snow, green grass, sharp razor, black raven, hot oven*). Though there are syntagmatic co-occurrence data for these data in the BNC, it is difficult to assess how important frequency of co-occurrence is, in affecting the responses. They could all be interpreted as being category 1 combinations. Indeed, Deese (1962) has recognized that word association responses to nouns may be of this nature. He comments, "It is possible that the syntagmatic associates to nouns ... reflect the defining characteristics of the noun" (1962, p.82).

There are disambiguation problems in the above responses, indeed much more so than in the previous analysis (Table 6.3). It has already been noted that it is difficult to tell whether frequency effects play a major role in affecting the responses, or whether semantic stereotypy is the driving force behind the responses. Further, there are several cases where it is not clear if the respondant provided the adjective as postpositional or in attributive position to the noun: *old age/age old, sharp razor/razor sharp, wide world/ world wide.* One instance in which the adjective is without doubt provided postpositionally is in the combination *malice aforethought. Aforethought* is one of the

[36] It should be noted that 2 of these are nearly in the "L" frequency category - *oven* and *honey*

[37] It is interesting to see that the polyseme which was uppermost in the minds of the respondants for this word is different to that required by the most frequent collocating adjective of the word in the BNC. See the comments in chapter 2, section 2.2 and the possibly significant role of polysemy in affecting intuitions about collocations.

[38] This is a good example of a culturally/environmentally inspired response.

very few adjectives which only occurs in this position, and indeed exclusively with this noun according to the BNC. Further, it is not clear how the subjects might have perceived the stimuli: the response *black* to *raven* may indicate a *black* (adj) *raven* (noun) response or possibly *raven black* which the BNC occasionally tags (adj)–(noun). *Snow white* can be a proper noun and whether the respondants had the fairy tale character in mind, or the weather cannot be known[39]. It is also possible that the resulting combinations are compound adjectives, or adverbs, e.g. *razor-sharp, worldwide, world-wide.* If it is indeed the case that there is a strong left to right tendency in the production of syntagm responses (Clark 1970, p.283) and if indeed these are co-occurrence 'phrasal' responses (which is debatable), then it may be the case that the stimulus word has sometimes been treated as the first word in the combination, making our classification of the stimulus words as nouns, suspect. In sum, it seems that the typical adjective responses to the noun stimuli in the Moss & Older are not that helpful in shedding any new light on the subject of adjective-noun collocation patterns.

On the basis of the above findings (particularly the data in Table 6.3.), it is useful to discuss Clark's (1970) two rules once again. The first, the semantic realization rule is perhaps best illustrated by the response *water* to *tepid*, where *tepid* requires a [+] liquid response[40]. Because *tepid* typically only describes liquids, one can argue that this severely limits the options open to the respondants in the giving of an associate. However, many of the other words do not have such a narrow range of words with which they can combine. As such then, this 'rule' is not particularly helpful in interpreting many of the results – it does not have much predictive power, except for a small number of the stimulus words. With regards to the idiom completion rule, it has been noted that a couple of the associates are idioms, quite a number are severely restricted collocations, and some are less restricted collocations. Still others are compound nouns, and a few are sayings, i.e. they are similar to the combinations that Clark himself forwards as examples of this rule in operation. However, it has also been noted that some of the resulting collocations from the adjective stimuli are free combinations (e.g. *shallow water*) and a number seem

[39] Similarly whether the response *blue* to *lagoon* has been influenced by the film (1980), the lagoon with this name in Iceland or been inspired semantically is impossible to tell.
[40] According to BNC it rarely has non-liquid collocates, though there are exceptions, e.g. *support, tomato.*

not to be phrasal syntagm responses, but rather 'prototypical' nouns having the essential attribute of the adjective (e.g. *round ball*). While then, the two rules of Clark explain some of the cases (particularly the second – the idiom completion rule, if not interpreted very literally), there are responses which seem not to be covered by either of his two rules.

4. Summary

In closing this chapter, it is useful to highlight the key findings. There is evidence supporting Fox's view that the typical/frequent partner of a word is available to respondants if the stimulus word is the 'restricted' word in a (frozen) restricted collocation (e.g. *blond hair, blunt instrument, cosmetic surgery, drunken stupour* etc). However, there are cases where the degree of restriction seems looser than she allows, i.e. the resulting collocation is not a 'frozen' collocation. For example, *old* can describe many things, though the primary response is *man*, which is a highly frequent collocate. Furthermore, it is sometimes the case that highly frequent collocates are provided and the resulting collocation is not classified as restricted by Benson et al. e.g. *shallow water, hollow tree*. Many things can be *shallow*, and many things can be *hollow*[41]. This suggests that strong connections may exist between words not classified as restricted collocations and that the notion that restricted collocations as a class have privileged psycholinguistic representation may be questionable.

In addition, a number of the responses can be interpreted as indicating that co-occurrence frequency might not be affecting the response, but rather a key semantic feature of the word drives the association. While corpus linguists have implied that the influence of the denotational meaning of a word will lead to a mismatch between lexical intuitions and corpus data, it should be noted that this is not necessarily the case. On several occasions the adjective in the resulting collocation has its typical dictionary meaning and the combination is very frequent (e.g. *parallel lines, straight line, curly hair*). There is though, some evidence to support the view that the denotational meaning of the word will

[41] Similarly, many adjectives can describe *water* or *trees*.

affect the dominant response with the result that the dominant response *is* different from the most frequent collocation (e.g. *senior citizen, casual clothes, heavy weight*). There is then, some support for the corpus linguist argument that the denotational or salient meaning of a word may, at times, 'negatively' affect the quality of intuitions about the frequent collocates of words. However, it must be remembered that the above analysis is from free word association data. It remains to be seen whether respondants will actually be able to 'sideline' the denotational meaning of a word when asked to provide its most frequent collocate.

Some of the responses to the word stimuli suggest that respondants access material that, one would assume, would be formulaic language, according to Wray's theory: there is, for example, no paradigmatic variation in *humble pie* for example, and very little in some of the restricted collocation responses (e.g. *drunken stupour*). However, as noted earlier, these responses can be explained either by appealing to the fact that there are no other frequent viable non-idiom collocates of the stimulus words, or by arguing that some words in idioms are strongly connected to the meaning of the stimulus word, as noted earlier with the responses to *humble* and *upper*. The type of material that seems more 'locked away', in the sense that it is not typically being produced, is a collocation having its stereotypical meaning, but occurring in larger chains of language (e.g. ***a broad range*** *of NP,* ***in the early days*** etc.).

As noted in section 2.3.1 of the chapter, and borne out by the data in section 3.2.1.3, high frequency adjectives do not, typically, elicit dominant noun responses in free word association tests, and so the data fail to inform us whether high frequency adjectives do have strong connections to frequent noun collocates in the minds of native speakers. This needs to be investigated, as it is typically argued that many frequent content words are delexicalised in usage (Sinclair 1991a, p.113). The research reported on in chapters 7 and 8 aims to fill this gap in our knowledge.

Chapter 7 – Testing Productive and Receptive Knowledge of Noun Collocates of Frequent Adjectives

1. Introduction

In this chapter I report on two experiments conducted with native speakers and non-native speakers of English, testing their word associations to frequent adjective stimuli. The first experiment reported (experiment 2) is a productive one, and is, using the terminology of chapter 6, a 'controlled word association test'. The task is 'controlled' in two ways: both in the type of response allowed (a particular part of speech) and in the perceived frequency of the response: the subjects were required to produce a collocate that they believed to be a high frequency partner of the stimulus word. The objective base standard against which the associations were compared is the BNC, with Altavista search engine confirmation as explained below[1]. The second experiment reported in this chapter (experiment 3) is a retrodictive test of collocation knowledge. The same respondants are provided the same stimulus words and three options for which is the most frequent collocate of each word. This is also a type of controlled word association task, and enables us, inter alia, to investigate whether corpus linguists are correct in their view that typical collocates can be recognized, rather than produced (as noted in chapter 1, section 3.5).

2. Experiment 2

2.1. Main research question and hypothesis

This experiment explores whether native speaker teachers of English and advanced level Arab speakers of English are able to think of the most frequent collocating nouns of high frequency adjectives in a decontextualised setting. The null hypothesis is that they can, i.e. that there will be no significant difference between their responses and what the corpus data tell us are the most frequent collocates. Several factors underpin this

[1] The need for independent verification of corpus data was argued in chapter 2, section 4.

hypothesis. The first, following the discussion in chapter 2, is that the corpus is assumed to be representative of generic language in the UK. The second is the assumption that respondants automatically encode frequency information (see chapter 4): that being so, there should be no real difference between the respondants' ideas of the most frequent collocates and the corpus data. Thirdly, it should be remembered that corpus linguists have provided very little support for their position that lexical intuitions are weak, as noted in chapter 1.

The alternative hypothesis, which I shall term the 'corpus linguist hypothesis', is that high frequency collocates cannot always be produced. Such a hypothesis largely ignores the literature indicating strong frequency estimation abilities from not only linguistic, but also non-linguistic studies. The reasons forwarded for the alternative hypothesis, as noted in chapter 1, are that the effects of delexicalisation, and the saliency of the denotational meaning of the word may 'work against' frequency effects in providing typical usage examples. Additional arguments supporting the alternative hypothesis, are, broadly speaking, accessibility related theories, founded upon how a word is stored in the mental lexicon (as discussed in chapter 3, section 4.3, with regards to the Bybee and Wray theories), and how the effects of that availability affect our perception of frequency (as discussed in chapter 4 with the Tversky & Kahneman theory). There is only slim support for these accessibility related theories in relation to lexical collocations - the *small* analysis of Nordquist (noted in chapter 3, section 4.4) and the non-production of some frequent collocates embedded in larger chains of language, as reported in chapter 6, section 3.2.1.3). The experiment reported below is designed to investigate this matter further. If the null hypothesis is defeated and respondants are not able to provide frequent collocates, then it will be necessary to address in more detail the theories noted above, and their respective strengths and weaknesses in accounting for the data.

An additional interest in this study is to see whether NNSs are actually *better* at providing the most frequent noun collocates of the adjective stimuli compared to the native speakers, for the reasons that there may be less holistically stored material in their lexicons (as noted in chapter 3, section 4.3, where Wray's theory on the storage of

collocations in native speakers and non-native speakers was discussed). If holistic storage works against accurate frequency judgements in collocation, then a lexicon in which there is less holistically stored material should, logically, give respondants fuller access to all the relevant information required to make the judgement accurately.

2.2. Subjects

2.2.1. Group 1 – Native speakers (NSs)

The native speaker cohort who participated in the test were all male EFL/EAP lecturers at King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. All were qualified and had considerable experience teaching English as a foreign language. Ten of the subjects were British. The other nationalities were: American (5), Irish (2), Canadian (1), Australian (1) and South African (1). It has been noted that level of education may be a factor in affecting frequency estimation in word frequency estimate studies, and that native speakers should not be assumed to represent a homogenous group (Schmitt & Dunham 1999, pp.401, 402). The above noted group was homogeneous educationally, and furthermore, all worked in the same teaching environment.

2.2.2. Group 2 – Non-native speakers (NNSs)

The second group of subjects consisted of fairly 'fluent' bilinguals, who, we may safely hypothesize, knew the adjectives (as they are high frequency) and also knew the typical collocating nouns, but may have different 'wiring' between these words than the native speakers. As far as was possible, the NS group and the NNS group were similar in every way except for the language difference.

KFUPM has a multinational faculty. The university differs from all others in Saudi Arabia in two important ways: it is an English medium university and it is a technical university. All of the NSs who completed the task teach English (EAP) at the university. The NNSs who work at KFUPM teach a variety of subjects reflecting the university's

technical emphasis, e.g. Engineering, Computer Science, Maths etc. The lecturers who teach the technical subjects are mostly from Arab countries or Pakistan, India. The majority of these have post-graduate degrees from the UK or USA. Even though the NNSs may not have lived in an English speaking country for a number of years, it is not the case that their language is 'dormant'. All of the NNSs work in an English-speaking environment, and ideally for the purposes of comparison, live and work in the same environment as the NSs. The NNS subjects were twenty Arab faculty. Of the twenty respondants who completed the task, 19 had earned PhDs in the sciences from USA or UK; the remaining subject had a Master's degree from the USA. This meant that the education level of this group was higher than that of the NSs (of whom several have Masters degrees, but none have PhDs). All of the NNSs had lived abroad for at least 1 year, the majority for several years. Of the 20 respondants who completed the task, 16 were Saudi, 2 Jordanian, 1 Iraqi and 1 Syrian. All were male. All were mother tongue Arabic speakers.

2.3. Designing the experiment

2.3.1. Justification for methodology adopted

This experiment is a single response (doubly) controlled word association task[2] - there is no context to help the respondants in providing their frequent associates. Our interest is in strong associations between words, and this test, on the basis of the discussion in chapter 6, section 2.2.2, appears to be the best way to test this knowledge.

Schmitt (1998a), however, in discussing how best to elicit collocation knowledge is critical of such an approach, arguing that the requirement to provide a collocate for a word is not a typical language activity. He prefers the elicitation type task (used by writers such as Nordquist (2004) and Gilquin (2005a and b)), where a respondant is asked to provide a sentence containing the stimulus. He comments, "it seems desirable to elicit collocations embedded in discourse rather than in isolation" (Schmitt 1998a, p.31).

[2] 'Doubly', in the sense that the response has to be both a noun collocate *and* a highly frequent one.

Schmitt's argument that such an approach is, "more natural and realistic" (1998a, p.31) fails to recognize that the elicitation task that he advocates is still extremely artificial in nature. As Roland & Jurafsky (2002) note, "'test-tube' sentences are not the same as 'wild' sentences" (p.327), as they lack discourse coherence and context. Indeed, Roland & Jurafsky (2002, p.334) suggest that because this is so, it would be wrong to expect the resulting data to match language data from corpora.

Schmitt, in his own testing of non-native subjects' collocational knowledge of a word, found collocates from corpus data for different words and then identified typical subject areas within which the collocations were used. For example, he identified the collocates of *massive*, and categorized them into the fields of 'war' (e.g. *massive attack*), 'economics' (e.g. *massive expansion*) and 'statistics' (e.g. *massive increase*). After informing the respondants of these semantic fields, he required them to provide 3 sentences including the word *massive*. This procedure certainly helps the respondant by providing context within which to use the word. However, such an approach is not suitable for our purposes, for the simple reason that it precludes the testing of a key aspect of the knowledge of a word – its semantic preferences. Are these known (consciously) by native speakers? Do native speakers of the language actually know that *massive* is typically used in these three fields? This is the kind of knowledge (i.e. semantic preference knowledge) that corpus linguists doubt speakers of the language possess. This is the knowledge that we are testing. To find out what people know, they need to be free to look into their own minds and see what they find, rather than being sent down certain pre-determined tracks.

2.3.2. Criteria affecting the choice of word stimuli

The choice of words used as stimuli is clearly of great importance. It was decided, on the basis of Clark's (1970) left to right syntagm production argument (noted in chapter 6, section 2.3.2) to ask for the word that occurs *after* the stimulus, not *before* it, i.e. to provide a 'stimulus-slot' format, as opposed to a 'slot-stimulus' format. Such a format seems an ideal way to test knowledge of attributive uses of adjectives. Further, while

work has been conducted on adverbs and verbs (Greenbaum 1988) and adverbs and adjectives (Granger 1998), very little research seems to have been conducted on adjective-noun combinations.

It was decided to provide the subjects with common (high frequency) adjectives as word stimuli. The reasons for providing high frequency items were various. Following on from the comments made in chapter 2, section 3.1, it is clearly desirable to have the maximum possible corpus data against which to measure subjects' associations, and this is facilitated with working with frequent items. Secondly, frequent words are often used in a delexicalised way, and corpus linguists believe that delexicalisation negatively affects our intuitions about typical combinations, as noted in chapter 1, section 4.1: this belief needs to be tested. Thirdly, it helps investigate more carefully the Bybeean fusion hypothesis, as discussed in chapter 3, chapter 4.3. According to that theory, the prime candidates for fusing are collocations that are used very frequently. Finally, this investigation fills the knowledge gap noted in chapter 6, section 3.2.1.3. Frequent adjectives do not, typically, elicit noun collocates in free word association tests.

Following on from the discussions in chapter 2, section 3.1, and chapter 5, section 2, it was decided that the stimulus words should, as far as possible, satisfy the following criteria:

1. They should have a frequency of over 100 words per million – i.e. be classed high frequency.

This was achieved for all the words, see Table 7.2, column 2.

2. They should be well distributed throughout different text types in the BNC and in an American corpus.

Appendix 3 shows the distribution of the words in the BNC (British) and Brown (American) corpora. The words chosen were all well distributed throughout the different texts.

3. The generic high frequency nouns, specifically *thing, time, people* and *way* should not be high frequency collocates for more than a few of the responses. If a respondant could successfully provide *thing* as a high frequency collocate to all the stimuli, for example, the experiment would be of little value.

Table 7.1 below indicates which of the adjectives used as stimuli had the above mentioned nouns among their most frequent collocates.

Table 7.1. The stimulus words and the most frequent nouns in English. If one of the frequent nouns is among the 20 most frequent collocates of the adjectives, this is marked with a *.

	People	Thing(s)	Time(s)	Way(s)
A. Different	*	*		*
B. Difficult		*	*	
C. Full			*	
D. Good		*	*	*
E. Great				
F. Important		*		
G. Large				
H. Main		*		
I. Old	*			
J. Particular			*	*
K. Personal				
L. Possible			*	*
M. Real	*	*	*	
N. Recent			*	
O. Similar				*
P. Small				
Q. Special				
R. Strong				
S. Various	*	*	*	*
T. Young	*			

4. The stimulus words should not have high frequency collocating nouns that are very different across written and spoken corpora (BNC total versus Spoken sub-corpus).

Of the twenty words used, 15 have the BNC (complete) most frequent collocate of the adjective in the top 3 collocates of the adjective in the spoken sub-corpus. The complete

corpus and the sub-corpus concur on the most frequent collocate for 10 of these words. The words for which there is a difference are: *difficult, particular, personal, similar* and *strong.*

5. The adjectives should not have high frequency collocating nouns that are very different between British and American corpora (BNC total versus Brown).

This was more difficult to achieve, and was achieved with only partial success. On 12 occasions the BNC most frequent collocate was among the top 5 collocates of the Brown corpus, and for 7 of those, the same word was the most frequent collocate in both corpora. The words for which there is a difference are: *difficult, good, main, personal, possible, similar, special* and *strong.*

6. Altavista search engine data should confirm the BNC findings about the most common collocate of the stimulus word.

This requirement was particularly important in the design of experiment 3, and is discussed further in section 3.2, below. Suffice to say here, Altavista had to confirm that the BNC most frequent noun collocate was more frequent than the BNC tenth and twentieth most frequent noun collocates. This was the case for all the words, with the exception of *real. Real estate* had a higher frequency than *real world* or *real name* in Altavista, compared to BNC. However, the word was retained as Tognini-Bonelli (1993, 2001) has discussed this adjective at length, and it was thought desirable to test these views.

As can be seen from the above, it was not always the case that all of the stimulus words could satisfy all of the above criteria, and differences between corpora were examined, after the completion of the experiment, to investigate any responses which seemed particularly unusual, as noted in section 2.7 below.

The initial basis for the choice of words (all adjectives, as noted above) was Leech et al's (2001) word frequency data, from the BNC[3]. However, unlike that list, frequency *alone* did not determine inclusion in the list of stimuli. Frequent well-distributed items were discarded for various reasons. For example, some words were also common adverbs (e.g. *high, early*), nouns (e.g. *public, major, right*), or verbs (e.g. *open, clear*). In addition, some frequent adjectives (e.g. *available*) had strong tendencies to be used predicatively not attributively, and were, therefore, excluded.

How many words should be given to the respondants? Time trial tests indicated that respondants were fairly happy working with 20 stimulus words, which took around 5 minutes to complete. It was felt that this was a reasonable time imposition to ask of the non-native speakers whom I would approach in their offices, during working hours. Indeed, in practice it was found that when respondants knew that the task would 'only' take 5 minutes, they were happy to take part.

2.3.3. The issue of lemmatisation

Should words be lemmatized in this study (i.e. in the corpus and in the respondant data) or should we respect the individual word forms? The decision was made to select complete word forms as stimuli, using their own frequency data for the reasons given below.

Knowles & Don (2004) note that the lemma has traditionally been used to make generalizations about the words in a 'family': lexicographers group lexemes under the lemma in dictionaries. However, in recent years, the results of corpus analyses have called into question the assumptions underlying such a practice: there are sometimes important differences in the adjective partners of singular and plural nouns for example, or the noun partners of different forms of the same adjective (see below). On the basis of observations similar to these, Knowles & Don argue, "The concept of the lemma is most

[3] It should be noted, though, that the number of hits provided in Table 7.3 below, is for the specific word form, and *not* the lemma, as provided in Leech et al. (1991). Justification for adopting a word approach, rather than a lemma approach is outlined below, in section 2.3.3.

useful at a general level in highly abstract discussions of a language, but seems to be of doubtful value for detailed studies of real texts" (2004, p.72). Different forms of the same adjective and different forms of the same noun may have different collocational profiles. This is discussed below, excluding the less relevant discussion about forms of verbs, which can be found elsewhere[4].

There are cases where a particular word can collocate with different lexemes of a word to similar extents. For example, Schmitt (1998a, p.37) notes that *rare* commonly collocates with *breed, breeding* and *breeds*. Another example, from the BNC, is *difficult* which has a similar number of instances of *question* and *questions* as collocates[5]. Sinclair & Renouf (1988) believe that singular and plural forms of many nouns, "share a similar range of meanings and usage patterns" (1988, p.147), so why should we not lemmatize? The problem is that Sinclair & Renouf's comment is a general one: there are occasions where there are differences between singular and plural nouns in terms of their most common collocates. For example, Schmitt (1998a, p.37) notes that *massive losses* is common (in the COBUILD Bank of English corpus), whereas *massive loss* is not. These differences between the word forms and their collocate preferences can be seen particularly clearly in the case of fixed expressions, where no interchangeability between the different forms is allowed. For example, Sinclair notes (1991b) that the common collocates of *eye* and *eyes* are quite different; for example, the latter is used more figuratively, and some of the expressions using *eyes* where *eye* cannot replace it are: *all eyes will be on* and *in the eyes of*. Sinclair (1991b, p.495) also notes that the more literal use of *eye* occurs in certain specific fixed phrases e.g. *an eye for an eye,* resisting paradigmatic variation with *eyes*. He comments regarding the different collocates of the single and plural forms, "there is hardly any common environment between them" (1991b p.496). A simple example illustrating the difference between singular and plural forms with regard to adjective-noun collocations in phrases, is provided below in Table 7.2, below. While *great deal* (adj-noun) is very common (2673 hits in the BNC), there are only two occurrences of *great deals* in the entire corpus. The reason for this is that 'great deal' (as an adjective-

[4] For example, see Sinclair and his analysis of *decline* (1991a, p.51ff), Stefanowitsch & Gries (2003), and Hunston & Francis (2000, p.254).
[5] There are 118 instances *difficult question* in the BNC, and 88 instances of *difficult questions*.

noun combination) typically occurs in the frame *DET + great deal + prepositional phrase*, i.e. *a great deal of NP*. Comparative and superlative forms of the adjective *great* also show that the different lexemes have quite different partners: the five most frequent noun collocates of these forms, excluding proper nouns, are provided below in Table 7.2. Clearly, there are some interesting differences. In terms of phraseology, *greater part* typically occurs in *....the greater part of the NP*, *great importance* in *...of great importance* and *greatest hits* in *N [possessive] greatest hits'* e.g. *Slade's greatest hits.* It should also be noted that two of the nouns (*number* and *importance)* collocate with two of the adjective forms.

Table 7.2. Top 5 noun collocates of *great*, *greater* and *greatest* according to the BNC (Numbers in brackets = instances of collocation)

Great	**Greater**	**Greatest**
Deal (2673)	Part (254)	Number (95)
Majority (389)	Degree (250)	Importance (67)
Success (370)	Detail (197)	Hits (63)
Interest (322)	Emphasis (188)	Need (48)
Importance (317)	Number (168)	Difficulty (47)

As Barnbrook (1996) notes, it may or may not be the case that using the lemma as the node word will have an effect on which collocates are deemed frequent. However, he goes on to say that, "the default separation produced by basic forms of collocation analysis may not be a disadvantage" (1996, p.105; Clear 1993, p.277 makes a similar point). Because of the evidence forwarded above, and because our interest is a psycholinguistic one, the decision was made to retain the specific forms of the words from the BNC data and from the subjects' responses.

2.3.4. The adjective stimuli

The items finally chosen for inclusion as stimuli in this task are listed in Table 7.2 below. The BNC search procedure adopted was as follows. Part of speech (POS) tagging was utilized to restrict the search to adjective occurrences. The number of instances was noted in column 2. Next, the number of different (one place to the right) collocates was noted (column 3)[6]. It should be noted that these figures include punctuation marks, function words, adjectives, etc., i.e. they are not the number of noun collocates. However, the figures are of some use in that they highlight that some words clearly have relatively fewer collocates than others (for example, the number of collocates for *difficult* is markedly low, relative to its number of instances in the corpus). To make a listing of the word's different noun collocates (according to raw frequency), the output from the BNC search was modified - pronouns, determiners, articles, prepositions, adjectives and proper nouns were excluded. The latter were excluded, as it was felt that they were too 'British' in orientation (e.g. *Old Trafford, Old Bailey*), and should not be in a test when not all of the native speakers were British, and where non-native speakers were also being tested. In addition, collocating nouns were excluded if they were not well distributed. So, for example, though there are 279 instances of *real wage* in the BNC, they occur in only 21 texts, so the collocate *wage* was omitted from the listing of the common collocates of *real*. For similar reasons, the nouns in the collocations *real output*[7], *small bowel*[8] and *old ref*[9] were also excluded from their respective adjective lists. Occasionally, it was felt necessary to omit nouns when they were not the head noun, for example *main opposition* and *main government*. For nearly all the instances of these occurrences another noun followed, e.g. *main opposition party*, *main government leaders*. If a noun collocate occurring in the list had the same number of instances as another noun collocate, then the ranking of these items was shared. A complete list of the frequency counts of the 20 most frequent collocating nouns of the adjective stimuli is provided in appendix 4.

---

[6] The collocation window was set at +1 right, for the reason that this is the slot that has to be filled by the respondant. Further, this facilitates a more accurate comparison with Altavista data, for the reasons noted in chapter 2, section 4.

[7] There are 129 instances in 11 texts.

[8] There are 196 instances in 9 texts.

[9] There are 1812 instances in 4 texts (ESRC grant abstracts), occurring in *Old Ref No:*

Table 7.3. List of adjective stimuli, their frequencies and basic information about their collocates

Stimulus word (adjective)	BNC instances	No. of different 'collocates'(to the immediate left, i.e.-1) *	Collocation Raw frequency range for top 20 noun collocations (and highest and lowest noun collocates in range). Numbers indicate number of instances in the BNC.
A. Different	47607	2493	1213 ways -- 181 colours
B. Difficult	21621	469	204 task -- 42 conditions
C. Full	27228	1480	592 time -- 124 report
D. Good	75812	2998	1861 idea -- 343 chance
E. Great	64369	3296	2673 deal -- 135 powers
F. Important	39265	1575	1048 part -- 122 aspects
G. Large	33036	2482	1868 number -- 104 room
H. Main	23870	1841	680 road -- 152 parties
I. Old	52275	3724	2358 man -- 132 system
J. Particular	21850	1888	377 interest -- 118 concern
K. Personal	17334	1275	654 computer -- 84 accident
L. Possible	33655	1408	162 way -- 46 changes
M. Real	22204	1808	679 world -- 92 threat
N. Recent	15688	1281	2777 years -- 72 reports
O. Similar	18295	1459	294 way -- 58 problem
P. Small	41865	3119	925 number -- 166 room
Q. Special	21662	1755	630 needs -- 100 features
R. Strong	15441	1239	209 sense -- 55 argument
S. Various	15293	1496	432 ways -- 68 countries
T. Young	30262	1531	3613 people -- 105 lad

* nb, this number includes non-noun collocates

In the above table therefore, choosing just one word as an example, we can see that the adjective *strong* occurs 15441 times in the corpus (tagged as an adjective). There are 1239 different words/characters following *strong* in the corpus in a +1 word right window collocate. The most frequent collocating noun of *strong* is *sense*. There are 209 instances of the collocation *strong sense* occurring in the BNC. The 20th most frequent collocating noun is *argument*: there are 55 *strong argument* instances in the BNC.

2.4. Method

The above words were scrambled into 5 different versions for test purposes. (See appendix 5 for a version of the test). The reason for making different versions was an attempt to counter chaining effects between the word stimuli, and to counter the effects of fatigue and/or loss of concentration on the part of the teachers doing the task. Subjects were informed in the instructions preceding the stimulus words to provide a noun in the slot next to each adjective, with the additional requirement that the noun should be one that he believed to be the most frequent collocating noun of the adjective in the English language as attested by the BNC. Only one response was required, for the reason that our interest is strong associates, and, as noted in chapter 6, section 2.2.2, the single word response is the appropriate testing instrument to investigate this.

The tasks were given to the two sets of subjects (NSs and NNSs) in their office hours. The subjects completed the task individually (whether they were alone in the office or with colleagues). The respondants were asked to read the instructions and were given an opportunity to ask for clarification. It was stressed verbally (on all occasions) that the subject should not write down the first word that came to mind, but rather the word that he believed likely to be the most frequent noun collocate of the prompt word in the BNC. The task was supervised, the subjects did not communicate, and neither did they consult any reference materials to help them complete the task. For practical reasons, four NNSs were not supervised when doing the task. However, they were asked to complete the task without using reference materials and to adhere to the instruction to complete the task

quickly. Five minutes was found to be adequate time to complete the task for all the subjects.

2.5. Quantitative analyses – justification for choice of test

The challenge in analysing the data from this experiment was to assess the extent to which the subjects had been able to correctly guess the most common collocate of the stimulus word. A simple 'correct/incorrect' judgement would lose a great deal of important information, since a subject might fail to think of the most common collocate but successfully produce the second or third most common. The statistical analyses adopted, should, ideally, give credit to 'good' responses, but more credit to 'better' ones – as measured by similarity to the BNC data. An ideal test, sensitive to this, is the Mann-Whitney U test. In this test the responses from the subjects can be ranked against the BNC data. The assumption is made that 20 subjects who could not all guess the top collocate might reasonably be expected to produce the top 20 collocates between them. Since the BNC always gets the top collocate, the respondants can't be ranked against it. Instead, the BNC data is treated as if it provided 20 different guesses. This is an artificial exercise in the sense that the 20 teachers tested (whether NS or NNS) could all come up with the same (top) response, whereas the BNC is required to produce 20 different choices, and in recognition of this, two versions of the analysis are carried out, according to the focus of the research question, and described in more detail below.

2.6. Results

2.6.1. Native speakers

2.6.1.1. Analysis 1 – Excluding ties

The NS results[10] were ranked against the BNC data, and a worked example of how this ranking was calculated is provided later on in this section. In the first analysis (where the

[10] For a full list of the responses see appendix 6.

focus was on the response types, not the subjects), N was calculated not as the number of subjects, but the number of different responses given. The number was, therefore, 20 at the most, if each subject gave a different response, though in practice often lower. For example, if three teachers wrote *day* next to the stimulus *difficult*, then one of these was counted and N was reduced by 2. The final N, the number of different responses, determined how many of the BNC's top collocates (listed in appendix 4) were used in the calculation. For example, if there were ten different responses, the ranking was conducted against the ten most frequent nouns for the stimulus word from the BNC. The reason for this was to investigate whether a subject who could not provide the top collocate, could provide the second collocate and so on. That is, the test was whether a cohort of *n* responses would match sufficiently closely the top *n* BNC responses for there to be no significant difference. A 'perfect' result would feature only one response from all the subjects, the top collocate in the BNC. If more than one different response was given, however, the null hypothesis would still hold if the responses *approximated* the most frequent response - something that can be decided statistically. It is important to stress, therefore, that by using the Mann Whitney U Ranking procedure to evaluate *n* responses against the top *n* collocates from the BNC, a perfect identification of the top collocate by all subjects is not required for the null hypothesis to hold.

Below, a worked example of the ranking procedure and calculation is provided to enable the reader to better understand the figures in Table 7.4 that follows.

There were 10 different responses to the word *difficult* from the native speaker teachers: *task, time, problem, problems, decision, times, situation, job, proposition* and *choice*.

The BNC top 10 nouns are (in order from highest to lowest with the number of instances of the collocation in brackets): *task* (204), *time* (139), *question* (118), *times* (115), *situation* (101), *questions* (88), *problem* (77), job (71), *problems* (69), *thing* (69). These words are ranked in Table 7.4 from 1 for *task*, down to 9.5 for both *problems* and *thing*. Note that the last two nouns are tied in terms of frequency, hence both attract the same

rank score (9.5) in the table. These words are in non-shaded boxes, indicating that they are the BNC group 1 responses.

When we introduce the subjects' responses, they are written besides the BNC listing, but are shaded to indicate that they are from a different group. It can be seen, for example, that next to the non-shaded *task* cell there is a shaded *task* cell, indicating that this noun (*task*) was produced by one of the teachers. However, there is no shaded box next to the *question* cell, which indicates that this frequent collocate in the BNC was not produced by the group 2 respondants (the NS teachers). At the end of the list there are three shaded words, indicating that these words were provided by the respondants but they were outside the 10 most frequent collocating nouns of *difficult* in the BNC.

The next step was to rank both sets of scores, as shown in the third row of the table. The most frequent collocate (number 1) scores the highest rank (19.5), and the least frequent noun collocate scores the lowest (1). Scores which are the same attract the same rank, in a way that shares the positions fairly.

Table 7.4. Worked example showing listing and ranking of the collocates of *difficult* from the BNC and the NS teachers

Non-shaded boxes – BNC data

Shaded boxes – Teacher data

N=10

	task	task	time	Time	question	times	times
Rank	1		2		3	4	
Shared Rank	19.5	19.5	17.5	17.5	16	14.5	14.5

	situation	situation	questions	Problem	problem	job	job
Rank	5		6	7		8	
Shared Rank	12.5	12.5	11	9.5	9.5	7.5	7.5

	problems	problems	thing	decision	choice	proposition
Rank	9.5	9.5	9.5			
Shared Rank	5	5	5	3	2	1

The next step was to add up the value of the ranks. In practice, SPSS was used from this point on, but the following calculation is described step by step below:

The Group 1 (BNC null set) rank is obtained by adding up the numbers in the third row from the table above in the non-shaded cells.

19.5 + 17.5 + 16 + 14.5 + 12.5 + 11 + 9.5 + 7.5 + 5 + 5 = 118

The Group 2 (Teachers observed set) rank is obtained by adding up the numbers in the shaded cells from the third row.

$19.5 + 17.5 + 14.5 + 12.5 + 9.5 + 7.5 + 5 + 3 + 2 + 1 = 92$

Calculation of U1 and U2

$U1 = (10 \times 10) + (10x(10+1))/2 - 118 = 100 + 55 - 118 = 37$

$U2 = (10x10) + (10x(10+1))/2 - 92 = 100 + 55 - 92 = 63$

Lower of U1 and U2 = U

U = 37 not significant, one tailed, p=0.05. The critical value for N=10 is 27, p=0.05.

This means that there is no significant difference between the teachers' lexical intuitions about the frequent collocates of *difficult* and the BNC data for this word.

Table 7.5, below, shows the sum of the ranks obtained for all of the stimulus words. The sums of the ranks for the two groups (BNC and NS teacher) are listed in columns 2 and 3 respectively, and it should be noted that in column 3 after the NS teacher rank score, the number of different responses provided by the teachers is given. In calculating whether U was significant or not, the critical values for N were consulted.

Table 7.5. Analysis of native speaker responses for adjective stimuli (ties out). N varies - see third column, number in brackets

Word	Group1 – Sum of the ranks (BNC)	Group 2 – Sum of the ranks (NS teachers)	U	Significance (one tail)
A. Different	326	139 (15)	19	0.01
B. Difficult	118	92 (10)	37	Not sig.
C. Full	160.5	92.5 (11)	26.5	0.05
D. Good	224	127 (13	36	0.01
E. Great	210.5	89.5 (12)	11.5	0.01
F. Important	207.5	92.5 (12)	14.5	0.01
G. Large	439	227 (18)	56	0.01
H. Main	173.5	79.5 (11)	13.5	0.01
I. Old	280.5	184.5 (15)	64.5	0.05
J. Particular	383	212(17)	59	0.01
K. Personal	495	171 (18)	0	0.01
L. Possible	268.5	137.5 (14)	32.5	0.01
M. Real	179.5	120.5 (12)	42.5	Not sig.
N. Recent	146.5	63.5 (10)	8.5	0.01
O. Similar	422.5	172.5 (17)	19.5	0.01
P. Small	505	236 (19)	46	0.01
Q. Special	239.5	111.5 (13)	20.5	0.01
R. Strong	321.5	143.5 (15)	23.5	0.01
S. Various	296	169 (15)	49	0.01
T. Young	102	69 (9)	24	Not sig.

The above results indicate that, with the exception of the words *difficult*, *real* and *young*, the null hypothesis is defeated. The teachers' lexical intuitions about the most frequent noun collocates of the other 17 adjectives do not sufficiently correspond to the BNC data to enable us to say that teachers' intuitions and BNC data are comparable. In the

discussion that follows, in section 2.7, reasons for these responses are discussed, and the theories supporting the experimental hypothesis are examined to see whether they could account for the data.

2.6.1.2. Analysis 2 – Retaining ties

Turning our attention to the respondants rather than the responses, another analysis was conducted, this time keeping multiple same responses in the calculation and weighting them accordingly. The rationale for this was to credit multiple hits on high frequency nouns (which analysis 1 excludes). For example, if 15 respondants chose the same high ranking collocate of a word, and the remaining 5 chose words which were low frequency, analysis 1 would give us a significant score: the teachers were *not* able to provide high frequency collocates. This is because the high frequency hits would only count as one score, not 15, the calculation being done on the number of *different* responses. This second analysis incorporates multiple same item choices. Such an analysis enables us to appreciate that certain particular collocates may be popular responses from a number of the respondants. In the following analysis, therefore, N (respondants) is constant at 20, and the responses were ranked against the BNC's 20 most frequent collocating nouns for the stimulus word. As opposed to the first analysis, this procedure allows for the possibility of the teachers' responses being given a higher rank than the BNC, as the BNC must provide 20 *different* noun collocates, while the teachers' responses (theoretically) may all, be, say numbers 1 and 2, or indeed even just number 1. The test is two-tailed accordingly. Table 7.6 contains the sum of the ranks for the two groups.

Table 7.6. Analysis of responses of native speakers to adjective word stimuli (retaining instances when the same response was provided by the respondants), N remains constant (20).

Critical values N=20, two tailed p=0.01(105), p=0.05(127)

Word	Group 1 - Sum of the ranks (BNC)	Group 2 – Sum of the ranks (NNS teachers)	U	Significance (two tailed)
A. Different	527	293	83	0.01
B. Difficult	331	489*	121	0.05
C. Full	475	344.5	134.5	Not sig.
D. Good	421.5	398.5	188.5	Not sig.
E. Great	590.5	229.5	19.5	0.01
F. Important	555.5	264.5	54.5	0.01
G. Large	529	291	81	0.01
H. Main	529	291	81	0.01
I. Old	420	399	189	Not sig.
J. Particular	514.5	305.5	95.5	0.01
K. Personal	610	210	0	0.01
L. Possible	529.5	290.5	80.5	0.01
M. Real	426	394	184	Not sig
N. Recent	562	258	48	0.01
O. Similar	569	251	41	0.01
P. Small	549.5	270.5	60.5	0.01
Q. Special	561.5	258.5	48.5	0.01
R. Strong	547	273	63	0.01
S. Various	512.5	307.5	97.5	0.01
T. Young	334	486*	124	0.05

* Note that for these two results the Group 2 rank is significantly higher than the Group 1 rank.

As can be seen, using this analysis procedure, on two occasions the ranks of group 2 were higher than that of the BNC (with the words *difficult* and *young*). This indicates a very good grasp by the respondants of the most highly frequent collocates of these two adjectives. Further, on four other occasions the differences between the ranks were statistically non-significant (i.e. the null hypothesis holds for the words *full, good, old* and *real*). This means that the NS teachers were able to produce either one or a small number of the most frequent noun collocates of these adjectives when their responses are considered together as a group. In analysis 1, the finding was that there was a significant difference between the teachers lexical intuitions and the BNC data for *full, good* and *old.* The fact though, that U is not significantly different from the BNC data in this analysis, indicates that a number of teachers provided some of the *same* frequent nouns of these adjectives. More specifically for *full* 3 respondants provided *time* (ranked no. 1) and two provided *range* (rank 4), for *good*, there were 6 respondants who produced the most frequent noun (*idea*), and 2 who produced the second most frequent collocate *news*. For *old* four teachers produced *man* (the most frequent collocate), 2 produced *people* (the third most frequent collocate) and 2 provided the fourth most common noun *woman*. For the remaining 14 words U was significant at .01 significance. For these 14 words then, the null hypothesis is defeated and the alternative hypothesis holds. The native speaker teachers of English tested do not have easy productive access to the most frequent noun collocates of most of the frequent adjectives tested.

2.6.2. Non-native speakers

2.6.2.1. Analysis 1 – Excluding ties

As with the NSs, the NNS data was analysed using the procedure set out in section 2.6.1.1. The results from this analysis are given below in Table 7.7.

Table 7.7. Analysis of non-native speaker responses for adjective stimuli (ties out). N varies - see third column, number in brackets

Word	Group1 - Sum of the ranks (BNC)	Group 2 – Sum of the ranks (NNS teachers)	U	Significance (one tail)
A. Different	440	226 (18)	55	0.01
B. Difficult	102	68.5 (9)	23.5	Not sig
C. Full	210	90 (12)	12	0.01
D. Good	196.5	103.5 (12)	23.5	0.01
E. Great	218	82 (12)	4	0.01
F. Important	164	89 (11)	23	0.01
G. Large	424	171 (17)	18	0.01
H. Main	192	108 (12)	30	0.01
I. Old	137	73 (10)	18	0.01
J. Particular	257.5	148.5 (14)	43.5	0.01
K. Personal	345	120 (15)	0	0.01
L. Possible	245	106 (13)	15	0.01
M. Real	248.5	157.5 (14)	52.5	0.05
N. Recent	345	120 (15)	0	0.01
O. Similar	301	105 (14)	0	0.01
P. Small	337.5	127.5 (15)	7.5	0.01
Q. Special	372	156 (16)	20	0.01
R. Strong	147	63 (10)	8	0.01
S. Various	202.5	97.5 (12)	19.5	0.01
T. Young	155	98 (11)	32	0.05

With the exception of the word *difficult,* U was statistically significant for all of the words. For seventeen of the words this was so at $p = 0.01$ significance, and for *young* and *real* at the $p = 0.05$ level, i.e. the null hypothesis was defeated on all but one occasion.

A second analysis was conducted, this time including ties. As noted before, this allows for the possibility that the NNS group provide a smaller number of different collocates than the 20 BNC collocates against which they are measured. The significance is two-tailed accordingly.

2.6.2.2. Analysis 2 – Retaining ties

Table 7.8. Analysis of responses of non-native speakers to adjective word stimuli (retaining instances when the same response was provided by the respondants), N remains constant (20).

Critical values N=20, two tailed p=0.01(105), p=0.05(127)

Word	Group1 – Sum of the ranks (BNC)	Group 2 – Sum of the ranks (NNS teachers)	U	Significance (two tailed)
A. Different	525.5	294.5	84.5	0.01
B. Difficult	337.5	482.5*	127.5	Not sig.
C. Full	504	316	106	0.05
D. Good	504	316	106	0.05
E. Great	568.5	251.5	41.5	0.01
F. Important	541.5	278.5	68.5	0.01
G. Large	577.5	242.5	32.5	0.01
H. Main	504.5	315.5	105.5	0.05
I. Old	388.5	431.5*	178.5	Not sig.
J. Particular	502.5	317.5	107.5	0.05
K. Personal	608.5	211.5	1.5	0.01
L. Possible	524.5	295.5	85.5	0.01
M. Real	445.5	374.5	164.5	Not sig.
N. Recent	606.5	213.5	3.5	0.01

Word	Group1 – Sum of the ranks (BNC)	Group 2 – Sum of the ranks (NNS teachers)	U	Significance (two tailed)
O. Similar	609	211	1	0.01
P. Small	597.5	222.5	12.5	0.01
Q. Special	548.5	271.5	61.5	0.01
R. Strong	507.5	312.5	102.5	0.01
S. Various	490	330	120	0.05
T. Young	394	426*	184	Not sig.

* Note that for these words, the Group 2 sum is higher than Group 1.

For *young*, *old* and *difficult*, the NNS group rank sum is higher than that of the BNC rank, and the different rankings are not statistically significant. For *real*, the NNS group rank sum was lower than the BNC group rank sum, and the differences between the two groups were also not significant for this word. These data indicate that a number of subjects were choosing (the same) frequent collocation partners for *young*, *old* and *real*. For example, seven subjects provided *task* to *difficult* (rank 1); ten gave the response *man* (rank 1) to the stimulus *old*; four gave *world* (rank 1) to *real*; and seven provided *man* (rank 2) to *young*. For 16 of the words U was significant, for 11 at p = 0.01 and for the remaining 5 words - *various, particular, main, full, good* at the p = 0.05 level.

2.6.3. Comparison of native speaker and non-native speaker responses

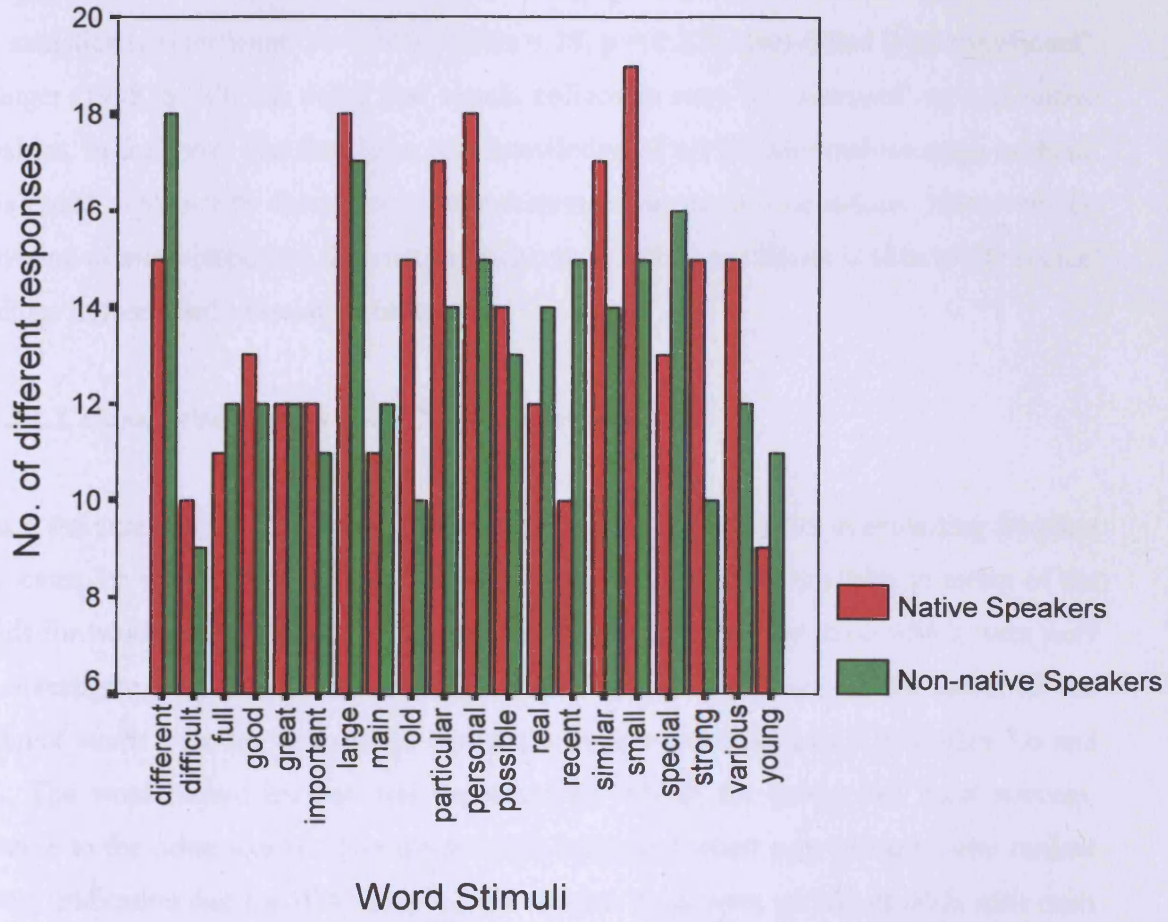
2.6.3.1. Quantitative analyses

2.6.3.1.1. Number of different responses

How similar and how different were the responses of the native speakers and the non-native speakers? It was noted in chapter 6, section 2.4.2, that in free word association tasks non-native speakers typically provide more (i.e. different) responses than native

speakers. However, it was also noted in chapter 6, section 2.4.3, that even advanced NNSs seem to be deficient in their knowledge of collocations, in which case we might expect a *smaller* number of responses to be produced by this group compared to the NSs. I investigated this by comparing the different number of responses from the two groups for each word using the Wilcoxon test to see if the difference was significant or not .

Figure 7.1. A comparison of the number of different responses provided by the native speakers and the non-native speakers



NNS < NS = 12

NNS > NS = 7

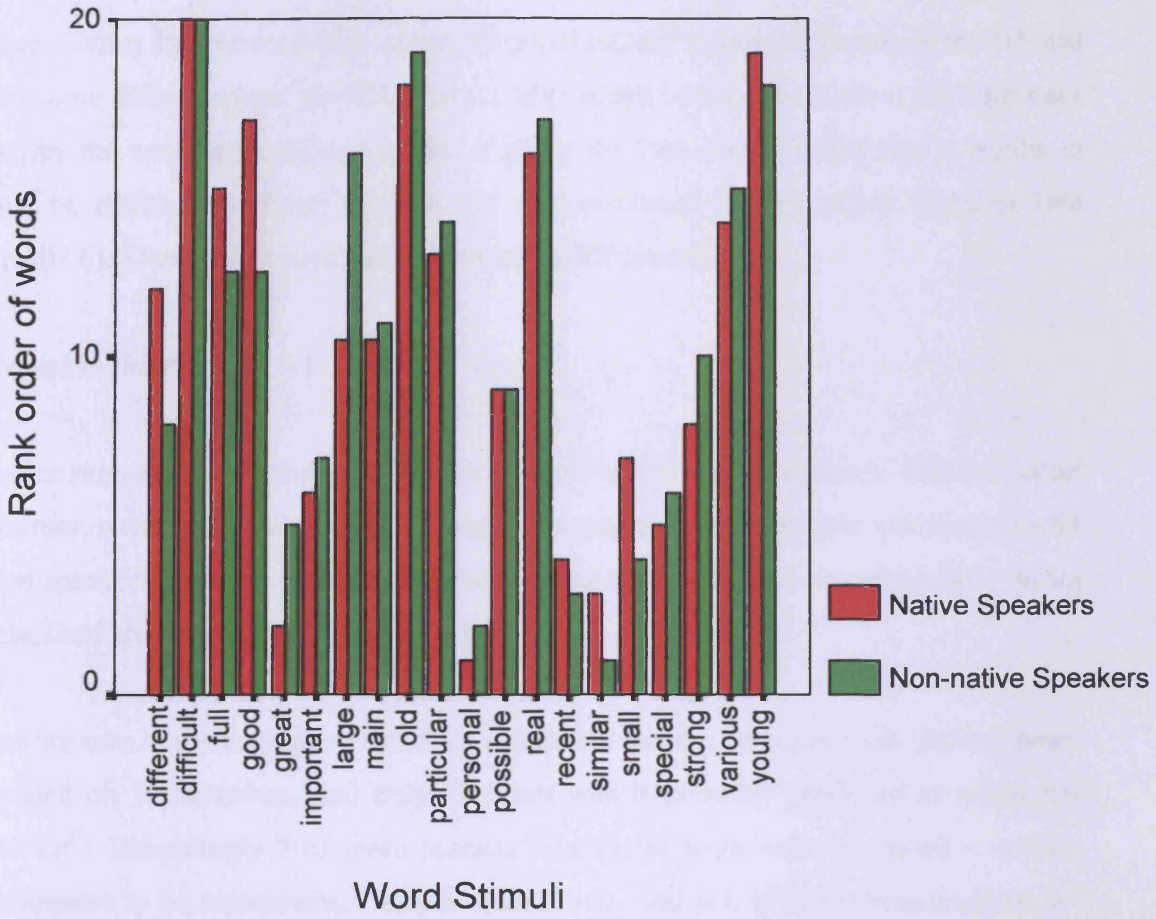
NS = NNS = 1

The data in Figure 7.1 indicate that on 7 occasions the NNSs provided more responses than the NSs (e.g. to the word *different* they provided 18 responses and the NSs provided 15 responses), and on 12 occasions they provided fewer (e.g. to the word *small*, to which there were 15 responses, and the NSs provided 19). For one word the same number of responses was provided by each group (*great*). The Wilcoxon test can be used to compare the different number of responses to each word from each group, and to see if there is a significant difference in the tendency to produce more or less responses from each group. This test not only takes into account the number of different responses from the two groups, but also the size of the differences. Using the test reveals that the differences are not statistically significant: z = 1.097, N-ties = 19, p = 0.273, two-tailed (Not significant). Granger (1998, p.148) has noted that certain collocates may be 'overused' by non-native speakers, in the sense that they have less knowledge of all the alternatives open to them. This could explain why there were fewer responses on certain occasions. However, the provision of *more* responses than native speakers on other occasions is akin to the typical findings in free word association tasks.

2.6.3.1.2. Comparison of NS and NNS group ranks

Was it the case that the two groups had similar success in, or failure in providing frequent collocates for the words? Do the NNSs responses correlate with the NSs in terms of the words for which high frequency collocates were produced, and for those which were not? To investigate this, the words were ranked according to the sum of the ranks of the different words from the two groups (using the figures from column 3 in Tables 7.6 and 7.8). The word ranked highest was the word for which the group had most success, (relative to the other words). The word which had the lowest sum of ranks was ranked lowest (indicative that the BNC data and the teacher data were greatly at odds with each other). After the words were ranked in this way for both the NS group and the NNS group, a Spearman Rho correlation was conducted to see whether there was a significant correlation between the two groups with regards to their ability to provide high frequency collocates for the words.

Figure 7.2. Group sum ranks of NS and NNS for word stimuli compared



An eyeball test on Figure 7.2, above, suggests that the respondants had 'success' with and 'problems' with the same words. The Spearman Rho correlation is .919, significant at p= 0.01 two tailed (i.e. highly significant). This means that the groups, generally, had similar success or failure in producing high frequency collocates for the same stimulus words. For example, both groups were, relatively speaking, quite successful in their ability to provide high frequency collocates for the word *difficult*. On the other hand, they were also quite *un*successful in producing high frequency collocates for the words *great*, *personal* and *similar*. There are, however, some interesting differences between the groups, as noted below in the qualitative analyses.

2.6.3.2. Qualitative analyses

Before looking into reasons why, generally speaking, the responses from both the NS and NNSs were different from the BNC data, I offer a few observations about the responses given by the non-native speaker group. If given the two sets of responses, it would, at times, be difficult to identify which list was produced by the native speakers (see appendix 6). There are, though, some interesting differences.

- The use of *thing/s*

Thing or *things* is among the top 20 collocates of 7 of the words: *different*, *difficult*, *good*, *important*, *main*, *real*, and *various*. *Thing/s* was produced on only four occasions by the native speakers – for *real*, *particular* and *various* (twice). *Thing* was, therefore, on the whole, used sparingly.

When we turn to the non-native speakers a different pattern emerges. This 'default noun' was used on 32 occasions, and only 12 times was it correctly produced as a frequent collocate[11]. Interestingly 7 of these 'correct' responses were with one word - *various*. This appears to be a somewhat 'indiscriminate' use, and yet, if that was entirely so, we might expect other 'generic' nouns (*way, people, time*) to be used that way too, but we do not. The finding that *thing* was overused, does however, add weight to the notion that even relatively fluent NNS speakers are not so aware of the collocates of words (see Bahns & Eldaw 1993; Farghal & Obeidat 1995; Nesselhauf 2003; Herbst 1996).

- The influence of the working environment

Some of the responses (quite a small percentage of the total) seem to have been influenced by the working environment of the respondants for the NNSs. While the native speakers seemed remarkably immune to 'workplace' effects on their responses,

[11] Because of the design of the experiment, the overuse of this word did not, generally, work to the advantage of the NNS respondants i.e. they were not able to perform 'better' than the NSs because of this.

this was not the case for the non-native group who produced, for instance: *difficult class, full grade, full mark, large classroom, large office, real number, similar proof, small class-room, various courses, young faculty*, and *large section*[12]. In addition, the noun *subject/s* was used on 9 occasions, and it is possible that this was, in effect, a default 'academic noun' response.

- Learner mistakes

In a few cases, the responses suggest 'learner' errors. Two of the learners provided *fashion* to the stimulus *old* (rather than *fashioned*), and one produced *fledge* to *full*. On two occasions *relation(s)* was used when *relationships* would have been more frequent (for the words *personal* and *special*). Interestingly, there may have been some orthographically induced clang responses. Firstly, two of the respondants provided *affairs* to *personal*. While this may be a genuine response (it is attested in the BNC[13]), it is possible that the respondants' knowledge of and familiarity with 'Personnel Affairs' (an important department in the university) may have influenced this response. Similarly the response *proof* to *full*, may be a result of confusion with *fool-proof*. A few responses seemed odd, (e.g. the response *past* to *old*), but some native speaker responses were also 'odd' at times (e.g. the response *outcome* to *small*).

- Cultural differences

There do not appear to be many culturally inspired differences between the two groups (see the discussion in chapter 6, section 2.4.1 on this subject). One possible exception to this, is the responses to the stimulus words *old* and *young*. *Old* received no [+ female] responses from the non-native group, whereas it received 3 from the native speaker group. Although both groups produced high frequency collocates to *young*, interestingly, the native speakers chose the number 1 collocate (*people*) 7 times, and this was chosen only once by the NNSs, whereas *young man* was produced on 7 occasions by the NNS

[12] Classes are given 'section' numbers in the university.
[13] In the BNC, there are 25 instances of *personal affairs*.

group. This may be because in Saudi Arabia, with the strict segregation of the sexes, one would talk of a *young man* or *young woman*, but not, typically, *young people*.

Having made the above observations it should be stressed that the responses were, on the whole, quite similar. In the discussion that follows, I investigate whether the responses constitute evidence for the explanations forwarded by writers who believe that the quality of our lexical intuitions is affected by factors other than frequency of co-occurrence.

2.6.4. The dominant responses

In chapter 6 an analysis of the dominant responses to free word association adjective stimuli was reported, and in section 3.2 of that chapter, the danger of making much of data for which there was little support was explained. As with free word association data, it would be unwise to make much of any single instance response in qualitative analyses: a psycholinguistic interest should focus on the dominant responses, as these may provide us with a better understanding of collocation representation. In Table 7.9 below, the 'dominant' response for each word is provided, dominant being defined as a response provided by at least 15% of the respondants. The two sets of subjects were grouped together for this analysis (i.e. N=40). It is also noted how many of the NS and NNS subjects produced the response, so that it can be observed whether one particular group was responsible for the dominant response or not. In the fourth column the relative ranking of the dominant response from the 20 most frequent collocates is recorded, according to the BNC, and numbered from 1-20, 1 being the highest rank. In the last column, the collocation is classified as restricted or not, by reference to Benson et al. (1986).

Table 7.9. Dominant responses, combining NS and NNS data. (Minimum 15% response attestation required for a particular word to be recorded in the table.)

Word	Dominant response (%age of respondants) N=40	Breakdown (No. of respondants providing the response)		BNC Collocation Rank of dominant response	Restricted?
		NS	NNS		
A. Different	People (15%)	5	1	7	N
B. Difficult	Task (25%)	3	7	1	Y
	Problem (20%)	5	3	7	Y
C. Full	Time (20%)	3	5	1	Y
D. Good	Idea (20%)	6	2	1	N
E. Great	Time (15%)	6	-	>20	Y
	Idea (15%)	2	4	>20	N
F. Important	Task (15%)	-	6	>20	N
G. Large	-	-	-	-	-
H. Main	Idea (22.5%)	5	4	>20	Y
	Street (17.5%)	3	4	6	Y
I. Old	Man (35%)	4	10	1	Y
J. Particular	-	-	-	-	-
K. Personal	-	-	-	-	-
L. Possible	-	-	-	-	-
M. Real	Time (20%)	7	1	8	N
N. Recent	Events (17.5%)	6	1	17	N*
	Event (17.5%)	4	3	>20	N
O. Similar	-	-	-	-	-
P. Small	-	-	-	-	-

Word	Dominant response (%age) N=40	Breakdown (No. of respondants providing the response) NS	NNS	Collocation Rank of dominant response	Restricted?
Q. Special	-	-	-	-	-
R. Strong	Man (22.5%)	2	7	14	N
S. Various	Things (22.5%)	2	7	11	N
T. Young	Man (30%)	5	7	2	Y
	People (20%)	7	1	1	Y

* It should be noted that Benson et al. (1986) do not usually have a plural noun entry – the singular noun entry for these words was checked and the adjective was not present in the dictionary.

There was a dominant response (defined as one which had at least 15% attestation) for 13 of the words. Of the 18 dominant responses provided[14], 5 are not frequent according to the BNC (i.e. they are outside the 20 most frequent collocates of the stimulus word). These are the collocations: *great time, great idea, important task, main idea,* and *recent event.* Of the 18 dominant responses, half are classified as restricted by Benson et al. The free collocations are: *different people, good idea, great idea, important task, real time, recent events, recent event, strong man,* and *various things.* Three of the free collocations are not very frequent according to the BNC data (*great idea, important task, recent event*). In the discussion below, in section 2.7, I refer to these associations, plus the complete record of the responses as given in appendix 6, and try to make sense of what is influencing them.

[14] Note that for some of the stimulus words there is more than one 'dominant' response.

2.7. Discussion

Is there evidence that the denotational, lone dictionary meaning of the word influenced respondants in their choices?

This is an important question to answer as it contains a key part of the corpus linguist argument as noted in chapter 1, section 4. The evidence is mixed about the importance of this factor in affecting the responses. Firstly, we examine the counterevidence. Perhaps surprisingly, there are cases where the non-denotational meaning of the stimulus word is in the resulting collocation. For example, in only 3 of the 21 different responses to *full* (*glass*, *stomach* and *tank*) does *full* have its prototypical meaning 'no space'. Interestingly, Sinclair (2004b, pp.21, 22) argues that in *full capacity* and *full range*, *full* is delexicalised, but both of these responses were provided on two occasions. Further, the meaning of *full* in *full time* (the dominant response) is not the typical/prototypical meaning of *full*. Another example, suggesting that the influence of the denotational meaning of the word is not so significant in 'driving' the responses, is the associations to *real*. Tognini-Bonelli (1993) believes that the word is "usually taken to mean 'existing in reality'" (p.118). This is not, however, the meaning of the word in the majority of the association responses. In only four of the 20 different responses (*thing, things, image, food*) does *real* have this meaning. What is more, the meanings of *real* when combined with the dominant responses (*time*, *world*) are quite idiomatic. One of the resulting collocations, provided twice by the respondants, (*real problem*) is given by Tognini-Bonelli as an example of a case where *real* has lost its, "existing in reality meaning" (1993, p.118). This loss of original meaning is, however, the norm in the responses, not the exception, as witnessed in the dominant responses and the other responses too (e.g. *real situation, real trouble*). It seems then, that in these cases, a combination is produced where the meaning of the stimulus word, (the adjective) is either delexicalised or does not have its denotational meaning. There is little evidence of such association types in the free association data, reported in chapter 6, section 7. Whereas Gavioli & Aston (2001, p.239) have questioned the accuracy of Ronald Carter's intuitions about the collocates of *real* in the light of corpus data, the responses of the native subjects reported here are

similar to BNC data: indeed, this is one of the very few words for which the BNC data and respondant data (native speakers) were statistically non-significant, and, crucially, the meaning of *real*, in all but 4 of the responses, is not its denotative meaning.

However, there is some support for the denotational, stereotypical meaning of the word affecting the responses on occasion. This is particularly clear in the case of *great*. The 'excellent' meaning of *great* is its stand alone meaning – if something is *great* it is stereotypically *good, excellent* etc. The dominant responses for this word were *time* and *idea*. In these collocations, the meaning of *great* is clearly the 'excellent' meaning. However, both of these responses are outside the 20 most frequent collocates of *great* and it should be remembered that in this task the respondants *are* being asked to provide a very frequent collocate[15]. In many of its most frequent collocations the meaning of *great* is 'large' (e.g. when it collocates with *majority, interest, importance, care* etc.) i.e., in combination with certain nouns, it does not have its stereotypical denotational meaning. One can argue that the denotational stand-alone meaning of *great* may have influenced respondants in their production of the dominant collocates[16] which are actually not very frequent according to BNC data.

But why is the evidence so mixed for the role of the denotational meaning of the stimulus word in influencing the associates? The effect of the denotative meaning *can* explain why the responses to *great* were as they were, but not the responses to *full* or *real*. What else might be affecting the production of the associates?

The adjectives used as stimuli in the experiment are all frequent, and many words are frequent because they are present in frequent phrases (see e.g. Coulmas 1979, p.239; Summers 1996, pp.262, 263; Stubbs 2002b, p.235). Some of the high frequency collocates, according to the BNC, occur 'embedded' with the stimulus word in a phrase,

[15] This is in contrast to the free association method, where the responses are not 'controlled' in any way.
[16] It should be noted, however, that the denotational meaning of the word is sometimes its meaning in a frequent collocation. The best examples from the dominant responses which support this are the collocations *main street* and *difficult problem*. In both of these cases the adjective has its denotational meaning: a *main street* is not 'minor' and it is 'large', and a *difficult problem* is a 'hard' problem, not an 'easy' one.

and others (usually with an article preceding the collocation) are complete phrases. All of the dominant collocations in Table 7.8 have one thing in common: they typically function as a complete unit, i.e. they are not embedded: they do not occur in larger phrasal chains. This is a crucially important finding. When we look at a large number of the most frequent collocates, they *do* typically occur in larger chains of language. There are three types of chain, as noted below.

Some of the frequent 'bare' collocations typically occur in frameworks of the type *DET ADJ NOUN of NP* (e.g. *a **large amount** of money*). Sinclair (1991a p. 89) notes that one of the types of relationships between the two nouns in this type of chain is that the first noun phrase is supportive of the second, and that as such, the second noun tends to be the most salient. For example, in the chain, *the usual kind of problem*, the second noun is more important in the information that it conveys than the first. Many of the noun collocates *not* provided by the respondants are these kinds of supporting nouns. For example, the dominant response to *different* was *people*. This is a reasonably frequent collocate. However, the typical syntactic pattern for the most frequent collocates of *different* is *DET different N of NP*. Indeed, according to the BNC *different types, different kinds, different parts* and *different aspects* occur around 90% of the time as the first noun phrase in the chain *DET different N of NP*. However, the case is very different for *different people* (the dominant response): it has no typical embedding patterns. Might it be that the most frequent nouns in the larger chains were not so available to the respondants, in their noun searches[17]? A similar case is the dominant response *task* to *important*. As with *different*, there is a strong tendency for many of the most frequent collocates of *important* to occur in the first noun position in *DET ADJ NOUN of NP* chains. For example, *important aspect* occurs in this chain in 90% of its occurrences in the BNC, *important feature* 66%, *important source* 65%, and *important aspects* 71%. The dominant response *task* does not have this tendency to be embedded and neither do the other combinations resulting from the other associations (e.g. *important person, important issue, important thing, important idea,* and *important information*).

[17] One might argue that because *people* is a more 'concrete' noun that it is more likely to be produced; however, as will be noted, a 'concreteness advantage' cannot be forwarded for many of the dominant responses.

Interestingly, it seems to be the case that there are instances where words from a particular semantic field fill the supporting noun slot. For example for *large*, if a 'number type' noun follows (e.g. *sum, quantity, amount*) it typically occurs in the frame *a large [NUMBER NOUN] of NP*, e.g. *a large amount of NP*. However, when we examine the respondants' answers, we find that the respondants did not typically provide words from the appropriate semantic field to fill the framework slot. Only six respondants provided 'number type' responses to *large* (*sums, amount, quantity, quantities*)[18] and the most common responses, *size* and *area,* do not, typically, occur in the supporting *NP* role in the *NP of NP* chain. *Small* is similar to *large* in this regard, in terms of its tendency to co-occur with 'number nouns' in the supporting NP role, and only 3 of the respondants provided quantity type nouns for *small* (*amount, amounts, quantity*). There were no dominant responses to speak of, with a wide variety of unrelated nouns produced.

Secondly, some of the collocations occur in adverbial chains. For *recent*, items from a particular semantic field (time period) fill the slot in the adverbial chain *in recent [time period]*, e.g. *in recent years*, *in recent months* etc). For *recent*, if the noun that follows this word is a time related noun, then it typically occurs in this chain. The dominant response to *recent* was *events* (7 responses, and there were also 5 *event* responses); however, the most frequent collocates are time related and they are very strongly embedded in the framework noted above, e.g. *recent years* (84%), *recent months* (86%), *recent weeks* (88%). The fact that there was only one 'time' association – *times* – to the word *recent* indicates that while *recent* is clearly a time related adjective, it did not easily elicit very frequent time related nouns. Like *recent*, *similar* has very frequent noun collocates that typically occur in adverbial clauses, e.g. *in a similar way, in a similar fashion, in a similar vein.* The dominant responses to *similar* were *ideas* (5), *thing* (5) and *things* (3): the resulting collocations are not found in the adverbial framework.

Thirdly, and finally, some nouns occur in collocational frameworks which are unique to that noun, in the sense that the other frequent noun partners of the adjective cannot fill

[18] However, not one respondant produced *number* or *numbers*, and *number* is the most frequent collocate by far. *Large number* occurs in the chain *large number of NP,* 95% of its occurrences in the BNC.

that framework slot. For example, *possible exception* typically occurs in the phrase *with the possible exception of*, while *with the possible way of*, with the possible solution of** are not typical. While this is a less important observation, it might explain why some frequent collocates were not provided.

Following on from these observations, it would seem that a sound explanation for the dominant productions (and omissions), would be the possibility that either the noun, the (stimulus) adjective, or the 'bare' collocation within the larger chains is not so salient or accessible as collocates which do not occur in the larger chains. For example, in the chain *in recent years*, either *recent*, *years* or *recent years* appears to be hidden, when respondants are searching for frequent collocates of *recent*. The respondants show a preference to provide a 'complete' collocation: for example, *good idea* is 'unit-like' (even without the determiner), and so too is *main street*. However, many of the most frequent noun collocates typically combine with the adjective in a larger chain of language, and they are incomplete as bare collocates: for example, *similar vein* is not complete in any real sense, and *large number* typically has a supporting role e.g. *a large number of problems*.

It would seem, therefore, that it is not the denotational meaning of the stimulus word that is the critically important driving force in affecting the quality of our lexical intuitions, but rather accessibility – and if a 'bare' collocation is typically embedded in a larger chain, then it seems to be not so accessible. Rather than arguing that respondants produce nouns which, in effect, delexicalise the stimulus adjective, it appears to make more sense to argue that some nouns are simply more accessible than others.

It may also be that this explanation could help in determining how it is that a particular meaning of a word becomes salient. Could it be the case that the availability of nouns for a particular adjective may affect our perception of the adjective's typical meaning? This explanation is consistent with Wray's theory that segmented material is analysed and the constituent material in formulaic language is less analysed. For example, returning to the *great* example noted earlier, the denotational, stand-alone, salient meaning of *great* is

good/excellent. Why is this the salient meaning and why is the 'large' meaning not salient? It could be because *great* means 'large' when it is in combination with a set of nouns in holistically stored formulas. Because these have not been segmented, the constituent parts have not been analysed, and as a result, the *great* = 'large' meaning of *great* does not enter the productive lexicon.

Can this explanation for what is going on in these responses account for why the native speaker responses to *difficult* and *real* were so similar to the BNC data[19]? It can. *Difficult* is quite different from the other adjectives used in experiment 2 because none of its most frequent noun collocates show any tendency to occur in fixed, invariable phrases. This, we would assume, assists the respondants in providing frequent associates. If we believe that an availability heuristic is being employed in the frequency searches, we would hypothesize that the responses would indeed be more accurate and less biased in such a case. With regards to *real*, while *real world* is sometimes in the chain *in the real world* it is only so in 36% of its occurrences in the BNC. This suggests that the words in this chain may be segmented, and as such, this would make the individual components more accessible than cases where the 'bare collocation' shows much more dominant embedding tendencies. The same is also true for *real time*, which occurs in *in real time* in (only) 41% of its occurrences in the BNC. Interestingly, while the NS respondants produced a set of associations that did not differ significantly from the BNC data for *real*, a very common collocate of *real* was not produced – *terms*. The 'bare' collocation *real terms* occurs 97% of the time in the chain *in real terms*. It is not surprising, given the argument above, that while the associates to *real* were, generally, very similar to the BNC data, this, the third most frequent collocate was not produced once.

There is one piece of evidence from the data which does not fit in well with the above proposed explanation and for which I can offer no solution, that is, the poor responses to *personal*. *Personal*, like *difficult*, shows no particular tendency for its nouns to be embedded, yet only one of the responses from 40 respondants was in the top 20 nouns

[19] I exclude discussion of the responses to *young* here, as the semantic preference restriction [+ animate] seems to be the key reason for this response.

(that being *problems*). Are the responses closer to BNC spoken data? They are not: of the 40 responses only 1 is in the top 20 BNC spoken sub-corpus. Various possibilities for the disparity were checked, however, none seemed to be fruitful. The Collins COBUILD English collocations CD was searched for typical collocates of *personal*, but they were no more similar to the responses than the BNC data. The most frequent collocates of *private* (the obvious synonym of *personal*) were checked in the BNC, and the highest ranking MI and z-score collocates were also checked. However, nothing convincing was found to either explain, or justify the responses.

Another question which should now be asked is whether there is any evidence that the NNSs had greater success in producing collocates which occur in the above noted frameworks. According to Wray (2002), NNSs may have a different mental representation of collocations to NSs: in particular they may have less holistically stored language in their lexicons, if they have learned a language later on in life, and have adopted an analytical approach to their studies. Two NNSs provided *interest* to *particular*. This combination, according to the BNC, is present in *of particular interest* in 60% of its occurrences. Two NNSs also produced *morning* to *good*. This is interesting. The phrase is phatic and, one would assume, a prime candidate for formulaic language status. Further, no native speakers produced the collocation *good morning* though it is frequent. However, this evidence is so marginal that it would be unwise to make much of it. As a consequence then, the results do bring into question Wray's theory about different collocation representation for NNSs. There is though a simple explanation. While the native speaker lexicon may contain holistically stored formulaic language, the NNS lexicon may contain fused formulas. If this were the case, then neither group would have such easy access to the frequent partners of the adjective stimuli in this test. Unfortunately, there is no way to test this.

It was argued in chapter 6, section 2.3.2, that we would not expect respondants to provide collocates to words in which the resulting combination would be classed an idiom. Unfortunately, this experiment sheds no further light on this matter, as so few of the frequent collocates of the stimulus words combine with the stimulus word to form an

idiom. There were occasional idiom responses: *small fry* and *old hat*, but the respondants did not, generally, produce idiom collocates and none of the very frequent adjective-noun combinations are idioms.

This experiment suggests that the noun collocate associations provided to adjective stimuli by respondants are, generally speaking, not the same as those from the BNC data. It has been argued that accessibility problems in particular, may be responsible for: (a) the under/non-production of non-salient nouns in *NP of NP* chains (e.g. *large amount, different way*); (b) the failure to provide typical collocates of words in adverbial chains (e.g. in *recent years, in a similar way*) and (c) the failure to provide nouns which are unusual, in the sense that other noun partners of the adjective do not typically fill the noun slot in the frame (e.g. *with the possible exception of*). It would seem from this experiment that the better candidates for formulaic language status (i.e. less analysed language) are not dyad collocations, but rather fixed or semi-fixed phrases or language chains. This is because, at times, the respondants do seem to access a strong (frequent) collocate of a stimulus word (e.g. *good idea*); however, the data is fairly convincing in showing that the noun items in the frameworks are not so accessible.

It follows from the above, that the previous explanation for the failure to access the frequent collocates of *small* (Nordquist 2004) may not have captured what was happening in that experiment. Nordquist noted that number collocating nouns of *small* were not produced by respondants in her elicitation task as much as predicted by her corpus data (a finding with which this experiment concurs, even though the methodology she employed in testing the knowledge of the collocates was different). Her explanation was that the 'number' collocates of *small* were fused with *small*, because of their high frequency of co-occurrence, and that the words in the resulting fused combinations had become autonomous. The experiment reported on in this chapter suggests that we may need to look at units bigger than the dyad in determining what may or may not be produced by respondants. High frequency dyads are sometimes produced (e.g. *good idea, different people, main street*) and they were also produced in the free association data, reported in chapter 6, section 3.2.1.2. Hypothesizing that frames, rather than dyads are formulaic and

hence less accessible seems better able to explain the data in this experiment, and is perhaps the key reason why corpus data and elicited data may differ on the subject of adjective-noun collocations.

It was predicted that the NS and NNS data would differ. However, very few differences between the two groups have been found, and it has been suggested that while the initial learning experiences of the two groups may have been different (i.e. holistic v analytical), prototypical holistic storage and fused storage may be impossible to distinguish in this experiment.

3. Experiment 3

3.1. Main research interest and hypothesis

This section of the chapter reports an experiment designed to investigate whether respondants can *identify* the most frequent collocate for the same word stimuli against which they were tested in the second experiment, when the test is of a multiple choice design. If provided with 3 collocates for a word (taken from the BNC, and differing in their relative frequencies), can respondants choose the most frequent collocate? In this experiment retrodictive knowledge of collocations was tested, a knowledge type which corpus linguists have suggested exists, but have provided no empirical evidence for[20].

The null hypothesis for this experiment was that, based on the results from the second experiment, respondants would not be able to identify the most frequent collocate. Though it may be the case that frequency information is automatically encoded, on the basis of the findings from experiment 2, there appears to be a bias affecting the ability to produce a frequent collocate. In a multiple choice task, it is hypothesized, chance alone will influence the ability to choose the most frequent collocation, particularly so, as many of the collocations typically occur in larger chains and this will have the effect of making

[20] It was noted in chapter 1, section 3.5, that some corpus linguists believe that we can confirm corpus evidence of frequent collocates, but not produce those collocates; however, it was argued that this should be tested not assumed.

them appear less frequent than they are. As a consequence, the frequencies of such collocations will be underestimated, as was suggested in chapter 5 for the '*bad terms*' collocation.

The experimental hypothesis is that respondants *will* be able to recognize the most frequent collocate of the stimulus words. There is slim support for this position. An analysis of the Granger 1998 study, for example, (discussed in some detail in chapter 1, section 3.1) established that corpus data generally confirmed the NS choices of the most frequent collocates of 11 amplifiers. Further, the results from experiment 1 suggested that the majority of the respondants could identify the most frequent collocate of a word. However, it should be remembered that none of the most frequent collocations in those sets (i.e. *personal computer, personal experience, bad news, bad idea*) were typically embedded in larger chains.

The results from this experiment should enable us to see whether the employment of an availability heuristic in frequency judgement may result in biases in recognition, as well as in production, as argued in the discussion section (section 2.7) in experiment 2 in this chapter.

3.2. Designing the experiment

The words used in this experiment are the same adjectives as those in experiment 2 (and the subjects are also the same). Certain principles had to be considered when deciding on how the different collocates should be chosen for inclusion in the multiple choice test.

One possibility would be to choose the most frequent collocation, and then choose a collocation that is 50% as common as this, and then choose a third which is 50% as common as the second[21]. On paper this looks reasonable: however, it is problematic. The difficulty is that such a procedure may, in fact, result in the respondant having to differentiate between the first and second most common collocates of a word in some

[21] There is nothing special about 50%, any percentage difference could be used.

cases and the first and tenth in others. For example, the most common collocate of *old* is *man* (2358 hits), and the second most common collocate is *age* (1261 hits): the second most common collocate is approximately half as common as the first. Choosing the next collocate on the basis of 50% frequency of the second, returns the collocation *old days* (567 hits), which is ranked the 6th most common collocation of *old*. For *particular*, on the other hand, the most frequent collocate is *interest* (377) and the collocate which is around half as frequent is *circumstances* (179 instances, ranked 10.5). The noun collocate which is half the frequency of *circumstances* is *cases* (91 instances, ranked 25). An additional problem with such an approach is what to do when there are no words which fall in the required slot. For instance, the most common collocate of *great* is *deal* with 2673 instances in the BNC, but then there is a huge jump to the next most frequent collocate *majority* (389 hits), clearly well below the 50% requirement. This procedure, then, does not seem feasible.

A second option might be to take a collocating word from a specified range of frequency for each item: for example, a collocation which has more than 500 instances, a second with between 200 and 300 instances and a final collocation with less than 100 instances. This would ensure parity across the different sets. However, this too, is problematic. For three word sets the collocation with between 200 and 300 instances is the most common collocation (*strong, difficult, similar*), whereas for another four it is the 10th most common (*important, main, small and young*). This procedure, like the first, seems unworkable in practice.

It was decided that the only appropriate way to choose the items for inclusion in the multiple choice sets would be to choose the collocations on the basis of their *relative* frequency collocation ranks for the stimulus words. The first rank would always be one of the choices. But what of the other ranks – what ranks would be chosen? Since the respondants had been tested against the 20 most common collocations in the productive experiment (experiment 2), it seemed most reasonable to give them choices within these 20 ranks, at equidistant ranking – i.e. ranks 1, 10 and 20. Clearly, if 'rank' determined the choice of words this does not guarantee that the differences in frequency between the

three different choices in one set will be comparable to the three different choices in another. For example, the 1st, 10th and 20th ranks of *good* (with numbers of instances in the BNC in brackets) are: *good idea* (1861), *good example* (591) and *good chance* (343). The most frequent collocation is around three times more frequent than the 'intermediate' collocation, and around five and a half times more frequent than the least frequent collocation from the ranks. However, for *great*, the figures for the most frequent, 10th most frequent and 20th most frequent are quite different: *great deal* (2673), *great man* (189) and *great powers* (135). The most frequent collocation in this set is approximately 14 times more frequent than the 10th most frequent and nearly 20 times more frequent than the least frequent collocation in the set. The difference between the most frequent and the intermediate frequency collocation was calculated for all 20 sets of collocations, and it was found that, on average, the most frequent collocation in the set was four times more frequent than the second most frequent collocation. As noted in section 2.3.2 above, Altavista concurred that of the three choices, the most frequent according to the BNC is also the most frequent in an exact-phrase search of the Internet.

3.3. Method

The three different collocate options for each adjective collocation were randomly ordered with continuation dots before and after the collocation, recognizing that many of them occur in the middle of the sentence or in a particular collocational framework (see appendix 7 for a copy of the test sheet). The subjects were asked to tick the box next to the most frequent collocate as attested in the BNC. Five different versions of the test were made in an attempt to combat fatigue, chaining etc. No time limit was given on this task, though all respondants completed it in a shorter time than the first task, i.e. <5 minutes.

3.4. Subjects

Of the 20 NS subjects who took part in experiment 2, 19 participated in this experiment. The time between the two experiments was several months. Of the 20 NNS subjects who completed the first experiment, 17 did the second. The time difference between the

administration of the two tasks varied from a couple of weeks to a couple of months, for this group of subjects.

3.5. Results

3.5.1. Native speakers

In Table 7.10, below, the BNC noun collocate ranks 1, 10 and 20 are listed from highest to lowest order for each adjective[22]. For example, *different ways* is more frequent than *different groups* than *different colours*. Next to each collocate is the number of votes given by the native speakers. So, of the 19 respondants 11 believed that *different ways* was the most frequent, 5 believed that *different groups* was the most frequent and 3 believed that *different colours* was the most frequent in the set. The way to interpret how accurate the native speakers were in their judgements is to see how often the largest figure occurs at the top of the list of the three collocations.

[22] It will be noted that there is a slight disparity to the claim that the 1st, 10th and 20th most frequent collocates were provided for all of the adjectives. For *full*, the 19th not 20th most frequent collocate was provided, and for *main*, the 18th, not the 20th most frequent collocate was provided. For *real*, the first, ninth and 18th most frequent collocates were provided and for *small* the 19th not 20th most frequent collocate was provided. These errors were due to the fact that a list not amended to exclude poorly represented collocates, was used in constructing the task. However, with the exception of *real*, as noted previously, Altavista concurs with the data about which of the three is the most frequent, and the numerical differences between the collocates provided and those that should have been used are not considered sufficiently large to challenge the validity of the findings. See appendix 4 for details of the frequency of the words used.

Table 7.10. Native speaker choices for the most frequent collocate of the adjective stimuli. Ordering of collocates is BNC order (highest to lowest order). Numbers next to words = number of votes

N=19

	A. Different	B. Difficult	C. Full	D. Good	E. Great
BNC	Ways 11	Task 5	Time 17	Idea 11	Deal 16
rank	Groups 5	Thing 8	Board 0	Example 6	Man 2
	Colours 3	Conditions 6	Report 2	Way 2	Powers 1

	F. Important	G. Large	H. Main	I. Old	J. Particular
BNC	Part 15	Number 11	Road 8	Man 7	Interest 8
rank	Element 1	Areas 7	Aim 5	Friends 12	Circumstances 4
	Aspects 3	Room 1	Features 6	System 0	Concern 7

	K. Personal	L. Possible	M. Real	N. Recent	O. Similar
BNC	Computer 12	Way 4	World 14	Years 6	Way 4
rank	Responsibility 6	Alternative 9	Name 2	Survey 11	Lines 5
	Accident 1	Changes 6	Estate 3	Reports 2	Problem 10

	P. Small	Q. Special	R. Strong	S. Various	T. Young
BNC	Number 9	Needs 6	Sense 1	Ways 17	People 13
rank	Children 8	Care 3	Views 11	Groups 1	Person 4
	Companies 2	Features 10	Argument 7	Countries 1	Lad 2

The most appropriate statistical analysis for the data is the Chi squared (χ^2) one sample test (also called goodness-of-fit test). The null hypothesis for this experiment is that, according to chance, the responses will be evenly spread between the alternative options, i.e. that the 3 options should attract a roughly equal number of votes (6 or 7 as there are 20 respondants). It should be noted that this is the *opposite* hypothesis than that given in

experiment 2, and the justification for holding it comes from the results from the first experiment. The experimental hypothesis is that the respondants would be able to identify the most frequent collocate.

In using the chi-squared test we must be sure that the expected number for each cell be no less than 5. This condition was satisfied. Items in cells are independent, and the actual numbers obtained are used. Because df = 1 (number of categories minus 1), Yates' correction for continuity is applied (see below).

3.5.1.1. Analysis 1 (Majority choice)

In Table 7.11 below, the respondants' answers are compared with the most frequent collocate in the BNC for each word.

Table 7.11. Words which the majority of NS respondants considered to be the highest BNC rank

	Observed 1	Observed 2	Observed 3	Total
BNC rank 1	13	5	2	20

For 13 of the 20 words the most frequent collocation (according to the BNC) was chosen by the majority of the respondants (if we compare the number of votes for the most frequent collocation, with the other choices). On 5 occasions the second most frequent of the three collocations provided was chosen as the most frequent by the majority of the respondants, and on 2 occasions the least frequent collocation (according to the BNC) was believed to be the most frequent by the majority of the respondants. However, we are not really interested in 1 compared to 2 or 3, but rather 1 compared to 2 and 3 together. The respondants are being asked to choose the most frequent collocate, and so when they did not do so, it is not of any real importance whether they chose the 2^{nd} or 3^{rd} most frequent option – the key thing is that they did not choose the most frequent. For this

reason, observed 2 and 3 can be grouped together. If we do this the table changes as noted below and χ^2 can be calculated.

Table 7.12. Words which the majority of NS respondants considered to be the highest BNC rank (combining observed 2 and 3).

	Observed 1	Observed 2+3	Total
BNC rank 1	13	7	20

$\chi^2 = 9.02$, df = 1 (χ^2 is significant at 6.64, p = 0.01, one-tailed). Therefore this result is significant. However, because df =1 in the above, Yates' correction for continuity should be applied. The resulting figure is $\chi^2 = 7.65$ (Chi square is significant at 6.64, p = 0.01, one tailed). Therefore this is still significant, even after the correction statistic is applied. This means that the null hypothesis is defeated. Though there are occasions when the respondants fail to choose the most frequent collocate, the most frequent of the three options for the twenty adjectives was chosen in a statistically significant way.

3.5.1.2. Analysis 2 (Votes)

In addition to the above analysis, another analysis was conducted where the number of votes was added for each of the three choices of collocations (most frequent, medium frequency, lower frequency). There is a possibility that this calculation would not result in a significant score for χ^2, if it were the case that the responses were fairly evenly split between the medium and low frequency choices, and that the actual number of votes for the most frequent collocation was not so high. Such would be the case, for example, if seven of the respondants believed that the most frequent choice was the most frequent, and six believed that the medium frequency was the most common and another six that the lower frequency choice was the most frequent. As with the second analysis above, the number of votes for 2 and 3 ranks are combined.

Table 7.13. Number of votes that each word received, (combining observed 2 and 3).

	Observed 1	Observed 2 + 3
BNC rank 1	195	185

χ^2= 55.2, df =1, highly significant. Incorporating Yates' correction for continuity χ^2= 54.43, df =1. This is also highly significant. It should be noted that although the figure under Observed 2 and 3 is very similar to Observed 1, this box, would, by chance have 2/3rds of the votes, not half of them.

3.5.2. Non-native speakers

In Table 7.14, below, the BNC ranks are listed from highest to lowest order as was the case in Table 7.10. Next to each collocate is the number of votes given by the NNS teachers who were attempting to identify the most frequent collocate. So, of the 17 respondants, 15 believed that *different ways* was the most frequent, 0 believed that *different groups* was the most frequent and 2 believed that *different colours* was the most frequent of the three collocations.

Table 7.14. Non-native speaker choices for the most frequent collocate of the adjective stimuli. Ordering of collocates is BNC order (highest to lowest order). Numbers next to words = number of votes. (N=17)

	A. Different	B. Difficult	C. Full	D. Good	E. Great
BNC	Ways 15	Task 10	Time 17	Idea 11	Deal 8
rank	Groups 0	Thing 4	Board 0	Example 6	Man 8
	Colours 2	Conditions 3	Report 0	Way 0	Powers 1

	F. Important	G. Large	H. Main	I. Old	J. Particular
BNC	Part 6	Number 11	Road 12	Man 4	Interest 9
rank	Element 2	Areas 2	Aim 2	Friends 9	Circumstances 4
	Aspects 9	Room 4	Features 3	System 4	Concern 4

	K. Personal	L. Possible	M. Real	N. Recent	O. Similar
BNC	Computer 15	Way 12	World 12	Years 11	Way 9
rank	Responsibility 2	Alternative 3	Name 1	Survey 6	Lines 1
	Accident 0	Changes 2	Estate 4	Reports 0	Problem 7

	P. Small	Q. Special	R. Strong	S. Various	T. Young
BNC	Number 11	Needs 10	Sense 0	Ways 15	People 7
rank	Children 3	Care 4	Views 4	Groups 1	Person 10
	Companies 3	Features 3	Argument 13	Countries 1	Lad 0

3.5.2.1. Analysis 1 (Majority choice)

In table 7.15, below, the respondants' choices are compared with the most frequent collocate in the BNC for each adjective.

Table 7.15. Words which the majority of NNS respondants considered to be the highest BNC rank

	Observed 1	Observed 2	Observed 3	Total
BNC rank 1	15	2	2	19

Of the 3 choices available to them, on 15 occasions the majority of the respondants chose the collocation which was the most frequent according to the BNC. On two occasions the second ranked collocate was chosen as the first, and on two occasions the least frequent collocation was believed to the most frequent by the majority of the respondants. For the word *great* the respondants were tied whether *great deal* (8 votes) or *great man* (8 votes) was the most frequent collocation. This word was excluded for this first analysis because there was not a single favoured choice, hence the total of 19 in Table 7.15. The observed 2 and 3 are combined in Table 7.16, below.

Table 7.16. Words which the majority of NNS respondants considered to be the highest BNC rank (combining observed 2 and 3)

	Observed 1	Observed 2+3
BNC Rank 1	15	4

χ^2= 17.813, df = 1 (χ^2 is significant at 6.64, p = 0.01). Because df = 1 in the above, Yates' correction for discontinuity should be applied, χ^2 = 15.817. Therefore, this is still significant, even after a more conservative statistic is applied. What this means is that the respondants' choices can not be explained by chance. They are able to recognize the most frequent collocation for most of the 20 sets of words, in a statistically significant way.

3.5.2.2. Analysis 2 (Votes)

In addition to the first analysis, another analysis was conducted where the number of votes was added for each of the three choices (most frequent, medium frequency, lower frequency), i.e. the same procedure as native speaker analysis 2, section 3.5.1.2. The

number of votes for the 2 and 3 ranks are combined. This enables us to include the data for the responses to *great*.

Table 7.17. Number of votes that each word received, (combining observed 2 and 3)

	Observed 1	Observed 2 + 3
BNC rank 1	205	135

χ^2= 111.22, df =1, highly significant. Incorporating Yates' correction for continuity, χ^2 = 109.87. This is highly significant.

3.6. Discussion

This experiment demonstrates that the same subjects who (generally) failed to provide the most frequent nouns of the adjectives in a production task, were able to identify the most frequent collocation in a statistically significant way, when it was presented in a multiple choice design format, along with the tenth and twentieth most frequent collocations. This is empirical justification for the view that respondants do have retrodictive knowledge about the most frequent collocate of a word. The results from the two sets of subjects are statistically significant in all the analyses conducted. Despite these significant results the native speakers' responses were at variance with BNC/Altavista data on more occasions (i.e. 7 times) than the non-native speakers (i.e. 4 times, with an additional tie – *great*). The words for which the preferred native speaker choice was different from the BNC data were: *difficult, old, possible, recent, similar, special* and *strong*. The words for which the preferred non-native speaker choices were different from the BNC data were: *important, old, strong* and *young*. In the discussion below, I look at some possible reasons for these differences.

- Difference between spoken sub corpus and main corpus

On two occasions the spoken sub-corpus of the BNC differs on the relative rankings of the sets: *difficult* and *strong*. *Difficult thing* is ranked higher than both *difficult task* and *difficult conditions* in the sub-corpus, and *strong argument* is higher than both *strong sense* and *strong views* in the sub-corpus also. *Difficult thing* was preferred by the NSs and *strong argument* was preferred by the NNSs. It is possible, then, to explain these differences by arguing that the responses might have been influenced more by the spoken language.

- Noun ranks

As explained in chapter 5, section 2, one of the problems in conducting ranking or recognition tasks with collocations is that the frequency of the nouns in the collocations may be ranked rather than the relative frequencies of the collocations themselves. Because this is a possible explanation for differences between the responses and the corpus data, the raw noun ranks were checked against the collocation ranks. For six of the 20 different sets of collocations the noun in the most frequent collocation is *not* the most frequent noun of the nouns in the other collocations in a noun (POS) search in the BNC. This was the case for the *various, great, different, good, special* and *difficult* sets of collocations. For the *difficult* set, the NSs preferred the collocation *difficult thing* to *difficult task*, and *thing* is more frequent in the BNC than *task*, even though the collocation *difficult task* is more frequent than *difficult thing*. The possibility that the NSs ranked the noun, rather than the collocation in this case should be considered, and this might explain why *difficult task* was not chosen to be the most frequent collocation in its set by the NSs.

- z-score

A third possible explanation for the differences is that the respondants chose the collocate, not according to the raw frequency collocation frequency, but on the basis of

the strength of attraction between the items and it is possible that strength of attraction will differ from raw frequency figures (see the discussion on this in chapter 5, section 2). The z-score is a statistical measure which calculates the chance of a node (the word under investigation) and a collocate co-occurring. (It is calculated using the observed frequency, expected frequency and standard deviation of the collocating item). The z-scores for the different collocations in the sets were compared with the collocation frequency ranks. With 2 exceptions, the raw frequency collocation data and z-score *concur* on which collocate comes to the top of the ranks. The two exceptions are the *possible* and *important* sets. *Possible alternative* has a higher z score than *possible way* (35.8 as opposed to 24.9). What this means is that of the total number of times that *possible* and *alternative* co-occur with *way*, a greater proportion of them entail *possible* and *alternative* than *possible* and *way*, even though *possible way* is more frequent overall than *possible alternative* (the difference being due to the relative frequencies of *alternative* and *way* in other contexts). In the NS data, *possible alternative* scored 9 votes as opposed to the most frequent collocation (*possible way*) which scored only 4. It may be therefore that the strength of the relationship, rather than the raw frequency collocation frequency had an influence on the subjects' responses in this case. The other exception was the difference in the *important* set. In this set, *important part* is the most frequent of the collocations, but the z-score of *important part* is slightly lower compared to *important element* (36.4 compared to 37.2). This is a negligible difference, and not of any real consequence, as the NNSs showed a preference to choose *important aspect* not *important element.*

- Embedding

Of the twenty collocations which are the most frequent in their sets, five occur in an invariable collocational framework in over 80% of their occurrences in the BNC: *great deal (of), large number (of), strong sense (of), (in) recent years, (in a) similar way.* There are then, using the terminology employed earlier on in this chapter, three collocations found in supporting noun phrases, and two found in adverbial chains in the vast majority of their occurrences in the BNC.

Three of these collocations were poorly recognized as being the most frequent in their sets by the native speakers (*recent years, similar way* and *strong sense*). Did the native speakers fail to access the collocation frames within which these collocations occur? Evidence *against* this hypothesis would be the correct choice of *great deal* in its set and also *large number* from the *large* set. I argue below, however, that the recognition of these collocations can be explained in a way that enables us to retain the embedding explanation.

Great deal typically occurs in the chain '*a great deal of NP*' in the BNC. *Deal* is a good example of a supporting noun in such a chain. If it were the case that *a great deal of NP* was stored as a chain, given the arguments forwarded earlier, we would not expect the respondants to generate enough exemplars of its use to accurately rank its frequency. Of course it may be that the theory is wrong. However, there is an alternative explanation. *Great deal* is not only an adjective-noun combination, it is also a very common adverbial. The respondants may have had in mind the adverbial sense of the combination (e.g. *he laughs a* ***great deal***), rather than the adjective noun sense (e.g. *a* ***great deal*** *of money*). It is more difficult to explain the accurate ranking of *large number*, other than to note that the respondants were fairly evenly split on whether *number* or *areas* was the most frequent collocate, and *large areas* is less embedded than *large number* in the supporting *NP of NP* frame (67% versus 95%). It is possible, therefore, to draw on the embedding theory to explain 3 of the 7 cases for which the NS responses differed from the BNC data. This adds some support to the slim evidence noted for the effects of embedding on *recognition* ranking abilities, forwarded in chapter 5, section 3.6.1. The NNS group had similar problems in correctly identifying *strong sense* and *great deal* as the most frequent collocations in their respective sets.

- Miscellaneous explanations

There may be a cultural explanation for one of the collocation differences for the NNS – the same cultural difference as noted earlier in this chapter, section 2.6.3.2. For the

stimulus word *young*, the majority of the NSs chose *young people*, whereas the majority of the NNSs chose *young person*. The collocation *young people* implies (implicitly) a mix of male and female (*young men* or *young women* would be more appropriate terms for single sex groups). *Young person*, however, avoids the inference of the mixing of the sexes, and although the BNC indicates that *young man* (2667 instances) and *young woman* (971 instances) are considerably more common than the gender-neutral *young person* (264 instances), the NNSs did not believe this to be the case. There may then be some Saudi/Arab/Muslim cultural influence upon this response, indicating, inter alia, the potential importance of societal values in judgements of frequency.

Regarding the other responses which were out of line with the BNC data, there seems to be no obvious reason for the differences. Somewhat surprisingly the NNSs favoured *important aspects* over *important part*. This seems surprising: *part* is more general, and we might expect the NNSs to be more attracted to it, rather than the considerably less frequent word *aspects*. There is an interesting difference with *special*. The NSs tended to prefer the noun *features* which would describe objects, e.g. cars, phones etc., whereas the non-native speakers preferred *needs*. *Special needs* (the most frequent collocation in the set) is typically used in educational contexts. A very surprising result (considering that *man* was by far the most common response in the productive task from both the NS and NNS groups to the word *old*) is that for both groups, the preferred collocation in the recognition task was *old friends* in which the adjective is non-inherent.

This experiment demonstrates NS and NNS abilities to correctly choose the most frequent collocation, from a list of three, for twenty different sets of collocations. This gives empirical justification to the belief noted in chapter 1, section 3.5, that native speakers, while they may not be able to provide frequent collocates, are able to recognize them. It has been noted, however, that there were some collocation sets for which the NS preference did not concur with corpus data: not all the high frequency collocations were correctly chosen. Several explanations have been offered to explain these differences, and one explanation that is particularly interesting is that recognition frequency estimation abilities (in addition to productive frequency estimates) may be biased by respondants

being unable to recognize that a particular combination occurs in a larger framework, and that, as a result, the combination is considered less frequent than an objective base suggests.

4. Summary

The results from the above two experiments are, broadly speaking, in line with the corpus linguists' view that productive knowledge of frequent collocates is weak, and receptive knowledge of frequent collocates is strong. However, the explanations forwarded for this difference are not the explanations forwarded by corpus linguists. Rather, it has been proposed that either the stimulus word (i.e. the adjective), the noun, or the 'bare' collocation in collocation frameworks are not so accessible as collocating words which occur in complete 'bare' collocates (e.g. *good idea*). The availability heuristic hypothesis of Tversky & Kahneman (1973), when interpreted in the framework of Wray's (2002) formulaic language model does seem able to explain many of the results, though it should be noted that the frame, rather than the existence of strong links between words in dyads seems to be the critical factor affecting the responses. Lending support to the tentative explanation offered in chapter 5 for the under-ranking of *bad terms*, it has also been argued that availability restrictions may also affect receptive / recognition estimates of frequency.

Chapter 8 – Testing Productive and Receptive Knowledge of Adjective Collocates of Frequent Nouns

1. Introduction

Experiment 2, reported in chapter 7, required respondants to provide high frequency collocating nouns to 20 frequent adjectives – the stimulus words were adjectives. It was found that the respondants' ideas about which nouns were frequent partners of the adjectives differed significantly from the BNC data for most of the words. It was also noted that the NS and NNS data were not very different from each other. In experiment 3 it was found that the respondants were able to recognize frequent collocates of the adjectives (rather than produce them).

But was it the case that the provision of the frequent adjective (rather than the noun) in experiment 2 may have made that task an 'unnatural' one[1]? There are some plausible reasons to hypothesize that providing nouns rather than adjectives as stimuli might (positively) affect the 'quality' of the responses in terms of the similarity of the associations to BNC data. In essence, the new research question addressed in this chapter is: will it make any difference if the respondants are given the noun and asked for the adjective collocate, rather than being given the adjective and asking for its high frequency noun collocate? In section 2.1 below, this possibility is discussed. A second productive task is then reported, similar in design to experiment 2, but in which the stimulus words were frequent nouns, rather than frequent adjectives. A recognition task is also reported (experiment 5) designed in the same way as experiment 3, to test retrodictive knowledge of collocates for the same noun stimuli.

[1] See below for further explanation of what I mean by this term.

2. Experiment 4

2.1. Rationale for experiment

Will the provision of noun stimuli (as opposed to adjective stimuli) assist the respondants in providing frequent collocates? Corpus linguists (as noted in chapter 1) have not really considered whether collocation knowledge may be affected somehow by the *type* of word (i.e. the part of speech of the word) given to respondants. In what follows, arguments for and against the idea that the provision of noun stimuli will positively affect the quality of collocation knowledge vis à vis corpus data are discussed.

The first argument in favour of the idea noted above, is that it is commonly recognized that the noun in an adjective-noun collocation is the base of the collocation. For example, Fontenelle (1998) comments, "in lexical collocations (of various types) a distinction is often drawn between the 'base' (the noun in the case of a V + N or A + N collocation) and the 'collocate' or collocator" (p.192; see also Benson 1989, p.6; Keller & Lapata 2003, p.461; Cowie 1998, p.222 who make similar points). In experiment 2 the respondants were given the collocate and asked to produce the base. According to Fontenelle, such a design places the cart before the horse. Providing a noun stimulus for respondants, i.e. supplying the respondants with the base, and asking for frequent adjective collocates would seem to be more 'natural' in such a view.

The second piece of support for the view that nouns should more readily elicit high frequency adjective collocates is somewhat related to the first, and is the argument that the meaning of an adjective is often determined by the noun which it describes. Aitchison (1994) comments that, compared to nouns, adjectives, "are less independent and often rely for their interpretation on the noun to which they are attached" (p.104). She notes, for example, that *mad* has different meanings dependent upon whether it describes a man, a dog, an idea or an evening (Aitchison 1994, p.59). It may be that the respondants struggled to provide frequent collocates in experiment 2 for the reason that they were given a 'dependent' word with no context, and if adjectives are indeed dependent on the

nouns they describe, the fact that they occurred alone, may not have given the respondants 'a fair chance' in the production of a frequent noun collocate. So then, these two points: the role of the noun as the base in the adjective-noun collocation, plus adjective dependency, may have made experiment 2 simply too difficult.

It is useful to stop here and ask whether the two points mentioned above are always true. The point of Fontenelle (1998), that the noun is the base of an adjective-noun collocation, seems reasonable in most cases. When one wants to describe a man or woman, for example, one may call upon any adjective that one so desires to describe that person, e.g. *a grumpy man* or *an intelligent woman*. In both of these cases the noun does indeed seem to be the base of the collocation, i.e. one starts with the noun, and the adjective helps provide more detail about that noun by classifying or describing it. However, as noted in chapter 7, section 2.7, there are cases when the adjective-noun collocation is typically only part of a larger framework. For example, in the collocation *great deal*, typically used in *a great deal of NP*, the meaning is *a large amount of NP*, but this meaning of *great deal* occurs only when *great deal* is in the chain *a great deal of NP* – otherwise *great deal* refers to a good bargain. So, when one wishes to speak of there being *a lot of X*, one can call upon the expression *there's a great deal of X,* but one cannot first select the noun *deal*, and then qualify it with the adjective *great*. In such a case it seems difficult to argue that the noun is the base of the collocation: the two words (*great* and *deal*) seem to combine to form a single unit, or perhaps, more particularly, and on the basis of the argument in chapter 7, the whole framework (*a great deal of NP)* functions as a single prefabricated unit. A similar case to the *a great deal of NP* case, also coming from the data reported in chapter 7, would be the collocation *similar vein*, where again, the notion of orientation from the noun to the adjective seems unlikely. The Longman Dictionary of Contemporary English (LDOCE) gives the relevant entry for this as follows: "'in a...vein' in a particular style of speaking or writing about something". The word *vein* only has this denotation in this structure: the structure determines the meaning. While then the view of Fontenelle (the noun being the base in adjective-noun collocations) seems valid in the majority of cases, it seems that there may be exceptions – particularly with collocating items inside set phrases.

The second observation above, that adjectives are less 'independent' than nouns, seems plausible, but perhaps not uniformly. A large number of the most frequent nouns are abstract, and are, perhaps, equally 'indeterminate'. Lewis (1997) argues that words such as *mind*, *way* and *thing*, "hardly have an existence independent of the multi-word phrases and expressions in which they occur" (p.24). If this is so, then the provision of a noun (rather than an adjective) in the experiment will not *necessarily* assist respondants in providing frequent collocating partners. If the word is typically delexicalised, or has different meanings it may not be easy for the respondants to provide a collocate for the 'right' polyseme of the word, i.e. the polyseme in the most frequent collocation, For example, as noted earlier, in chapter 1, in the discussion on Fox's (1987) experiment, respondants often provided body part collocates of *feet*, rather than collocates which gave the word its measurement meaning – the meaning in the most frequent collocations. As a result, their lexical intuitions were found to differ from the corpus data used in Fox's research. Further, it is important to note that sometimes an adjective determines the meaning of the noun rather than the other way around. For example, with regard to the noun *matter*, the adjective *inorganic* signifies the matter to be a substance; the adjective *serious*, that the matter is a topic or issue; and the adjective *printed* that the matter is a book/newspaper etc. The above factors, plus the idea of sequential processing and Clark's (1970) view that there is a left to right tendency to syntagm production in word association data, weaken the case that the provision of the noun will assist the respondants in providing typical collocates.

2.2. Research questions

Can respondants provide frequent adjective collocates to a set of noun stimuli, corresponding to BNC data?

The null hypothesis is that they can, i.e. that there will be no significant difference between the respondants' associations and the BNC data. The justification for this belief is that the corpus is deemed to be representative, and the respondants are assumed to automatically encode frequency information. Further, on the basis of the above

discussion, though there may be exceptions, it seems that in the *majority* of cases the provision of the base of the collocation should help respondants in providing typical collocates. As such, it is hypothesized that the responses will be much closer to the BNC data, than was the case in experiment 2 reported in chapter 7.

The alternative hypothesis is that the two sets of data will differ from each other. Following on from the discussions in chapters 1, 3 and 6, and the discussion in section 2.1 above, there are a number of possible reasons why the two sets of data (i.e. corpus data and elicited data) may be different. Firstly, the denotational meaning of a word may affect the respondants' ideas about typical collocates, but it may be that in its most frequent collocations the stimulus word is delexicalised or has a secondary meaning. In addition, it may be the case that the respondants are not aware of the dominant semantic prosody of a word as discussed in some detail in chapter 1, section 3.3. Another possibility is that the respondants will provide a collocation for a polyseme of the word that is not the polyseme which co-occurs with the dominant collocational partners (e.g. *straight* might be provided as an associate to *ruler*, where *ruler* refers to a measuring instrument, rather than *great* being provided as an associate where *ruler* is a political office[2]). If this latter kind of 'error' occurred, it would add support to the argument that respondants do indeed have problems in being consciously aware of the most common meaning of a particular word (see chapter 1, section 3.2), or, more particularly, the most common meaning of the word in its most common collocation. A final explanation may be that accessibility restrictions play a role in negatively affecting the 'quality' of the responses. In experiment 2 it was this explanation which appeared best able to account for the data. This latter possibility leads on to a secondary research question: Will the collocating partner of a word embedded in a collocational framework be produced?

This is a difficult question to answer. If it is the case that it is peculiarly the first noun that is less salient in *ADJ NOUN of NP* frameworks, then, if this noun is provided, frequent collocates qualifying it in the framework may be provided, assuming that the respondants are able to access the collocate within the framework. However, if the adjective *and* noun

[2] This example comes from the Moss & Older data discussed in chapter 6, section 3.2.2.2.

(as a pair) are hidden away in this larger chain, then the provision of the noun may not so readily activate the partner of the word in the combination. Regarding adverbial chains (e.g. *in recent years*), it was noted in chapter 7 that the stimulus word *recent*, did not, typically, elicit the noun in the adverbial chain. If the word *years* in *in recent years* is the base of the collocation, then when it is provided as a stimulus word it may be more likely to elicit *recent*. However, in some of these collocations, one can argue that the key word is not the noun, but rather the adjective: *in recent years* means 'recently'. As such, the provision of the noun may not help in such a case.

2.3. Choice of noun stimuli

Many of the nouns which are the frequent partners of the adjectives in experiment 2 are what Schmid (2000) terms 'shell nouns'. Schmid defines shell nouns as a subset of abstract nouns, "used by speakers to create conceptual shells for complex and elaborate chunks of information" (2000, p.6). He gives the examples of *case, chance, fact, idea, news, point, problem* etc. as typical examples (ibid, p.3). He believes that these nouns can be frequent (ibid, p.6), typically have unspecific meanings (ibid, p.15), and are similar to anaphoric pronouns (ibid, p.16) in that their context determines their meaning. Further, he says that they occupy the middle ground between full content nouns and pronouns with an anaphoric reference (ibid, p.15) in discourse. He notes that such nouns have been neglected in linguistics (ibid, p.4), and it also seems that they were (sometimes) neglected by the respondants in the associations provided in experiment 2, particularly when they were the supporting noun in a *NP of NP* construction. To facilitate a better comparison with experiment 2, it was decided to provide a majority of abstract (shell) nouns as the stimuli.

A large number of nouns were considered for inclusion in the list (around 50). Only nouns which occurred around 10,000 times or more in the BNC were considered (i.e. around 100 instances per million words). These nouns are highly frequent, and similar, with regard to their frequency, to the adjective stimuli in experiment 3. Words were not lemmatized, for the same reasons as mentioned in chapter 7, section 2.3.3.

In the initial BNC searches it was noted, early on, that many of the high frequency nouns that could be provided as stimulus words have *good* as a high frequency adjective. Because of the possibility that a respondant could provide *good* for all of the stimuli, only 4 nouns were chosen which had this adjective as a high frequency collocate according to the BNC data, and these are noted below in Table 8.1 (column 4). There is some reason to believe that in experiment 2 the NNSs fell back upon the generic noun *thing* in their associates, and it is quite possible that they could use *good* as a 'generic' adjective in this experiment, so, for this reason many eligible nouns were excluded from the list. In addition, checks were made on the other adjectives, in order to confirm that no one particular adjective could be used 'successfully' on many occasions. In addition to the details for *good*, the adjectives which could 'successfully' be provided as associates five or more times (i.e. they were in the top 20 collocates of the 20 nouns) are given below in Table 8.1 together with the nouns which they can describe.

Table 8.1. The stimulus words and the adjectives which are most commonly among the 20 most frequent collocates for these words.

	First	General	Good	Great	Important	Other	Second
A. Amount							
B. Approach	*	*					*
C. Basis							
D. Chance	*		*	*			*
E. Details					*	*	
F. Evidence			*			*	
G. Fact					*		
H. Future				*			
I. Importance				*			
J. Information		*					
K. Kind	*	*				*	*
L. Matter					*	*	
M. Moment	*						
N. Part	*				*		*
O. Problem	*			*		*	*
P. Purpose		*				*	
Q. Range							
R. Role					*		
S. Sense		*	*	*			
T. Word	*		*				

Because this task was also to be given to NNSs, as much as was possible, noun stimuli with a number of comparatively low frequency adjectives were excluded. So, for example, *agreement* was excluded on the grounds that among its most frequent 20 collocates are: *mutual, definitive, unanimous, tacit* and *bilateral. Attempt* was also excluded (because of the common adjective collocates *abortive* and *futile*) as was *effect* (collocates include *cumulative* and *detrimental*), *decision* (collocates include *unanimous*

and *informed*) and *mind* (collocates include *subconscious, conscious, unconscious, enquiring, suspicious* and *unsound*).

It was envisaged that a receptive multiple-choice task would also be given, at a later stage, to the respondants who took part in this experiment (i.e. the respondants would be asked to choose the most common collocation from a choice of three for each stimulus word). Therefore, a secondary check on the BNC ranks was made using Altavista. The first, tenth and twentieth most common collocations in the BNC were checked against the search engine data, to ensure that the same rank order was obtained. When there was a difference over the most frequent collocate the stimulus word was abandoned. This disqualified the words *event* and *knowledge* from the list. In three cases there was a difference over the orderings of the 10^{th} and 20^{th} most frequent words in the BNC with Altavista data (for the nouns *range, sense* and *problem*). In these cases the stimulus word was not abandoned, as this difference would not materially affect the outcome of the task, i.e. differentiating the most frequent collocation from the other two options.

Following on from the discussion in chapter 2, and in addition to the comments made above, it was decided the stimulus words should, as far as was possible, satisfy the criteria outlined below:

1. They should be well distributed throughout different text types in the BNC and in an American corpus.

This was not difficult to achieve: see appendix 8 for the details of the distributions of the words.

2. The words should not have high frequency collocating adjectives that are very different across written and spoken corpora (BNC complete versus the spoken sub-corpus)

Eleven of the nouns chosen as stimuli have exactly the same most frequent adjective collocate in the BNC (complete) and the spoken sub-corpus (the nouns were: *amount,*

basis, chance, fact, future, importance, information, kind, matter, role, sense). For six of the remaining nine words, the most frequent collocate in the BNC (complete) was among the three most frequent collocates according to the spoken sub-corpus data. The only words for which there was not a general concurrence on the most frequent collocates between the two sources of data were *approach, evidence* and *word.* These words were, however, retained because they satisfied the other criteria and it was not possible for *all* of the words to satisfy *all* of the criteria.

3. The stimulus words should not have high frequency collocating nouns that are very different between British and American corpora (BNC complete versus Brown).

Eight of the nouns have the same most frequent collocate in the BNC and Brown (*future, importance, matter, moment, range, role, sense, information*). For another 7 of the stimulus words, the most frequent collocate from the BNC (complete) is among the top 5 most frequent collocates in the Brown corpus (*amount, approach, evidence, kind, part, problem, word*). For the words *basis, chance, details, fact, purpose* the most frequent collocate either has no instances in the Brown corpus or falls outside the most frequent 5 collocates. Because of the data scarcity problems encountered in checking these very frequent collocations it was not felt to be problematic to include these items as stimulus words, particularly as they satisfied the other requirements noted above.

In chapter 2 it was noted that a corpus can provide skewed results about word and/or collocation frequency data. The BNC, like any other corpus, is vulnerable to this problem. Therefore, after the words were eventually chosen, whenever it was found that a single source was providing a very large number of instances of a certain collocate, the collocate was omitted from the frequency listing. The collocations which were omitted because of this consideration were: *following details, pertinent details, supporting evidence, dark matter, dry matter, good range, correct word* and *whole word.*

An additional amendment to the BNC data was to combine hyphenated and non-hyphenated instances (for many of the *range* words, and for *common sense, general purpose, special purpose, up to date information, long term future, not too distant future, well known fact, day to day basis, part time basis, one to one basis*). This also meant that the number of instances of other collocations had to be adjusted accordingly (e.g. the total for *known fact* was reduced, because *well known fact* instances were added to the *well-known fact* instances). It was not felt to be a problem that some of the resulting 'collocations' were, in fact, compound adjectives (e.g. *close* the adjective, being connected with *range* the noun, to produce *close-range* the compound adjective), because the respondants were only being asked to provide a high frequency collocating adjective: the resulting part of speech of the 'collocation' was not important. In one case, a spelling difference arose: The number of instances of *foreseeable* and *forseeable* were combined for the instances of *for(e)seeable future*. Finally, *criminal evidence* was omitted because of its predominant usage in the Police and Criminal Evidence Act 1984.

The list of stimulus words, and details of their collocates are provided below in Table 8.2. The 'range' column (column 4) indicates the most frequent and the 20th most frequent noun collocates (non-noun collocates were omitted from the list), together with the number of instances of the most frequent collocation and the number of instances of the 20th most frequent collocation in the BNC. Appendix 9 provides the full details of the collocating items of the stimuli words.

Table 8.2. Noun Word Stimuli word details for Experiment 4

Stimulus word (noun)	BNC instances	No. of different collocates (1 to the left)	Raw frequency figures for the most frequent collocate and 20th most frequent collocate according to the BNC.
A. Amount	12646	280	769 certain -- 50 increasing
B. Approach *	13668	995	219 new -- 41 whole
C. Basis *	14360	681	436 regular -- 39 full (-) time
D. Chance *	12392	232	343 good -- 22 main
E. Details	11521	232	742 further -- 19 written
F. Evidence *	21116	701	276 further -- 75 oral
G. Fact *	36560	241	226 actual -- 16 established
H. Future	11381	182	585 near -- 11 indefinite
I. Importance *	9573	241	317 great -- 53 practical
J. Information *	38326	1471	1131 further -- 76 available
K. Kind	21181	332	375 different -- 12 funny
L. Matter	15285	242	207 different -- 14 straightforward
M. Moment *	20814	269	270 last -- 23 opportune
N. Part *	48631	1152	1047 important -- 134 lower
O. Problem *	28559	795	401 major -- 60 growing
P. Purpose *	9154	247	183 general (-) -- 26 true
Q. Range	18432	656	2743 wide -- 45 huge
R. Role *	17993	683	625 important -- 53 supporting
S. Sense	20339	683	1125 common (-) -- 48 broadest
T. Word	18535	546	308 last -- 32 particular

* = Shell noun according to Schmid (2000, pp.381-402).

Compared to the adjective stimuli table (Table 7.2), it should be noted that the frequencies of the nouns are, on the whole, lower (average frequency of the adjectives =

31,931, average frequency of the nouns = 20,023). In addition, the number of collocates of the nouns is lower (average = 543) than for the adjective stimuli (average = 1,930). It could not be foreseen whether or not the actual frequency of the frequent stimuli words would have any effect upon which collocates would be provided. Though it may be that the smaller number of collocates aids respondants, this too is unclear. For example, results from Nelson & McEvoy (2000) do not give any support to the idea that more frequent words will have more (i.e. different) associations than less frequent words in free association tasks. They comment, "more common words...do not seem to have more or stronger connections to other words" (2000, p.517).

2.4. Method

Unlike experiment 2, where nouns (only) could be expected in response to adjective stimuli in the 'adjective slot' format, one of the problems in setting up this experiment was how to prevent non-adjective responses being provided in a 'slot noun' format. The BNC POS tagging system was used to identify the frequent collocates not classified as adjectives and a list was made of all of the non-legitimate high frequency collocates which the respondants might provide, but which were not permitted. The list was put on the test paper, preceding the task alongside instructions explaining that comparative, superlative and adjective phrase responses were acceptable responses. The test paper is provided in appendix 10, and the note to the respondants covering the points noted above, is provided below:

> **Note**: Please remember that your response should be an **adjective** (e.g. **old**). Your response can also be a comparative form of an adjective, e.g. **older**, or a superlative form e.g. **oldest**. Sometimes you may want to provide an **adjective phrase** if you think that it is very frequent, e.g. **'knock-on'** (as in 'knock-on effect') or **'half-hearted'** (as in 'half-hearted attempt'). **This is acceptable**. If you want to use an adjective / adjective phrase more than once you may do so.
>
> Please note that the following words are not permissible: **my, your** etc.; **this, that, any, another, one, more, every, much, such, few, all, little, own, some, enough, each, same.**

The 20 noun stimulus words were randomly organised in four different versions, all four containing the same words but in different orders. The reason for making different versions was an attempt to counter chaining effects between the stimulus words, and to minimise the effects of fatigue and/or loss of concentration of the subjects doing the experiment. The task specified that only one response be given by the respondants (for the reasons given in chapter 6, section 2.2.2).

Respondants were asked to complete the task quickly, in around 5 minutes. Native speakers were not supervised during the task, as experience in conducting the first productive experiment (i.e. experiment 2), indicated that they understood what was required of them. Non-native speakers were typically supervised: they needed more guidance about what to do, and additional explanation about what was required of them. The fact that not every conceivable response was acceptable meant that when an illegitimate response was provided, the NNS respondants were asked to provide another response. Care was taken to simply reiterate the instructions with the same examples, in explaining what was required.

2.5. Subjects

2.5.1. Native speakers (NSs)

None of the respondants who took part in this experiment had participated in experiments 2 and 3. The native speakers were male EFL lecturers at KFUPM, Saudi Arabia. All were qualified English language teachers and had considerable experience teaching English as a foreign language. All were native speakers, and as was the case for experiments 2 and 3, the majority were British, indeed, the number of Britons who did this task was slightly higher than in experiment 2 (14/20 compared to 10/20). The other nationalities who took part in the experiment were: Irish (3), American (2) and Australian (1). The teachers were approached in their office hours and the task was explained to them. The majority of the subjects completed the task at their convenience and posted me their responses.

2.5.2. Non-native speakers (NNSs)

None of the respondants who did this task were involved in experiments 2 and 3. The non-native speakers were all native Arabic speakers, teaching at KFUPM. All had lived in an English speaking country (predominantly USA) for a minimum of 4 years. All had earned PhDs from USA, Canada, UK or Australia. This meant that the education level of this group was higher than that of the NSs (of whom all have bachelors/masters degrees, but none have PhDs). The non-native speakers who took part in the experiment were from a variety of departments in the university, reflecting its engineering orientation, e.g. Chemical Engineering, Petroleum Engineering, etc. As with experiments 2 and 3, the majority of the subjects were Saudi, though fewer in number than those experiments (9 as opposed to 16). The other participants were: Egyptian (3), Palestinian (3), Algerian (2), Jordanian (2), and Lebanese (1). The majority of the respondants were supervised when doing the task (17 of the 20). Three respondants posted me their completed tasks, for practical reasons.

2.6. Results

2.6.1. Native speakers

Upon completion of the tasks, the respondants' answers were compared to the BNC collocation output data (for details, see appendix 9). The resulting sums of the ranks were subjected to two MWU analyses. The first involved the elimination of ties from the respondants' answers and the reduction of N accordingly, in the number of different BNC collocates used in the calculation. In the first analysis, then, the MWU analysis compared the set of *different* answers against the same number of BNC items. The second analysis retained the ties. For more background on the rationale for choosing this statistical procedure, the different focuses of the two analyses and the steps involved in calculating U, see chapter 7, section 2.6.1.1.

2.6.1.1. Analysis 1 – Excluding ties

Table 8.3 below, details the sum of the ranks and U for the 20 stimulus words.

Table 8.3. Native Speakers responses ranked against the BNC (N varies, see number in brackets, column 3)

Word	Group 1 – Sum of the ranks	Group 2 – Sum of the ranks (N)	U	Significance (1 tailed)
Amount	107.5	63.5 (9)	18.5	0.05
Approach	322	143 (15)	23	0.01
Basis	145	64.5 (10)	9.5	0.01
Chance	156.5	96.5 (11)	30.5	0.05
Details	267	139 (14)	34	0.01
Evidence	315	150 (15)	30	0.01
Fact	322.5	205.5 (16)	69.5	0.05
Future	32	23 (5)	8	Not sig.
Importance	92	44 (8)	8	0.01
Information	228.5	122.5 (13)	31.5	0.01
Kind	185.5	114.5 (12)	36.5	0.05
Matter	126	84 (10)	29	Not sig.
Moment	330.5	197.5 (16)	61.5	0.01
Part	130	80 (10)	25	0.05
Problem	133	77 (10)	22	0.05
Purpose	131.5	78.5 (10)	23.5	0.05
Range	34	21 (5)	6	Not sig.
Role	248.5	157.5 (14)	52.5	0.05
Sense	44.5	33.5 (6)	12.5	Not sig.
Word	189	111 (12)	33	0.05

According to the figures in Table 8.3 above, the null hypothesis holds for 4 of the 20 words: *future, matter, range, sense*. This meant that the respondants' ideas about the frequent collocates for these words and the BNC data were similar, and there is no statistical significance between the two sets of data.

For all the other 16 words the data from the two groups were significantly different. For *amount, chance, fact, kind, part, problem, role, word, purpose* this was so at the p=0.05 level, and for the remaining words *(approach, basis, details, evidence, importance, information, moment*) at the p=0.01 level.

2.6.1.2 Analysis 2 – Retaining ties

In this analysis tied associations were retained, i.e. if two or more of the respondants provided the same word, these responses were not discounted, as was the procedure in the first analysis. N is usually 20; however, occasionally it was the case that the native speakers produced responses which were not adjectives and were therefore rejected (for the words *details, importance*)[3]. When this was the case, N was reduced accordingly (as indicated in the third column).

[3] It was not realized, at the time of conducting the experiment, that these respondants provided illegitimate responses according to the test instructions.

Table 8.4. Native Speakers responses ranked against the BNC, admitting same responses

Significance levels (N=20, two tailed) p = 0.05=127, p = 0.01=105.

Word	Group 1 – Sum of the ranks	Group 2 – Sum of the ranks	U	Significance (2 tailed)
Amount*	367	453 (20)	157	Not sig
Approach	464.5	355.5 (20)	145.5	Not sig
Basis	578.5	241.5 (20)	31.5	0.01
Chance*	382.5	437.5 (20)	172.5	Not sig
Details	447.5	293.5 (19)	103.5	0.05
Evidence	495	325 (20)	115	0.05
Fact	456	364 (20)	154	Not sig
Future*	282.5	537.5 (20)	72.5	0.01§
Importance*	301.5	364.5 (18)	130.5	Not sig
Information	550.5	269.5 (20)	59.5	0.01
Kind	433.5	386.5 (20)	176.5	Not sig
Matter*	397.5	422.5 (20)	187.5	Not sig
Moment	490.5	329.5 (20)	119.5	0.05
Part	427	393 (20)	183	Not sig
Problem	418.5	401.5 (20)	191.5	Not sig
Purpose*	394	426 (20)	184	Not sig
Range*	276.5	543.5 (20)	66.5	0.01§
Role	463.5	356.5 (20)	146.5	Not sig
Sense*	281	539 (20)	71	0.01§
Word*	384.5	435.5 (20)	174.5	Not sig

Note * = Group 2 higher than group 1[4].

§ = Higher group 2 rank leads to significance.

[4] This is possible as the BNC has to give 20 different responses and this is being compared to the 20 responses of the teachers whether or not there are 20 *different* responses or a smaller number with multiple hits on the same word.

This analysis provides a very different picture from that of Table 8.3, and it does so because when the ties were retained for consideration in the calculation, they were usually high frequency collocates of the stimulus words. For the words *future, range* and *sense* the sum of the ranks of the NS teachers were higher than the BNC sum of ranks, to the extent that a significant difference is recorded, indicated by § in column 5. For 12 of the words, there is not a statistically significant difference. This means that for these 15 words, the NS data is either comparable to the other set of data, or the NSs are able to perform better than a notional norm represented by a reference group of fictitious idealised subjects who, between them, produce the top 20 collocates. For only five of the words the BNC data gives statistically different data to those of the NS teachers in this analysis, indicating that the teachers do not seem to be aware of the collocates of these words. These words are: *basis, details, evidence, information* and *moment*. Possible explanations for why these differences exist are forwarded in the discussion section below (section 2.7).

2.6.2 Non-native speakers

2.6.2.1. Analysis 1 – Excluding ties

Turning now to the NNSs, Table 8.5 below shows the relevant sums of the ranks and U for the NNS data in an analysis excluding multiple same responses.

Table 8.5. Non-native speakers responses ranked against the BNC (N varies; see number in brackets, column 3)

Word	Group 1 – Sum of the ranks	Group 2 – Sum of the ranks (N)	U	Significance (1 tailed)
Amount	116	55 (9)	10	0.01
Approach	177.5	75.5 (11)	9.5	0.01
Basis	301	105 (14)	0	0.01
Chance	98	73 (9)	28	Not sig
Details	244.5	106.5 (13)	15.5	0.01
Evidence	121.5	49.5 (9)	4.5	0.01
Fact	289	176 (15)	56	0.01
Future	35.5	19.5 (5)	4.5	Not sig
Importance	190.5	109.5 (12)	31.5	0.05
Information	264	142 (14)	37	0.01
Kind	241	110 (13)	19	0.01
Matter	261	145 (14)	40	0.01
Moment	254	152 (14)	47	0.01
Part	283.5	181.5 (15)	61.5	0.05
Problem	136.5	73.5 (10)	18.5	0.01
Purpose	228	123 (13)	32	0.01
Range	62	43 (7)	15	Not sig
Role	160.5	92.5 (11)	26.5	0.05
Sense	47	31 (6)	10	Not sig
Word	372.5	222.5 (17)	69.5	0.01

The null hypothesis holds for 4 of the words (*chance, future, range, sense*). This means that there is no statistical difference between the NNS data and the BNC for these words. It should be noted that three of these words are the same as the words for which there is no statistical difference in the NS data, Table 8.3 – *future, range* and *sense*. For the other

words there is a statistically significant difference at either p=0.01 or, for three of the words at p=0.05 (*part, role, importance*).

2.6.2.2. Analysis 2 – Retaining ties

Although every effort was made to ensure that the respondants were giving adjective responses, on five occasions N was reduced from 20 to 19, because of an oversight in the supervision, i.e. respondants provided words that were not adjectives. These responses were excluded from the resulting analysis, and this is indicated by N being 19 not 20 in column 3 for these words.

Table 8.6. Non-native speakers responses ranked against the BNC, admitting same responses.

Significance levels (N=20, two tailed) p = 0.05 = 127, p = 0.01 = 105

Word	Group 1 sum	Group 2 sum	U	Significance (2 tailed)
Amount	418.5	322.5 (19)	132.5	Not sig.
Approach	581	239 (20)	29	0.01
Basis	605	215 (20)	5	0.01
Chance*	316	504 (20)	106	0.05§
Details	499	242 (19)	52	0.01
Evidence	442.5	377.5 (20)	167.5	Not sig
Fact	524	296 (20)	86	0.01
Future*	302	517 (20)	92.5	0.01§
Importance	474.5	266.5 (19)	76.5	0.01
Information	516	304 (20)	94	0.01
Kind	528.5	291.5 (20)	81.5	0.01
Matter	462	358 (20)	148	0.01§
Moment	499	321 (20)	111	0.05
Part	426	394 (20)	184	Not sig
Problem	434	386 (20)	176	Not sig
Purpose	511.5	308.5 (20)	98.5	0.01
Range*	277	464 (19)	87	0.01§
Role	420	400 (20)	190	Not sig
Sense*	269	472 (19)	79	0.01§
Word	526	294	84	0.01

Note * = Group 2 sum of the ranks is higher than group 1[5].

§ = Higher group 2 rank leads to significance.

[5] This is possible as the BNC has to give 20 different responses and this is being compared to the 20 responses of the teachers whether or not there are 20 different responses or a smaller number with multiple hits on the same word. If the teachers are providing a number of the same high frequency adjectives then this raises the resulting sum rank.

For five of the words (including the four words mentioned in Table 8.5, for which there was no statistical difference between the BNC and the NNSs) the NNS sum of the ranks is higher than the BNC, in a statistically significant way: *chance, future, sense, range, matter*. For five other words (*amount, evidence, part, problem, role*) there is no statistical difference between the BNC data and the NNSs.

The 10 words for which the two sets of data remain statistically significant, with the BNC sum rank being higher, are: *approach, basis, details, fact, importance, information, kind, moment, purpose,* and *word*. There is a degree of overlap with the NS data (Table 8.4), in that of the 10 words noted here, there are four for which there was a significant difference in the NS group, analysis 2: *basis, details, information, moment*. The only word for which there was a statistically significant difference between the BNC and the NSs but *not* for the BNC and NNSs in the second analysis was *evidence*. There is no evidence that the difference between the NS data and the BNC can be explained because of differences between the BNC (complete corpus) and the spoken sub-corpus or differences with the Brown corpus, and so these words are subjected to a more careful analysis below[6].

2.6.3. Comparison of native speaker and non-native speaker responses

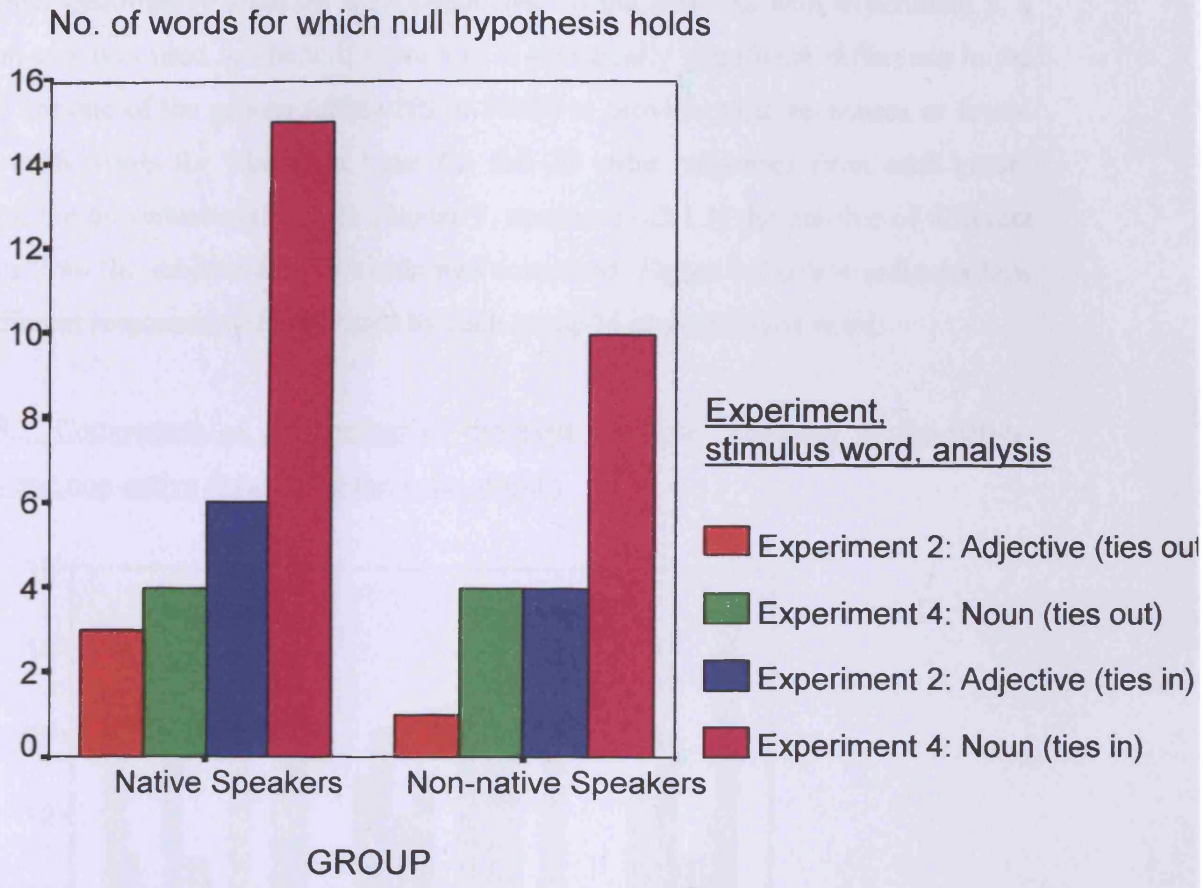
For most of the words in the first analysis (for both the NS and NNS groups) the null hypothesis can be rejected - there is a significant difference between the two sets of data. However, for the second analysis, the null hypothesis holds for the majority of the time for the NS group (15/20) and for half of the time for the NNS group (10/20). There is evidence here that NSs and, to a lesser extent, NNSs are aware of some frequent adjective collocates of high frequency noun stimuli, when the analysis admits the inclusion of multiple same responses. Clearly though, this latter result does not give warrant to the

[6] It should be noted that of the words which NSs had problems with, for two (*basis* and *details*) there is a difference between the BNC and Brown data. There are very few instances in Brown for collocations containing either of these words, and the respondants' associations are not more similar to the Brown data. In Brown, the most common collocates for *basis* are: *part-time (3), random (3), general (2), reasonable (2), regional (2), scientific (2), contingent-free (2),* and *statistical (2)*. For *details* the only adjective collocate is *everyday*, with two instances in the corpus. As noted in section 2.3, for *evidence* there is a difference between the BNC total corpus and the spoken sub-corpus, but the respondants' data do not fare any better against the sub-corpus.

idea that respondants can provide comparable data to the BNC, rather the results from the second analysis indicate that *some* high frequency collocates are known by a reasonable number of the respondants. For example, to the stimulus word *problem*, the NSs provided the following frequent collocates (number of responses in brackets): *big* (3), *serious* (4), *difficult* (5) and *common* (2). To the same word the NNSs provided the following (number of responses in brackets): *big* (4) *serious* (3) and *difficult* (6). It should be noted though, that neither of the groups was able to provide a set of responses comparable to the BNC data for the stimulus word *problem*, when the tied responses were excluded from the analysis (see Tables 8.3 and 8.5).

It would seem that the provision of noun stimuli (as opposed to adjective stimuli) has a generally beneficial/positive effect on the 'quality' of the respondants' associations in terms of the greater similarity to the BNC data, and this is so for both the NSs and NNSs. The graph below indicates the differences, numerically, where the null hypothesis was found to hold between the BNC data and the NS/NNS respondant data according to the 2 different analyses in experiments 2 and 4.

Figure 8.1. Comparison of BNC / respondant non-significant cases (from 20) for the two different analyses in the two productive experiments[7].



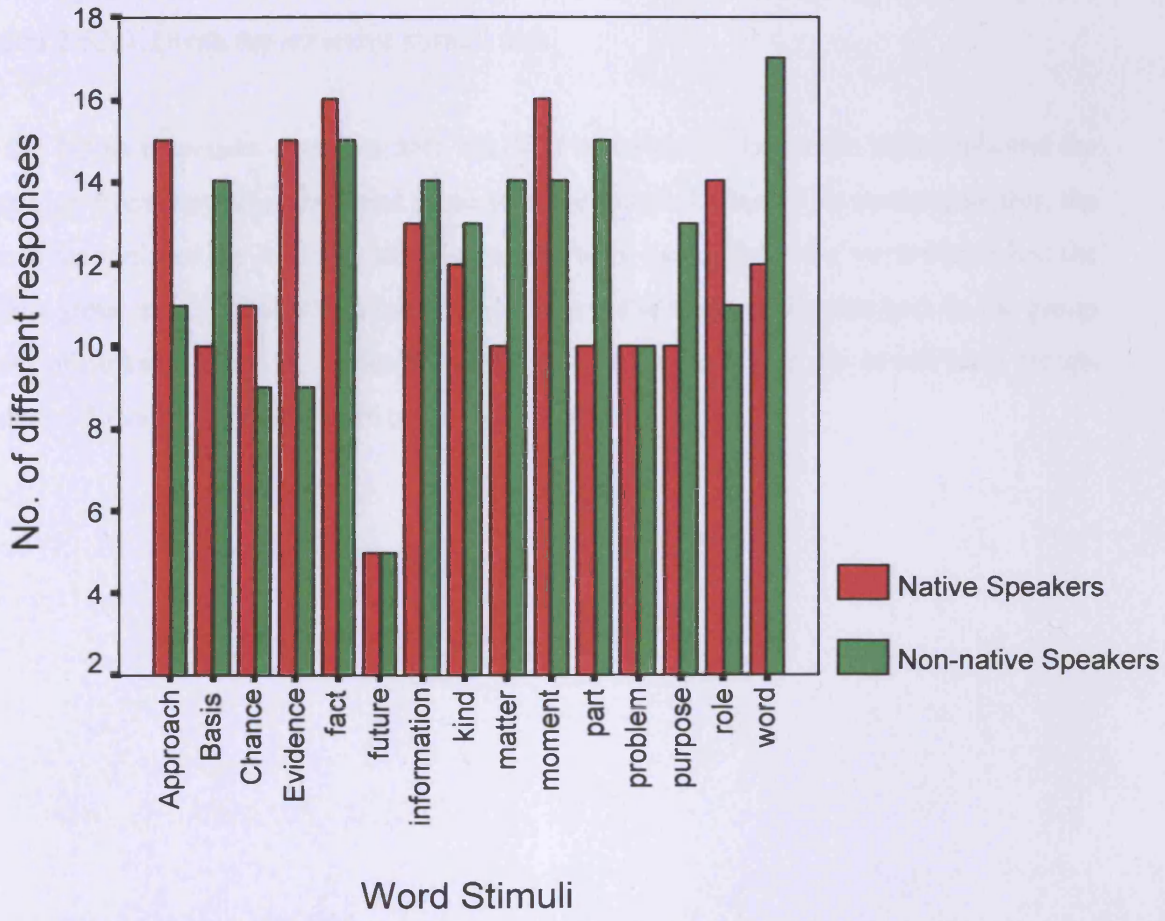
The data from Figure 8.1 suggest that it is easier to think of a high frequency adjective for a noun stimulus word (signified by green bars in the figure) than a typical noun collocate where the stimulus word is an adjective (the red bars in the figure). These data also indicate that while the respondants (when considered as a group) may not be aware of all, or most of the most frequent collocates, they do have knowledge of *some* very frequent collocates, i.e. the ties in data (signified by the blue and magenta bars) indicate that a number of the respondants are providing the same high frequency collocates. As can be seen, the NSs provided associations closer to the BNC data than the NNSs in 3 of the 4

[7] Or where the NS or NNS sum of ranks scored significantly higher than the BNC.

analyses, indicating that their lexical intuitions approximate more closely to the BNC data than do the advanced non-native speakers'.

Two further quantitative analyses were conducted on the data. As with experiment 2, a Wilcoxon test was used to check if there was a statistically significant difference in the tendency for one of the groups (either NS or NNS) to provide more responses or fewer. There are 15 words for which we have the full 20 valid responses from each group. Following the procedure outlined in chapter 7, section 2.6.3.1.1, the number of different responses from the subjects for the words was compared. Figure 8.2 below indicates how many different responses were provided by each group to each stimulus word.

Figure 8.2. Comparison of the number of different responses provided by the native-speakers and non-native speakers to the noun stimuli



NNS>NS = 6

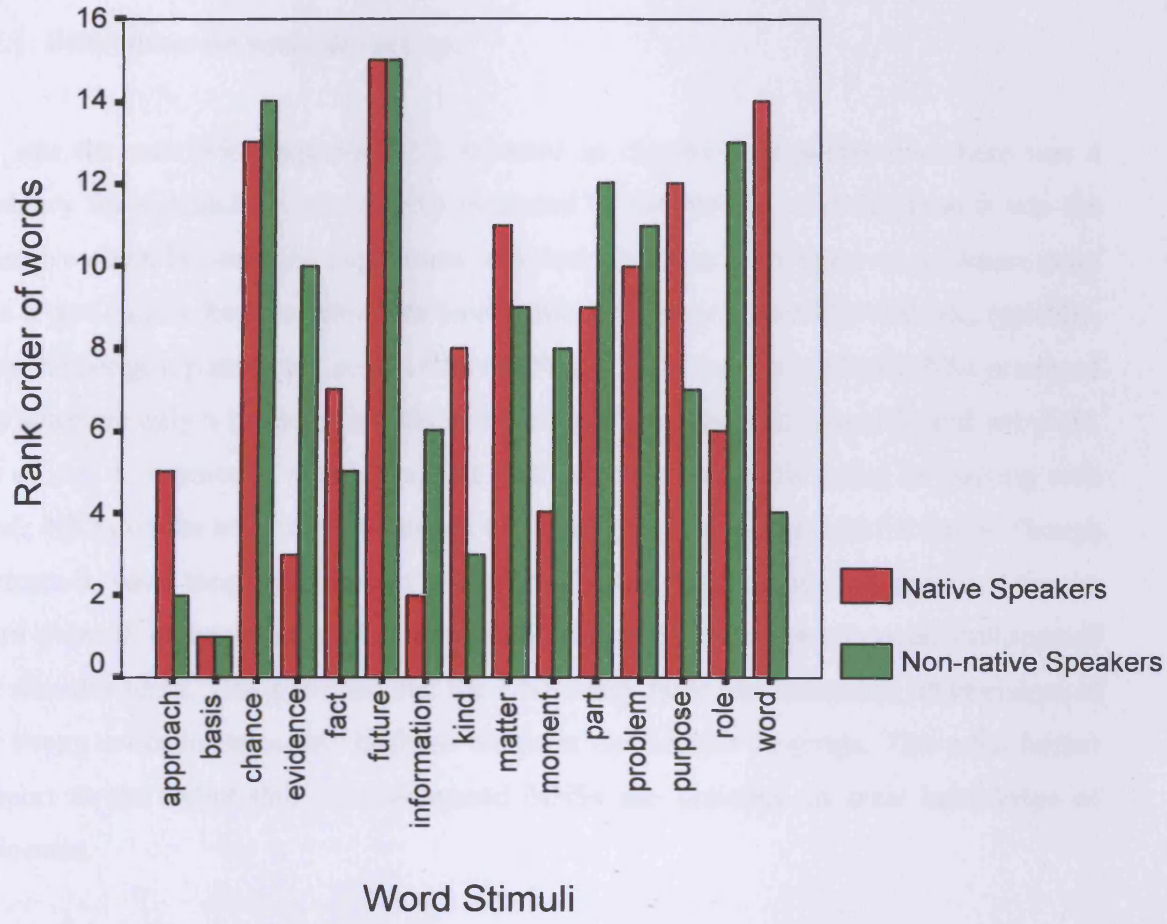
NNS<NS = 7

NS=NNS = 2

The data in the graph indicate that on 6 occasions the NNSs provided more responses than the NSs (e.g. to the word *word* they provided 17 responses and the NSs provided 12 different responses). On another 7 occasions they provided fewer responses (e.g. to the word *evidence*, for which they provided 9 different responses and the NSs provided 15). On two occasions the same number of responses was provided by each group (for the stimulus words *future* and *problem*). Using the Wilcoxon test reveals that the differences are not statistically significant: z=-.421, N-ties=15, p=0.624, two tailed (not significant). Therefore, it is not the case that the non-native speakers produced consistently more or fewer responses than the NSs. This finding is similar to the finding reported in chapter 7, section 2.6.3.1.1 with the adjective stimuli task.

Do the NNSs responses correlate with the NSs in terms of the words which elicited the most high frequency adjectives and those which elicited the least? To investigate this, the sum of the ranks of the different stimulus words were listed, from the word which had the highest group sum to that which had the lowest, on the basis of the numbers in the group 2 sum of ranks columns in Tables 8.4 and 8.6. Only those words for which both groups produced 20 valid responses were ranked (15 in total).

Figure 8.2. Group sum ranks of NS and NNS for word stimuli compared



A Spearman Rho correlation test indicates that the difference was not significant (Rho =.45). Clearly, there is a mild correlation, but this is not significant, and, as such, this finding differs from the data from Figure 7.2 (chapter 7, section 2.6.3.1.2), showing the results from the comparable analysis with the adjective stimuli responses. There are some similarities in Figure 8.2 between the two groups; for example, they had similar 'success' with *future* and *chance*, and similar 'problems' with *basis,* vis á vis the similarity of the responses to the BNC data. However, unlike the findings from the Spearman Rho calculation in experiment 2, there are noticeable differences too, as can be seen in the above figure and the quite different relative rank orders of, for example, *role* and *evidence* between the groups. It is worthwhile considering why it may have been that one group's responses were, relatively speaking, not as good as the other's for these words, and this is one of the subjects discussed below

2.7. Discussion

2.7.1. Differences between the groups

As was the case with experiment 2, reported in chapter 7, it seems that there was a tendency for a default associate to be produced by the NNSs – and this time it was the adjective *good*. Because the experiment was designed to exclude many cases where *good* was a particularly frequent collocate (see Table 8.1 above), the NNSs did not, typically, score higher group sum rank scores than the NSs because of this tendency. NSs produced this adjective only 6 times, as associates to: *chance* (1), *kind* (1), *sense* (3), and *word* (1). On all but 1 occasion it was a frequent collocate (the exception being its pairing with *kind*). NNSs on the other hand produced the word *good* as an associate 38 times. Though at times it was a frequent collocate (11 of the 38 *good* responses were provided for the word *chance)*, in the majority of cases when it was used, it was not a frequent collocate of the stimulus word. This indicates that the NNSs may have less conscious appreciation of the strong collocate associates of these words in the English language. This adds further support to the belief that even advanced NNSs are deficient in their knowledge of collocates.

At times the NSs seemed to draw on a wider range of semantically related possible adjective combinations to the stimulus word, whereas for the NNSs there was a tendency to provide the same word. This is seen most clearly in the responses to *role*. Whereas the NSs produced the frequent collocations: *important* (1), *major* (1), *key* (2), *leading* (3), *significant* (2) and *prominent* (1), eight of the NNSs produced *important* (the most frequent collocate) with only 2 other semantically related frequent responses provided (*major* and *vital*). This suggests (in line with Granger 1998) that there may be a tendency for NNSs to overuse certain combinations, and to be less sensitive to the alternative ways of framing a particular idea.

2.7.2. BNC – NS differences

In the ties-in analysis with native speakers (reported in section 2.6.1.2), it was found that there was a significant difference between the BNC data and the NS data for only five words (*basis, details, evidence, information* and *moment*). In this section I look at whether the various explanations forwarded to explain differences between corpus data and elicited data can account for these differences, starting with the possible explanation that the associations differed from the BNC data because the most frequent collocates of these words typically occur in larger chains.

For two of the above mentioned five stimulus words, an accessibility/availability argument can be forwarded to explain why frequent collocates were not provided - for *basis* and *moment*. A large number of the frequent collocates for *basis* are time related (*regular, daily, day-to-day, part-time, permanent, annual, temporary, weekly, monthly, full-time*). Very importantly, these occupy all of the first 7 most frequent collocate slots of this word. Only 2 NSs produced any of these adjectives: *regular* (2) and there was also 1 less frequent response in the same semantic class - *frequent*. All of the above noted time-related adjectives show a very strong tendency to occur in *on a(n) ADJ basis* framework. The topical classifying adjectives (e.g. *individual, regional, national, commercial*) also tend to occur in this chain. The dominant responses were, however, evaluative – *sound* (5), *firm* (4), and *solid* (3) - clearly all belonging to the same semantic field, and evaluating the strength of the *basis*, where *basis* means 'the facts, ideas, or things from which something can be developed' (LDOCE), rather than the meaning of *basis* in the above-noted chain, where it means the way or method of doing something. The adjectives produced by the NS respondants do not occur in the above noted collocation framework. It can be argued, then, that *basis* in the chain *on a(n) ADJ basis* is not very salient – it is locked up in an adverbial clause and the chain is stored holistically. An alternative explanation for these responses is that the denotational meaning of *basis* affected the subjects' responses, i.e. that *basis* as a lone word means *facts, ideas*, and the respondants provided adjectives which evaluate the reliability of those *facts, ideas*. Clearly the adjective is the key word in this chain: *on an annual basis* means *annually*, *on*

a regular basis means *regularly* – evidence that the adjective here is the 'key word' in the chain, rather than the noun.

Regarding *moment*, there is an interesting split in the respondant data. Some of the NS respondants provided an associate indicating the *timing* aspect of the meaning e.g. *last* (2), *right* (2), *long* (1), *brief* (1), and these are among the most frequent collocates. However, it seems that a particular positive semantic prosody is sometimes given to this word by the respondants in their associations, as seen in the responses: *magic, big, tender, golden, special, defining*[8]; there are no such positive prosodic instances among the word's most frequent collocates in the BNC, but there are frequent collocates which are negative (e.g. *worst*). The denotational meaning of *moment* is a time period, and so it is perhaps surprising that more respondants did not provide time-related adjectives. It is here that an accessibility restriction might be argued. According to the BNC, some of the very frequent time collocates typically occur in adverbial clauses e.g. *last moment* (in *at the last moment* in 71% of its occurrences in the BNC), *long moment* (*for a long moment,* 79%), and as a consequence it may be that the individual components making up these frameworks are not so salient, having become separate from the component works which make up the chain. It may be the case that respondants provided adjectives with the positive prosody as these collocates were more salient/available, though not so frequent.

Explanations for why the associations to *details, evidence* and *information* are not, typically, frequent collocates are more difficult to provide. There is one striking similarity for these words: they all have *further* as their most frequent collocate, and this word was not provided once by any of the respondants (NS or NNS). *Further,* though an adjective, has the same meaning in this chain as *more*, which is classified as a determiner in these combinations. It may be that the respondants discounted it as a valid answer for this reason. In addition, it is important to note that the adjective *further* is not in its 'prototypical' comparative form in such combinations, i.e. it is not related to *far*. Because of these factors *further*, as a potential collocate, may well have been 'overlooked'. We should probably not read too much into its non-production – though see the comments

[8] NNS data also had this as the dominant prosody: *happy* (3), *lovely, intimate, nice, meaningful, precious.*

about this in chapter 9, section 2.2.3. While the word does occur in bare collocations within certain chains/frames, e.g. *for further details* (43%), *for further information* (47%), this is not a particularly exclusive embedding and probably not invariable enough to posit that its existence in these chains hides it from searches.

For *details* there seems to be a semantic relationship between some of the non-frequent responses and the most frequent collocates. For example, the native speakers provided 'details are small' responses: *tiny* (1), *insignificant* (1) and *intricate* (1). None of these actual pairings are particularly frequent, but both *small* and *minor* are among the top 20 collocates.

For *information*, five NS respondants believed *important* to be a highly frequent collocate. This collocation is not a particularly frequent one, either in the BNC, the BNC spoken sub-corpus, or on the Internet[9]. Interestingly here, the respondants' answers do not have the same function as the majority of the frequent adjective collocates. In addition to the 'information is important' responses, some of the NS responses give the word an 'information is valid' meaning (e.g. *right, true, correct*). These two types of response (information is important/valid) differ from some of the BNC most frequent adjective partners, which seem to classify the noun (e.g. *new, detailed, additional, confidential, available* etc.) rather than evaluate it.

Evidence is one of the few words for which the NNS intuitions were actually better than the NS intuitions, according to the ties-in analysis. It seems quite clear what is going on in the native speaker responses. The word *evidence* can, broadly speaking, have legal, scientific or evaluative adjectives attached to it. Clearly, the salient meaning for the native speakers was its legal, as opposed to scientific meaning, as evidenced in the responses: *conclusive, circumstantial, corroborating, admissible, false, fresh* (2 of these are in fact in the top 20 for frequency). For the NNSs, one might expect, given their hard science training, that the frequent 'scientific' adjectives would have been produced e.g.

[9] According to Altavista searches, *important information* had around one quarter of the number of hits as *further information*, and one third of the hits as *general information.*

scientific, experimental, empirical. However this was not the case: it seems that the *type* of evidence was taken for granted, but its *validity* is the key factor in influencing the respondants. *Clear, strong* and *hard* between them account for nearly three quarters of the NNS responses.

Summing up, it has been argued that embedding can be called upon to explain why the responses to *basis* and *moment* were different to corpus data. However, it would seem that this explanation is *not* able to account for some of the other observations above. In particular, it seems that, on occasions, the NS respondants attached a dominant prosody to a word (e.g. *moment* is positive) not present in BNC data. Further, it appears to be the case that the respondants attached a particularly salient meaning to a word and its collocates (e.g. *evidence* is legal) not reflected in the BNC data. An additional factor mentioned here is that some adjectives are more 'adjective-like' and prototypical than others. *Further* may not have been produced as a collocate for this reason. These are very important observations, as they suggest that accessibility restrictions are only one factor that may account for differences between the corpus data and the elicited data in this experiment.

2.7.3. The dominant responses

As with the analysis of experiment 2 in chapter 7, it makes sound psycholinguistic sense to focus on the dominant responses, i.e. the items produced, rather than the items not produced, or produced only once or twice, in helping us to understand how collocations may be represented in the mind.

In Table 8.7 below, the dominant responses (i.e. responses provided by at least 15% of the respondants, NS and NNS data combined) are given. Although *N* (i.e. the number of responses) is usually 40, when it is not, this is noted in the first column. In the fourth column the frequency rank of the collocation according to the BNC data is provided, and in column 5 the collocation is classified as restricted or not, according to whether it is present in Benson et al. (1986).

Table 8.7. Dominant responses, combining NS and NNS data.

Stimulus word	Dominant response (%age of respondants)	Breakdown (No. of respondants providing the response) NS	NNS	BNC Collocation Rank of dominant response	Restricted
Amount (39)	Large (44%)	11	6	3	Y
Approach	New (15%)	6	-	1	N
Basis	Sound (15%)	5	1	15.5	Y
Chance	Good (30%)	1	11	1	Y
	Last (22.5%)	7	2	3	Y
Details (38)	-	-	-	-	-
Evidence	Clear (27.5%)	4	7	5	N
	Strong (17.5%)	2	5	11	Y
Fact	-	-	-	-	-
Future	Near (62.5%)	12	13	1	Y
	Bright (22.5%)	5	4	7	Y
Importance (37)	Great (27%)	8	2	1	Y
Information	Important (20%)	5	3	>20	N
	Useful (15%)	4	2	6	N
Kind	-	-	-	-	-
Matter	Important (27.5%)	6	5	5	Y
	Serious (15%)	4	2	4	Y

Stimulus word	Dominant response (%age of respondants)	Breakdown (No. of respondants providing the response) NS	NNS	BNC Collocation Rank of dominant response	Restricted
Moment	-	-	-	-	-
Part	Small (20%)	5	3	10	Y
Problem	Big (17.5%)	3	4	7	N
	Serious (17.5%)	4	3	5	Y
	Difficult (27.5%)	5	6	11.5	Y
Purpose	Main (17.5%)	5	2	2	N
Range (39)	Wide (56%)	13	9	1	Y
	Long (18%)	3	4	4	Y
Role	Important (22.5%)	1	8	1	N
Sense (39)	Common (67%)	13	13	1	Y
Word	Right (15%)	5	1	3	N

There are dominant responses (a response provided by at least 15% of the respondants) for 16 of the 20 stimulus words. On several occasions one particular stimulus word elicited more than 1 response with 15% attestation, so there were, in fact, 24 dominant responses. Only 1 of these combinations is not frequent according to the BNC - *important information,* mentioned earlier in section 2.7.2. Of the 24 dominant responses, 7 are the most frequent according to the BNC data (*new approach, good chance, near future, great importance, wide range, important role, common sense*), and 8 of the 24 dominant collocations are not included in Benson et al. (1986), indicating that we should view them as free collocations: *new approach, clear evidence, important information, useful information, big problem, main purpose, important role* and *right word.* In what follows, these responses are discussed with reference to the different theories forwarded to explain

the differences between corpus data and elicited data. Appendix 11 lists the complete record of responses obtained.

2.7.3.1. Embedding and choices

There are three instances among the dominant responses where the resulting collocation is usually the supporting ADJ NOUN collocation in *NP of NP* collocational frameworks, i.e. *large amount*, *small part* and *wide range*. The first two cases are particularly interesting as the dominant response adjectives were provided as stimulus words in experiment 2. The data showed that one NS produced *part* as an associate to *small*, and no NNSs provided this word in experiment 2. However, in the reverse procedure of the experiment, *part* elicited *small* 8 times (5 NS and 3 NNS)[10]. Two native speakers produced *amount*, when presented with the stimulus word *large* in experiment 2, and 17 respondants (11 NS and 6 NNS) produced *large* as an associate to *amount* in experiment 4. It is not the case that the more frequent word elicited the less frequent (*part* is more frequent than *small*, but *large* is more frequent than *amount*), but it should be noted that both *large* and *small* have many more collocates in the BNC than *amount* or *part*[11]. This finding, that the elicitation tendency is uni-directional, is a very interesting one. It does not sit comfortably alongside the notion that the dyads are stored holistically with the consequence that the individual components have weaker links to the words (as argued by Nordquist 2004). If this were the case, then it should not matter which of the words was provided as a stimulus word. Clearly though, there is a stronger link in the minds of the respondants from the noun to the adjective-noun collocation (framework) than there is from the adjective to the adjective-noun collocation (framework) for these words.

The other case where the dominant response is an adjective typically occurring in 'ADJ NOUN of NP' collocational chains, is the response *wide* to the stimulus word *range*. *Wide range* occurs in *wide range of NP* 97% of its occurrences in the BNC. The noun

[10] Interestingly, in free word association, the primary response for *part* is *of* (in the EAT), clearly indicating an awareness of its supporting role in *NP of NP* structures.
[11] According to the BNC *amount* has 280 collocates and *part* has 1152 collocates. *Large* has 2482 collocates and *small* has 3119 collocates, though it should be remembered that this includes all types of collocates.

range, together with the nouns *amount* and *part* are interesting in that they are nouns rarely used except as the supporting nouns in *NP of NP* chains, i.e. they do not really have an independent existence, occurring as the first noun in *NP of NP* around 70% of their occurrences according to the BNC. As such, we can argue that respondants are 'forced' to search the supporting NP in *NP of NP* chains when considering the typical partners of these words. As a consequence, such cases must be distinguished from other cases where there is a *genuine* alternative search option open to respondants (see the comments made earlier about *basis* and *moment*, where there are frequent non-collocational framework adjective alternatives available to the respondants).

The observation just made, that the presence or absence of non-chain choices may affect the quality of our lexical intuitions is of crucial importance. Though not typically forwarded as a reason for the mismatch between elicited data and corpus-data, such an explanation seems able to account for some of the data from this experiment.

NP of NP chains are only one type of chain where 'bare' adjective-noun collocations may be embedded. Two of the dominant adjective responses provided occur with the nouns in bare collocations typically embedded in other chains/frames: *great*, provided as a dominant collocate to *importance,* and *near* to *future*. *Great importance* typically occurs in the chain *of great importance* (e.g. *the book club was of great importance*) in the BNC, and *near future* typically occurs in *in the near future*.

In the BNC, *importance* is the fifth most frequent noun collocate of *great*, but in experiment 2, no respondants provided *importance* as a collocate to the word *great.* Like the cases of *large amount* and *small part* (mentioned above), the noun (*importance* in this case) more readily elicited the adjective, than the adjective (*great*) elicited the noun (*importance*). The majority of the most frequent adjective collocates of *importance* occur in the chain *of ADJ importance*, as indicated by the number in brackets: *particular* (79%), *crucial* (83%), *paramount* (90%), *considerable* (73%), *vital* (73%), *central* (71%), *prime* (92%), *fundamental* (77%), *major* (86%). For the adjectives *greatest* and *utmost* the dominant pattern is '*of the ADJ importance*': *greatest* (60%), *utmost* (81%). Again, like

the cases with the supporting noun in *NP of NP* chains, it can be argued that the respondants are forced to access *of ADJ importance* chains in their searches, because adjective collocates so rarely occur outside this framework. As such, like the cases of *range, amount* and *part*, the dominant responses to *great* should be distinguished from cases where there are 'genuine' non-framework options available for the respondants to produce as typical collocates.

The dominant responses to *future* are interesting. Unlike the cases above, there is a genuine non-collocational framework adjective option here, as well as a frame option. The most frequent 'time' adjectives of *future* typically occur in *in the ADJ future* or *for the ADJ future* frames (e.g. *near, not-too-distant, immediate*). The dominant response was *near*, and in the BNC this adjective shows strong embedding in the pattern *in the near future* (88%). However, there are quite a number of non-time adjectives which do not appear in this pattern, and yet which are also frequent collocates (e.g. *political, bright, uncertain*). The second most frequent response, after *near,* was *bright*. It would seem then, that there were two dominant access routes available to the respondants – the framework route and the non framework option. According to the 'choices' explanation forwarded above, more of the respondants *should have* provided the non-framework option – *bright*. However, over twice as many respondants provided *near* as produced *bright*, and this does challenge the 'choices' explanation to a certain extent: access to the framework seems better than one would expect, given the alternatives available without breaking into a framework.

2.7.3.2. Denotational meaning and polysemes

The possibility that the denotational meaning of the word 'negatively' affected the respondants in providing their dominant responses (vis á vis the BNC data) does not seem to be a particularly convincing explanation for explaining why the responses differed from BNC data[12]. For example, the dominant response *common* to *sense*, does not give

[12] Though this is able to explain the responses to *basis*, as mentioned previously, i.e. the dominant responses give the word its 'facts/ideas' meaning – the stand alone meaning.

the word *sense* its 'feeling', or 'five senses' meaning, which word association data suggests is the most salient meaning of this word (the dominant response in Moss & Older (1996) is *smell*). Like the results from experiment 2, it seems that respondants can 'bypass' the denotational meaning in looking for common adjective partners.

There is little evidence that the respondants had the wrong polyseme in mind when providing their frequent adjective collocates. That is, the respondants, did not, generally, provide collocates of the stimulus word which gave it a different meaning to its meaning in the most frequent collocations. For example, *matter* was not, typically, given its 'substance' meaning by the respondants, rather, the majority of the responses gave it its 'subject' meaning. This suggests that respondants are, generally, aware of the most frequent meaning of a word (see discussion on this in chapter 1) in providing their responses.

Like the case with the frequent collocates of the adjectives in experiment 2, there are no frequent idiom partners for the words in this experiment. While a few of the resulting combinations were idioms (e.g. *fat chance, sixth sense*), there is not enough evidence in this experiment to shed further light on the matter of idiom representation as discussed in chapter 6. As with the dominant responses in experiment 2, there are a couple of cases in the data where the dominant response creates a combination which is a compound adjective (e.g. *long-range, common-sense*). While *long-range* is more frequent than *long range* in the BNC, *common sense*, is much more frequent than *common-sense* [13].

2.8. Summing Up

This experiment has provided some interesting data. In particular, it has been noted that the supporting nouns that were not typically elicited in experiment 2, though they were frequent, *did* elicit their collocating adjectives in this experiment. These nouns do not typically occur outside the supporting noun role in *NP of NP* frames. When they were

[13] BNC data indicates the following: *long-range* (214), *long range* (86); *common-sense* (155), *common sense* (970). Also note that *general-purpose* (65) is less frequent than *general purpose* (118).

provided as stimulus words, respondants did seem able to access the framework and provide typical adjectives. It has been argued that this should not necessarily be interpreted as counterevidence to the embedding argument as forwarded in chapter 7. It can be argued that because respondants are forced to provide an associate there is no other option but to access the frameworks, and this reason can also explain the responses to *importance* and *future*. However, when there is a 'genuine' search option available to the respondants, it seems as if embedded collocates are less visible than they 'ought' to be; i.e. there is some kind of bias operating, which we can assume to be connected with the use of an availability heuristic in the search procedure. This is seen particularly clearly in the responses to *basis* and *moment*, and a little less so in the responses to *future*.

However, it has been noted that an appeal to embedding *alone* cannot explain all of the data from this experiment. Other factors do seem to play a role in affecting the responses. In particular, it seems that semantic prosody may play a role in affecting the responses (e.g. the responses to *moment*) or that an environmentally determined saliency might affect the responses (e.g. the responses to *evidence*). The role of the denotational meaning of the word in affecting the responses does not seem so significant: it may be that this affects the ability to define a word, but not the ability to produce frequent collocates. Very few differences have been noted between the NSs and the NNSs: there appears to be little support for the idea that there are differences in the mental representation of the words and their collocates, between these two groups of subjects

3. Experiment 5

A second experiment was conducted with the same respondants who participated in experiment 4 to see whether they were able to recognize the most frequent collocate of the stimulus words, when provided with a choice of three collocate options. This is the same procedure as was adopted investigating the associations to adjective stimuli. The reason for conducting this experiment was to investigate whether the ability to *recognize*, rather than *produce* a frequent collocate was affected in any way by accessibility restrictions.

3.1. Hypothesis

Based on the assumption that respondants automatically encode frequency information, and that the corpus used is representative, it was hypothesized that respondants would recognize the most frequent collocation for each of the stimulus words. It is less likely, on the basis of the findings from experiment 4, that the embedded collocations will be considered to be less frequent than they are according to corpus data. As such then, it was hypothesized that the data resulting from this experiment would be closer to the BNC data than were the data from experiment 3.

3.2. Method

In this task, the same respondants who took part in experiment 4, reported above, were presented with three collocations of the same 20 stimulus words. The three collocations provided as alternative choices for each word were the highest frequency, a medium frequency and a relatively lower frequency adjective collocate (1, 10, and 20) of the word, according to the BNC. See chapter 7, section 3.2 for an account of the issues involved in selecting suitable collocates. The subjects were asked to tick the box of the most frequent collocate. Four different versions of the test were made in an attempt to combat fatigue, chaining etc. No time limit was given, but all respondants completed the task in a shorter time than required to complete experiment 4. The test paper is provided in appendix 12.

3.3. Subjects

The same subjects (both NS and NNS) who took part in experiment 4, took part in this experiment. The time between this experiment and experiment 4 varied from a week or so to a couple of months. The native speakers were not supervised and half of the NNS subjects were supervised.

3.4. Results

3.4.1. Native speakers

In Table 8.8 below, the BNC collocate ranks for the stimulus words are listed from highest to lowest order. For example, *certain amount* is more frequent than *maximum amount* which is more frequent than *increasing amount* (ranks 1, 10, and 20 respectively)[14]. Next to each collocate the number of votes given by the NSs who were attempting to identify the most frequent collocate is given. So, of the 20 respondants who took part in this experiment, 15 believed that '*certain amount*' was the most frequent, 1 believed that '*maximum amount*' was the most frequent and 3 believed that '*increasing amount*' was the most frequent.

[14] Due to an oversight in the preparation of the task, there were 4 occasions when the choices given were *not* the first, tenth and twentieth: for *basis* rank 21 instead of rank 20 was provided for *problem* and *purpose* rank 11 instead of 10 was provided, and for *word* rank 19 instead of 20 was put in the list.

Table 8.8. No. of respondants choosing the different collocates in each word set - NS group (N = 20).

	A. Amount	B. Approach	C. Basis	D. Chance	E. Details
BNC	Certain 16	New 11	Regular 18	Good 19	Further 15
rank	Maximum 1	Traditional 9	Regional 1	Greater 0	Financial 3
	Increasing 3	Whole 0	Casual 1	Main 1	Written 2

	F. Evidence	G. Fact	H. Future	I. Importance	J. Information
BNC	Further 14	Actual 9	Near 16	Great 7	Further 13
rank	Available 6	Well-known 7	Better 4	Utmost 13	General 4
	Oral 0	Established 4	Indefinite 0	Practical 0	Valuable 3

	K. Kind	L. Matter	M. Moment	N. Part	O. Problem
BNC	Different 18	Different 10	Last 8	Important 14	Major 14
rank	Second 0	Small 4	Given 7	Small 5	Further 2
	Funny 2	Straightforward 6	Opportune 5	Lower 1	Growing 4

	P. Purpose	Q. Range	R. Role	S. Sense	T. Word
BNC	General 10	Wide 19	Important 3	Common 17	Last 15
rank	Special 6	Limited 1	Significant 14	Broad 1	New 3
	True 4	Huge 0	Supporting 3	Broadest 2	Particular 2

According to the above data, on only two occasions did the majority of NSs choose a collocate as the most frequent, which was at variance with the BNC data: for the words *role* and *importance*. Reasons for these differences with the BNC data are suggested in section 3.5, below. To see whether the native speakers were, as a group, significantly more likely to choose the most frequent collocate than either of the other two on offer, a chi-squared goodness of fit test was employed.

Table 8.9. Native speaker analysis - by which option recieved the greatest no. of votes

	Observed 1	Observed 2 and 3
BNC Rank 1	18	2

According to the data from Table 8.9, χ^2 =28.88. Incorporating Yates' correction for continuity, χ^2 =26.38, df=1, highly significant (p=0.01 level of significance = 6.64).

A second analysis was also conducted, this time taking into consideration the number of votes which each response received, as recorded in Table 8.10 below.

Table 8.10. Native speaker analysis – by the number of votes.

	Observed 1	Observed 2 and 3
BNC Rank 1	266	134

χ^2 =198.13. Incorporating Yates' correction for continuity, χ^2 = 196.64, df=1, highly significant (p=0.01 level of significance = 6.64).

This analysis also indicates that the ability to choose the most frequent collocate was statistically significant.

3.4.2. Non-native speakers

Two respondants failed to choose an answer at all for one stimulus word. Because of this, in Table 8.11 below the word *chance* has only 19 responses and so too does the word *information*. The absence of these scores did not materially affect the results.

Table 8.11. No. of NNS respondants choosing the different collocates in each word set (N = 20, except for *chance* and *information*).

	A. Amount	B. Approach	C. Basis	D. Chance	E. Details
BNC rank	Certain 16	New 19	Regular 20	Good 17	Further 18
	Maximum 3	Traditional 1	Regional 0	Greater 2	Financial 0
	Increasing 1	Whole 0	Casual 0	Main 0	Written 2

	F. Evidence	G. Fact	H. Future	I. Importance	J. Information
BNC rank	Further 15	Actual 3	Near 17	Great 17	Further 1
	Available 5	Well-known 15	Better 3	Utmost 2	General 13
	Oral 0	Established 2	Indefinite 0	Practical 1	Valuable 5

	K. Kind	L. Matter	M. Moment	N. Part	O. problem
BNC rank	Different 19	Different 14	Last 18	Important 14	Major 20
	Second 1	Small 2	Given 2	Small 5	Further 0
	Funny 0	Straightforward 4	Opportune 0	Lower 1	Growing 0

	P. Purpose	Q. Range	R. Role	S. Sense	T. Word
BNC rank	General 15	Wide 17	Important 10	Common 20	Last 12
	Special 5	Limited 2	Significant 10	Broad 0	New 7
	True 0	Huge 1	Supporting 0	Broadest 0	Particular 1

According to the above data, on two occasions the majority of the NNSs chose a collocate which is not the most frequent according to the BNC data: for *fact* and *information*. In addition, one of the words scored a tie: *role*, with 10 respondants believing *significant* to be the most frequent adjective partner, and the other 10 believing *important* to be so. To see whether the non-native speakers were, as a group, significantly more likely to choose the most frequent collocate than either of the other two on offer, a chi-squared goodness of fit test was employed.

Table 8.12. Non-native speaker analysis - by which option received the greatest no. of votes (Note that Observed =19, because there was a tie for *role*)

	Observed 1	Observed 2 and 3
BNC Rank 1	17	2

According to the data from Table 8.12, χ^2 =26.98. Incorporating Yates' correction for continuity χ^2 = 24.5, df=1, highly significant (p=0.01 level of significance = 6.64).

Table 8.13. Non-native speaker analysis - by no. of votes

	Observed 1	Observed 2 and 3
BNC Rank 1	302	96

It should be noted that there were 398 responses, and that Table 8.13, includes the votes for *role*, excluded in Table 8.12. According to the responses χ^2 =328.05. Incorporating Yates' correction for continuity χ^2 = 322.28, df=1, highly significant (p=0.01 level of significance = 6.64).

3.5. Discussion

The results from this experiment indicate that the majority of the respondants (both NS and NNS) had very little trouble in identifying the most frequent collocate in each of the twenty sets, and both groups performed better than their peers did in the corresponding recognition task where the most frequent noun of the different adjective stimuli had to be identified (experiment 3, chapter 7). In the discussion that follows, explanations for the rare occasions when the majority of the NSs and NNSs did not choose the most frequent collocate from the sets, are examined.

3.5.1. Embedding

Of the 20 most frequent collocations in their respective sets, four are in collocational frameworks in over 80% of their occurrences (*certain amount (of)*, *(on a) regular basis*, *(in) actual fact* and *wide range (of)*). The NNS respondants disagreed with the BNC data over the relative frequency of just one of these collocations. The majority of NNSs preferred *well-known fact* over *actual fact* as the most frequent collocation in the *fact* set, and though the majority of NSs chose *actual fact* (9), as a group, they too were quite split on whether this collocation or *well-known fact* (7) was the more common. In the chain '*in actual fact*' the word *fact* is rather redundant – the expression means 'actually', and the key word in the chain is *actual*. This is quite unlike the meaning of *fact* in *well-known fact*, in which *fact* has its prototypical meaning – 'truth'. Only 2 of the 40 respondants provided *actual* as an associate to *fact* in experiment 4 and this finding, together with the generally poor recognition that *actual fact* is the most frequent adjective-noun collocation containing *fact* suggests *either* that the denotational meaning of the word *fact* influenced the respondants in their choice (in that they preferred a combination in which *fact* had its stand-alone meaning), or that they failed to recognize that *actual fact* occurs in *in actual fact*, and as such, they did not generate enough exemplars from their mental lexicon searches. The NNSs in particular, preferred to choose a more 'complete' collocation, even though it is much less frequent. The other cases where the NS and NNS subjects failed to recognize the most frequent collocation are not among those which show dominant embedding, and different explanations must be forwarded to explain the failure to recognize them.

3.5.2. z-scores

In experiment 4, eight NS respondants provided the word *great*, and one provided *utmost* to the stimulus word *importance* (the most frequent collocate is *great*). However, in the recognition task, the majority of the NS respondants chose *utmost*. A check was conducted to see whether the respondants who produced *great* in experiment 4 were the same as those who believed it to be the most frequent here, and 4 of the 8 respondants

were consistent in their belief about the frequency of this word. (The single respondant who produced *utmost* in the production task also believed it to be the most frequent collocate in this task.) The two collocations are typically in the same frame (*of ADJ importance*) though *utmost* strongly attracts *the*. A possible explanation for why the majority of the NSs preferred the collocate *utmost* is that the z-score of *utmost*, when collocating with *importance* is higher than that of *great* (from higher to lower, the highest z-scores for the collocates of *importance* in the BNC are: *paramount, utmost, relative, crucial, vital, great*). In simple terms this means that *utmost* strongly predicts the presence of *importance* whereas for *great* this relationship is not so strong/exclusive. The strength of attraction between the two words *utmost* and *importance* may have affected the respondants' choices more than the raw frequency co-occurrence effect.

3.5.3. Miscellaneous

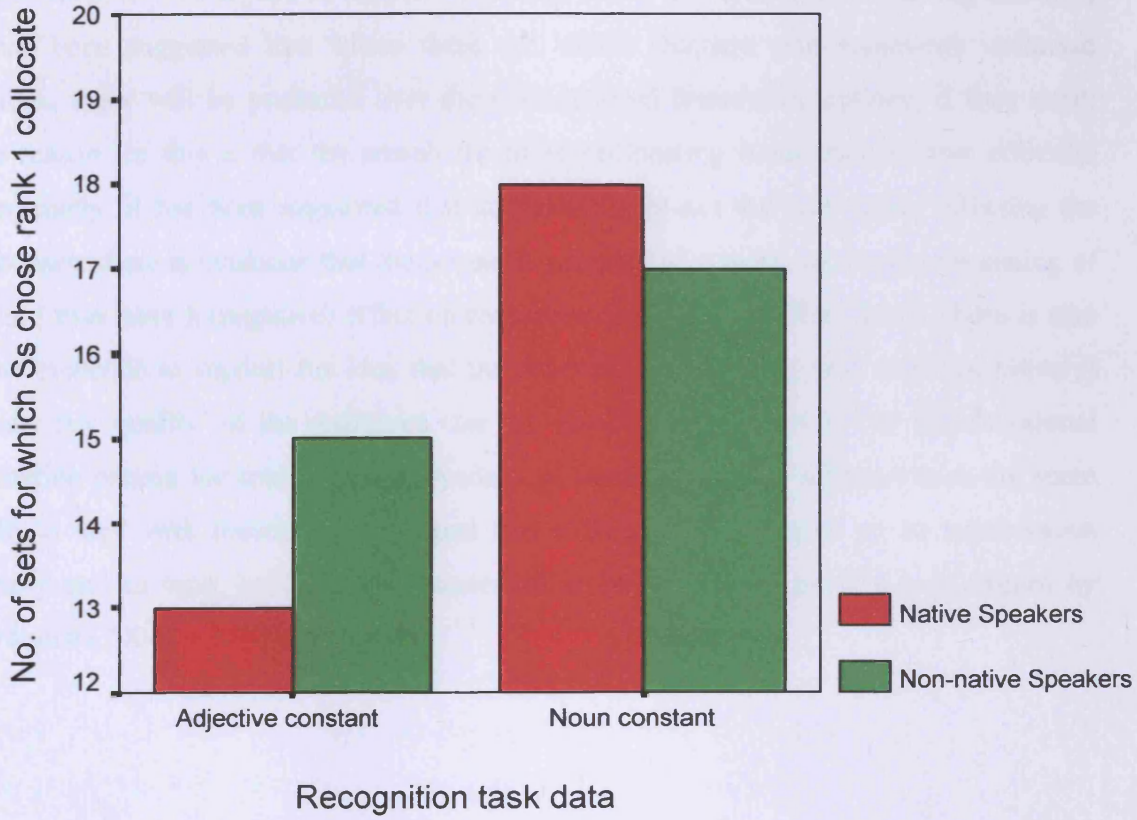
The NSs showed a strong preference for '*significant role*' over '*important role*', and the NNSs were split 50/50 on the same choices, though *important role* is around 4 times more frequent than the other option in the BNC. It was noted earlier on in this chapter (section 2.7.1) that in contrast to the NNSs (8 of whom produced *important* as the most frequent partner of *role*), only 1 NS produced this word in the productive task. *Important* has the highest z-score of the adjectives of *role*, and there seems to be no obvious reason why the majority of NSs chose as they did with this set of collocations.

The non-native speakers showed a marked aversion to choosing *further information* in the *information* set. *Further* has a much higher z-score than *general* for the information collocates – indeed it has the highest z-score. It should be noted that the majority of the respondants were quite happy to choose *further* as the most frequent collocate of *details*, so we cannot simply posit that they have an aversion to this word. It is a difficult case to explain.

3.6. Comparing the results from experiments 3 and 5

Figure 8.3 below shows how many of the 20 adjective sets and 20 noun sets, the majority of the subjects (NS and NNS) correctly identified the most frequent collocate for.

Figure 8.3. Overall comparison of results for NSs and NNSs in experiments 3 and 5



Clearly, when the noun is constant in the sets, the respondants' choices are closer to the BNC data, than when the adjective in the sets remains constant.

4. Summary

The results from experiment 4, reported in this chapter, suggest that both the NS and NNS teachers at KFUPM had more success in providing collocates in accord with BNC data when nouns rather than adjectives were provided as stimulus words. Further, the

results suggest that the respondants who were asked to recognize the most frequent collocation from a set of three, where the noun was constant in the set, concurred more often with the BNC data than was the case when the adjective was constant in the sets (experiment 3). As such, these results suggest that the accuracy of lexical intuitions are affected by the part of speech of the stimulus word – and this is seen most clearly in the second analysis of experiment 4. It seems that search restrictions, in line with the use of an availability heuristic in the frequency judgement, play less of a role in affecting the judgements, as was argued in chapter 7, section 2.6.4, in experiment 2. Having said this, it has been suggested that where there are viable frequent non-framework collocate options, these will be preferred over the collocational framework options, if they exist. The reason for this is that the search for these collocating items may be less effortful. Importantly, it has been suggested that accessibility is not the only factor affecting the responses: there is evidence that the semantic prosody of a word, or a salient meaning of a word may have a (negative) effect on responses (vis á vis the BNC data). There is also some evidence to support the idea that the denotational meaning of a word (negatively) affects the 'quality' of the responses (for the associations to *basis*). The uni-directional elicitation pattern for some frequent dyads (e.g. *small part, large amount*) does not seem to fit in well with theories that suggest that a frequent dyad takes on an autonomous identity and as such, becomes less connected to its constituent parts, e.g. as argued by Nordquist (2004).

Chapter 9 – Conclusion

1. Overview of key findings

In this study 5 experiments have been reported which were designed to help shed light on the elicited-data corpus-data debate, investigating lexical intuitions about adjective-noun collocations. The native speaker respondants who participated in experiments 2-5 were language teachers, and it is language teachers in particular, who find themselves in situations where they have to make spontaneous judgements about how a word is used, or are asked about a word's typical collocates etc. In this chapter, the key findings of the research are summarized, the limitations of the study are noted and some comments are made about how the findings of the research may be of help to language teachers.

As described in chapter 1, some corpus linguists question the ability of native speakers to provide reliable instances and examples of language use. The basis for holding this position was the belief that intuitions and corpus data clashed. Because corpus linguists believe corpora to be a more reliable record of language, the intuitions were believed to be at fault.

After reviewing the literature on collocation, and discussing the psycholinguistic representation of collocations, it seemed that the most obvious candidate for holistic storage in the lexicon was the restricted collocation. Fox (1987) had singled out frozen collocations as the exception to the rule that lexical intuitions about the frequent collocates of words were unreliable. However, it was not found in experiment 1 that restricted collocations (as a class) had privileged psycholinguistic representation. Alternative ways of looking at collocation representation had to be considered. In looking at the frequency estimation research, there was some evidence that frequency judgements could be biased, and the idea that some collocates, or collocations, might be less available than others in tests investigating collocation frequency seemed a subject worth investigating. Further reason to look into this subject was Wray's (2002) prediction that some material in the lexicon would not be as accessible as other material in test

conditions. Such biases would be principled, rather than random, and Wray argued that biases in searches would exclude access to a particular type of language - formulaic language.

However, some existing data did not seem to fit in well with Wray's hypothesis that a stimulus word would not elicit strong collocating partners of that word when the two words were a chunk or part of a larger chunk. There *were* cases in which the stimulus word *did* seem to have a strong connection with a larger unit containing that word. For example, there was the attested ability to provide a partner to a word in a frozen collocation (Greenbaum 1988), the provision of phrasal and idiomatic uses of *give* and *take* in elicited experiments (Gilquin 2005 a and b), and also it was found that some stimulus words elicited high frequency dyad partners, when existing free word association data was reanalyzed (as discussed in chapter 6).

I argued that Wray's theory could be called upon to explain the data, but that the best candidates for holistically stored (i.e. formulaic) language were idioms and collocational frameworks – not dyads. Even with the idioms and frameworks, it was not the case that the words in these larger chains were *always* unavailable: the data did not allow such an interpretation, and it was suggested that an important factor affecting the 'quality' of the intuitions was the availability of 'choices' in respondants' searches.

Regarding those choices, on the basis of the existing free word association data, it was argued that if there was a choice between idiom and non-idiom partners, non-idiom partners would be produced, or frequent idiom partners in which the components were more analysed. The matter of choice also came up in experiment 2, reported in chapter 7. The respondants did not, generally, provide collocates which typically occur in collocational frameworks of various types: the dominant responses tended to be complete collocations (e.g. *good idea, main street, recent event, difficult task*), rather than 'incomplete' collocations (e.g. *similar vein, large amount, real terms*) which were often highly frequent, and should have been known on the grounds that frequency of co-occurrence information is automatically encoded. It was argued that such results were

consistent with the use of an availability heuristic in assessing instances in memory, and consistent with the idea that some of the collocates of the words were less available than others. In experiment 4, the provision of the noun stimulus seemed to assist respondants in their searches, and the respondants *were* able to provide some collocates of the stimulus words that predominantly occurred within collocational frameworks. However, it was argued that this was so because there were no non-framework options. When there were, the non-framework options were, generally, preferred. For example, the dominant response to *basis* was *sound basis* compared to the most frequent collocation *regular basis* (typically occurring in *on a regular basis*), and there was a preference for 'moment is special' adjectives to precede *moment*, rather than the dominant collocation *last moment* (typically found in *at the last moment)*. In recognition experiments too there was some support for the embedding argument. In experiment 3, this explanation was forwarded to account for the NSs failure to recognize that *recent years, similar way* and *strong sense* were the most frequent collocations in their respective sets. In experiment 5 there was also some evidence to support this theory, where the NNSs failed to recognize *actual fact* to be the most frequent collocation in the *fact* set.

The research reported in this thesis suggests that chunks larger than the dyad are better candidates for formulaic status, and further, that the availability of choices may be important in affecting responses, and in explaining why intuitions about frequent collocates of high frequency words were often not elicited in the controlled word association tasks.

It was found in some of the dominant responses in experiment 2, that the stimulus word in the collocation was semantically bleached, or did not have its typical dictionary meaning in the combination (e.g. *full time, real time*). This brought into question the validity of the argument forwarded by corpus linguists that the denotational meaning of a word would drive intuitions about its typical meaning and uses, e.g. that *full* would be more associated with the meaning 'no space', rather than 'not temporary'. Though there were cases where the corpus linguist view could account for some of the responses (e.g. the dominant response *time* to *great,* and *sound* to *basis*), it was argued that the

embedding explanation could *better* explain such cases. Further, it was suggested that one way to explain why it is that a particular meaning of a word becomes psychologically salient is that its non-framework, i.e. dyad occurrences contribute to this saliency, but that frame occurrences, being less analysed, do not.

However, it is important to note that the accessibility explanation, which does seem able to account for *some* of the data, is insufficient, in itself, to explain *all* of the data. There are cases where the respondants' associations *should* have been more in line with the BNC data than they were, in cases where no availability restrictions were posited. There is some evidence that respondants might not be aware of the typical prosody of a word, or that they might attach a prosody to a word which the BNC data does not reflect (e.g. *moment* is positive). Further, it was occasionally found that the salient meaning of a word in the respondants' minds differed from the meaning in the most frequent collocations (e.g. the native speakers provided 'evidence is legal' collocates, but these were not the most frequent collocates in the BNC). As such, then, a more refined theory is needed to explain all of the data, and, given the complexity of the patterns found, developing such a theory is beyond the scope of this thesis. In what follows, I note the limitations of the current research and make suggestions for further research, which may help shed more light on the corpus-data elicited-data debate.

2. Limitations of the study and suggestions for further research

This thesis should be viewed as a step towards investigating the differences between corpus data and elicited data. Below, the limitations of the study are discussed, and a number of suggestions are made for further research.

2.1 Assumptions

2.1.1. Elicitation tests

In this research I have aligned the findings from controlled word association tasks with free association data (Moss & Older 1996), and with elicited data from sentence production tasks (Gilquin 2005 a and b, Nordquist 2004). It would be worthwhile investigating whether these different types of elicitation experiments result in different data. For example, if the stimulus words from the free association test were provided as stimulus words in controlled association tests, in which the respondants had to provide high frequency collocate responses, would the responses be noticeably different? This may be the case. It was found in experiment 2 that collocates were sometimes produced which delexicalised the adjective. This type of collocate response was not at all typical in the free association data (chapter 6). If the same stimulus words in chapter 6 were provided in a controlled association task it may be that there would be fewer noun collocate responses that, stereotypically, semantically entail the adjective quality (e.g. *green grass*). Respondants might reject stereotypical, though infrequent collocates (e.g. the response *ball* to *round*) and provide quite different responses. Another consideration is whether a sentence production task might provide different data. Though there is some evidence that the responses to *small* were similar in the controlled word association task and the Nordquist (2004) sentence production task, it would be unwise to generalize this finding. Testing the same words in different types of elicitation tasks is a worthy focus of further research. To align the data from different methodological approaches, as I have done, may hide some interesting facts.

2.1.2. Collocations

Unlike word frequency estimation research, there are two additional factors that may play a role in collocation frequency estimation: the strength of association between items and the effect of the frequency of the variable item in the set of collocations. How do the issues of co-occurrence frequency, individual word frequency, and strength of attraction

interact? Occasionally, I have argued that respondants may confuse the latter two issues with co-occurrence frequency, although it is difficult to be sure of the importance of the roles that they play. If experiments were designed in which there was a choice between choosing a collocation **either** with a high z-score frequency or a high co-occurrence frequency, and respondants were asked to provide the most frequent, it would be interesting to observe whether strength of attraction and co-occurrence frequency might be confused. In the experiments reported in this research, it was usually the case that z-scores and co-occurrence frequency measures concurred, and yet this will not be the case for all potential stimuli. More research should be conducted to investigate this possibility.

2.2 Methodological issues

2.2.1. Different instructions

If respondants were actually told, prior to the task, that many of the high frequency collocates typically occurred in frames, would their responses be better? This is a very interesting question which only further research can answer. If such an instruction did have a positive effect on the quality of the intuitions, vis á vis the BNC data, then 'collocation blindness' in the classroom could perhaps be cured, by raising teacher awareness of some key issues raised in this thesis. For example, the important role of many frequent adjective-noun collocations in *ADJ NOUN of NP* chains could be stressed in lexical studies in teacher training, and the presence of frequent adjective-noun collocations in adverbial chains (e.g. *in recent years*) could also be highlighted. It may be possible to raise teachers' conscious awareness of the supporting ADJ NOUN collocations, as all that is required is the (mental) placing of *of* after the stimulus word e.g. *different NOUN of...., ADJ kind of....* It is possible that a consequence of such consciousness raising may be to positively affect the quality of spontaneous collocate offerings in classroom scenarios, material design, test creation etc. However, given the varieties of some of the other chains, those in which the adjective-noun collocation is *not* in a supporting frame (e.g. *of great importance, in the near future, at the last moment*), even if consciously aware of the importance of frames, it may still be difficult for

teachers to be consciously aware of the presence of embedded collocations in such frameworks. As mentioned in chapter 6, section 2.2.2, there is a common belief that the fast associations in free association tests are indicative of strong connections between words in the lexicon. However, it may be the case that the provision of *additional* time to complete controlled word association tasks (after consciousness raising about the importance of frames) would help respondants to access frames, as it has been argued that such searches are more effortful, and, therefore more time consuming.

2.2.2. Extra clues

Very little has been said in this research about the roles of *a* and *the* in collocation chains – they have, in effect, been ignored. If respondants had been provided with either of these determiners preceding the adjective, would the responses have been very different in experiment 2? For example, if the respondants had been provided with *a similar*, rather than *similar*, would this have helped them access collocates in adverbial clauses (e.g. *in a similar way*), as more of the chain is provided? Certainly, adding a determiner would exclude plural noun responses, and it might assist respondants into conducting framework searches. Setting up research in which additional parts of a frame are provided, may help us discover at what point collocation framework searches are 'triggered'.

2.2.3. Restrictions on the collocates permitted

Did the fact that the respondants had to provide a particular part of speech collocate, rather than just a frequent collocate, mean that their answers were more affected by their knowledge of adjectives/nouns rather than their knowledge of collocates *per se*? The fact that *further* was not produced by any of the respondants, though it is the most frequent adjective collocate of *details, evidence* and *information* suggests that this may have been the case. It is possible that the task design forced the respondants to be considerably more analytical than they would have been in a free association response, and therefore, as a consequence, that their processing was even further removed from typical language

processing. Setting up an experiment in which *any* response is allowed, so long as respondants believe it to be frequent, would be a worthwhile variation to be explored.

2.3. Subjects

2.3.1. NS subjects

Would a group of less educated NS respondants produce different responses? According to Wray, educated speakers have engaged in more analysis of their lexicons, and the NS respondants in this experiment were all well educated and worked as university lecturers. It is a possibility, therefore, that less educated respondants would provide associates which are quite different from the subjects tested in this research. It was noted in chapter 6, section 2.4.1, that there are differences between the responses in free association data from different groups of native speakers. Assuming that less segmentation has gone on in native speakers who are less educated, it may be that the respondants' data would be quite different: but this should be tested, not assumed. However, the tasks as they stand at present would probably not be suitable for non-educated NSs, as some metalanguage knowledge is required (particularly for experiment 4) and so a task designed somewhat differently would be required[1].

2.3.2. NNS subjects

2.3.2.1. Learning style

The NNS respondants in this experiment learned English in a classroom environment. Would a group of respondants who had learned English as a second language *without* classroom learning have responded in very different ways to these classroom taught learners? We would expect the classroom taught learners to have adopted a more analytical approach to language learning; however, it may be that greater analysis is

[1] Some friends in the UK were asked to conduct the controlled word association task with less educated speakers and several commented that many of the responses provided were not valid.

simply a consequence of learning language later on in life, and part of biological and cultural development, rather than to do with the shape of the input in the classroom. However, this should be tested not assumed.

2.3.2.2. Native language

Would a group of advanced NNSs from another language background have produced different responses? The idea that the native language affects collocation knowledge in L2 has been argued by Bahns (1993), who, in particular, has noted transfer of L1 collocation patterns into L2 collocation usage. Further research on this could use parallel corpora and investigate whether the L2 responses are affected by the frequency of the L1 equivalent collocation. For example, might it be the case that some of the NNS respondants produced *high* to *importance*, rather than *great*, because of transfer from Arabic? Useful insights into language representation and language transfer could be gained by investigating this subject.

2.3.2.3. Different stimulus words in different languages

The differences between the results obtained in experiments 2 and 3, compared to 4 and 5 suggest that it was easier to provide high frequency collocates of nouns than high frequency collocates of adjectives. Is this only so for languages like English where attributive adjectives occur before the nouns rather than postnominally, such as in French and Arabic? Does typical word order in the language have any effect on the ability to provide collocates from a particular form class of words? Again, this is an issue worth investigating.

2.4. Analyses

2.4.1. Reconsidering the less frequent responses

In experiments 2 and 4, no credit was given to responses which were outside the 20 most frequent collocates of the stimulus word, although it was sometimes the case that these responses were only just outside this range. This decision was made on the basis that the collocates having a >20 rank were, on average, at least 10 times less frequent than the most frequent collocate, and word frequency estimation abilities have been established at this difference in frequency (see chapter 4, section 2.1.1). However, it may be that a design measure more sensitive to the >20 responses, or one designed to group together the responses into different semantic classes (an approach adopted by Gilquin 2005 a and b) would have revealed some interesting additional data. For example, it may be found that the respondants are sensitive to semantic class preferences of the words, though not perhaps the most frequent collocations in that class.

2.4.2. Embedding explanation

I have argued that, given one of the two words which form a bare collocation in a collocation framework, the second word is not likely to be produced in a controlled association task. However, I have argued this only when the embedding is very frequent (around 80% of more) in the same chain. What though if there are a small number of variable items in the chain? For example, *possible way*, often occurs with a superlative before the collocation (e.g. *best, nicest, worst, strongest*). How does this affect the visibility of the collocation *possible way*? None of these instances are, individually, particularly frequent, but the pattern *the SUPERLATIVE ADJECTIVE possible way* is a very common frame within which *possible way* occurs. More research should investigate how invariable a chain must be, before it is considered to be formulaic, and whether the number of different items filling a slot in the frame may have an effect on the transparency of the items in the collocation.

3. The last word

Experiment 2 was a frustrating task for a number of the respondants (both NSs and NNSs) who knew that the task looked simple, but struggled to provide collocates that they believed were frequent partners of the stimulus words. One NS reported that he couldn't get *elephant* out of his mind when he was thinking of collocates for *large*, and a NNS complained to me, after he had completed the task, about the difficulties he had faced. "Well, look at 'different' for example", he explained, "it's very difficult to find a partner for it, because it is used in so many different ways". He then stopped, immediately realizing that the last word he had spoken was a frequent collocate of *different*, and he proceeded to change his response from *things* to *ways*, which is the most frequent collocate of *different* according to the BNC (1213 instances); his original response was retained in the analysis of the data. Both of the examples reported above suggest that the skills employed in the analytical production experiments differ considerably from those employed when language is used in a more spontaneous, natural way.

Aizawa et. al. (2001) comment, in the context of word frequency estimation that, "native speaker's intuitions about frequency are not always straightforward" (p.80). While I would agree with this comment, some progress has been made in this thesis towards making sense of frequency estimation in the subject of adjective-noun collocations, and while it is true that intuitions about frequency are not always straightforward, it has been argued that there may be a principled bias affecting collocate and collocation frequency judgements. Corpora should be seen as invaluable aids in making us more consciously aware of the phrasal patternings of many frequent adjective-noun collocates, which, it has been argued, are in the blind-spot of a teacher's mirror on language. Whether we can be taught to look over our shoulders, and be more consciously aware of the prevalence of such items in the language, remains to be seen.

References

Aisenstadt, E. 1981. Restricted collocations in English lexicology and lexicography. *ITL Review of Applied Linguistics,* 53, pp.53-61.

Aitchison, J. 1994. *Words in the mind,* Second edition. Oxford: Blackwell.

Aizawa, K., Mochizuki, M. & Meara P. 2001. Intuition of word frequency: What does it tell us about vocabulary knowledge? *Research Reports of the Faculty of Engineering, Tokyo Denki University – General Education Edition,* 20, pp.75-82.

Alario, F.- Xavier, Costa, A. & Caramazza, A. 2002. Frequency effects in noun phrase production: Implications for models of lexical access. *Language and Cognitive Processes,* 17(3), pp.299-319.

Arnaud, P.J.L. 1990. Subjective word frequency estimates in L1 and L2. Paper presented at the 9th World Congress of Linguistics, Thessaloniki. *ERIC Document* ED329120, pp.1-15.

Aston, G. 1997. Enriching the learning environment: corpora in ELT. In: Wichmann, A., Fligelstone, S., McEnery, T. & Knowles G. eds. *Teaching and language corpora.* Harlow: Longman, pp.51-64.

Aston, G. & Burnard, J. 1997. *The BNC handbook: Exploring the British corpus with SARA.* Edinburgh: Edinburgh University Press.

Attneave, F. 1953. Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology,* 45(2), pp.81-86.

Baayen, R.H., Piepenbrock, R & Gulikers, L. 1995. *The CELEX lexical database* [CD ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Backman, J. 1976. Some common word attributes and their relations to objective frequency counts. *Scandinavian Journal of Educational Research*, 20, pp.175-186.

Backman, J. 1978. Subjective structures in linguistic recurrence. *ERIC Document* ED 180195, pp.3-22.

Bahns, J. 1993. Lexical collocations: a contrastive view. *English Language Teaching Journal*, 47(1), pp. 56-63.

Bahns, J. & Eldaw, M. 1993. Should we teach EFL students collocations? *System*, 21(1), pp.101-114.

Balota, D.A., Pilotti, M. & Cortese, M.J. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition*, 29(4), pp.639-647.

Bandera, L., Della Salla, S., Laiacona, M., Luzzatti, C. & Splinnler, H. 1991. Generative associative naming in dementia of the Alzheimer's type. *Neuropsychologia*, 29, pp.291-304.

Barlow, M. 1996. Corpora for theory and practice. *International Journal of Corpus Linguistics*, 1(1), pp.1-37.

Barnbrook, G. 1996. *Language and Computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Baudot, J. 1992. *Fréquences d'utilisation des mots en français écrit contemporain.* Montréal: Presses de l'université de Montréal.

Beaugrande, R. de. 1996. The pragmatics of doing language science: The warrant for large corpus linguistics. *Journal of Pragmatics*, 25, pp.503-535.

Beaugrande, R. de. 1999. Reconnecting real language with real texts: text linguistics and corpus linguistics. *International Journal of Corpus Linguistics* 42(2), pp.243-259.

Benson, M. 1985. Collocations and idioms. In: Ilson, R. ed. *Dictionaries, lexicography and language learning.* (ELT documents 120). Oxford: Pergamon, pp.61-68.

Benson, M. 1989. The structure of the collocational dictionary. *International Journal of Lexicography,* 2(1), pp.1-13.

Benson, M., Benson, E. & Ilson, R.1986. *The BBI combinatory dictionary of English: A guide to word combinations.* Amsterdam: John Benjamins.

Berry, D.C. 1996. How implicit is implicit learning? In: Underwood, G. ed. *Implicit cognition.* Oxford: Oxford University Press, pp.203-225.

Bialystok, E. 1994. Representation and ways of knowing. In: Ellis, N. ed. *Implicit and explicit learning of languages.* London: Academic Press, pp.549-569.

Biber, D. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing,* 8(4), pp.243-257.

Biber, D. 1994. Using register-diversified corpora for general language studies. In: Armstrong, S. ed. *Using large corpora.* Cambridge, Mass: Bradford Book MIT Press, pp.179-201.

Biber, D. 1996. Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics,* 1(2), pp.171-197.

Biber, D., Conrad, S. & Reppen, R. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics,* 15(2), pp.169-189.

Biber, D., Conrad, S. & Reppen, R. 1996. Corpus based investigations of language use. *Annual Review of Applied Linguistics,* 16, pp.115-135.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus linguistics: Investigating Language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman grammar of spoken and written English.* Harlow: Longman.

Blair, I.V., Urland, G.R. & Ma, J.E. 2002. Using internet search engines to estimate word frequency. *Behavior Research Methods, Instruments and Computers,* 34(2), pp.286-290.

Bley-Vroman, R. 2002. Frequency in production, comprehension and acquisition. *Studies in Second Language Acquisition*, 24, pp.209-213.

British National Corpus - World edition, CD-ROM. 2000. Oxford: Humanities Computing Unit of Oxford University.

Boyland, J.T. 2001. Hypercorrect pronoun case in English? In: Bybee, J. & Hopper, P. eds. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp.383-404.

Brown, N. 1995. Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 21, pp.1539-1553.

Bush, N. 2001. Frequency effects and word-boundary palatalization in English. In: Bybee, J. & Hopper, P. eds. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp.255-280.

Bybee, J. & Scheibman, J. 1999. The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics,* 37(4), pp.575-596.

Bybee, J. & Hopper, P. 2001. Introduction to frequency and the emergence of linguistic structure. In: Bybee, J. & Hopper, P. eds. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp.1-24.

Byrd, P. 1995. Issues in the writing of grammar textbooks. In: Byrd, P. ed. *Materials writer's guide*. Boston: Heinle & Heinle, pp.45-63.

Cacciari, C. 1993. The place of idioms in a literal and metaphorical world. In: Cacciari, C. & Tabossi, P. eds. *Idioms: processing, structure and interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum, pp.27-55.

Cantos, P. & Sánchez, A. 2001. Lexical constellations: What collocates fail to tell. *International Journal of Corpus Linguistics,* 6(2), pp.199-228.

Carroll, J.B. 1971. Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior,* 10, pp.722-729.

Carroll, J.B., Davies, P. & Richman, B. 1971. *Word frequency Book.* New York: American Heritage Publishing Co.

Carter, R. 1983. "You look nice and weedy these days!" Lexical associations and the foreign language learner. *Journal of Applied Language Study,* 1(2), pp.172-189.

Carter, R. 1987. *Vocabulary applied linguistic perspectives*. London: Routledge.

Carter, R. & McCarthy, M. 1999. The English get-passive in spoken discourse. *English Language and Linguistics,* 3(1), pp.41-58.

Channell, J. 2000. Corpus-based analysis of evaluative text. In: Hunston, S. & Thompson, G. eds. *Evaluation in text: Authorial Stance and the construction of discourse*. Oxford: Oxford University Press, pp.38–55.

Chomsky, N. 1962. Paper given at the University of Texas 1958, 3rd Texas conference on problems of linguistic analysis on English. Austin: University of Texas.

Church, K.W. & Hanks, P. 1989. Word association norms, mutual Information, and Lexicography. *Proceedings of the Annual Meeting of Association for Computational Linguistics*, Vancouver, 76-83.

Clark, H.C. 1970. Word associations and linguistic theory. In: Lyons, J. ed. *New horizons in linguistics*. Harmondsworth, Middlesex: Penguin Books, pp.271-286.

Clear, J. 1993. From Firth principles: Computational tools for the study of collocation. In: Baker, M., Francis, G. & Tognini-Bonelli, E. eds. *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, pp.271-292.

Collins, P.C. 1996. Get passives in English. *English World-wide,* 15(1), pp.43-56.

Cook, G. 1998. The uses of reality: a reply to Ronald Carter. *English Language Teaching Journal,* 52(1), pp.57-63.

Cosmides, L. & Tooby, J. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition,* 58, pp.1-73.

Cowie, A.P. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics,* 2(3), pp.223-235.

Cowie, A.P. 1992. Multiword lexical units and communicative language teaching. In: Arnaud, P.J.L. & Bejoint, H. eds. *Vocabulary and applied linguistics*. London: Macmillan, pp.1-12.

Cowie, A.P. 1998. Phraseological dictionaries: Some East-West comparisons. In: Cowie, A.P. ed. *Phraseology*. Oxford: Clarendon Press, pp.209-228.

Coulmas, F. 1979. On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics,* 3, p.239-266.

Coulmas, F. 1981. Introduction: Conversational Routine. In: Coulmas, F. ed. *Conversational routine*. The Hague: Mouton, pp.1-17.

Cramer, P. 1968. *Word Association*. New York and London: Academic Press.

Cromm, O. 2001. On the relative influence of corpus and dictionary size in a study using non-parallel corpora. *Journal of Quantitative Linguistics*, 8(2) pp.137-148.

Crystal, D. 1998. Speaking of writing and writing of speaking [online]. *Longman Language Review*, 1. Available at: <URL:http://www.longman.com/dictionaries/llreview/lwrit1.html> [Accessed 1st December 2006].

D'Arcais, G.B.F. 1993. The comprehension and semantic interpretation of idioms. In: Cacciari, C. & Tabossi, P. eds. *Idioms: Processing, structure and interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum, pp.79-98.

Deese, J. 1962. Form class and the determinants of association. *Journal of Verbal Learning and Verbal Behavior,* 1, pp.79-84.

Desrochers, A. & Bergeron, M. 2000. Valeurs de fréquence subjective et d'imagerie pour un énchantillon de 1916 substantifs de la langue française. *Canadian Journal of Experimental Psychology*, 54(4), pp.274-325.

Dulaney, D.E., Carlson, R.A. & Dewey, G.I. 1984. A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General,* 114, pp.25-32.

Dutch Center for Lexical Information. 1995. *The celex lexical database*. Nijmegen: Dutch Center for Lexical Information.

Edinburgh Associative Thesaurus (EAT) [online] Available at URL <http://www.eat.rl.ac.uk> [Accessed Sept 2005].

Ellis, N. 1993. Rules and instances in foreign language learning: interactions of explicit and implicit knowledge. *European Journal of Cognitive Psychology*, 5 (3), pp.289-318.

Ellis, N. 2002a. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, 24, pp.297-339.

Ellis, N. 2002b. Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition,* 24, pp.143-188.

Ellis, R. 1993. The Structural syllabus and second language acquisition. *TESOL Quarterly* 27(1), pp.91-113.

Ellis, R. 1994. A theory of instructed second language acquisition. In: Ellis, N. ed. *Implicit and explicit learning of languages*. London: Academic Press, pp.79-114.

Engwall, G. 1994. Not chance but choice: Criteria in corpus creation. In: Atkins, B.T.S. & Zampoli, A. eds. *Computational approaches to the Lexicon.* Oxford: Oxford University Press, pp.49-82.

Farghal, M. & Obeidat, H. 1995. Collocations: a neglected variable in EFL. *International Review of Applied Linguistics,* 33(4), pp.315-331.

Fellbaum, C. 1993. The determiner in English idioms. In: Cacciari, C. & Tabossi, P. eds. *Idioms: processing, structure and interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum, pp.271-294.

Fiengo, R. 2003. Linguistic intuitions. *The Philosophical Forum,* 34(3 & 4), pp.253-265.

Fillmore, C. 1992. "Corpus linguistics" or "computer-aided armchair linguistics" In: Svartvik, J. ed. *Directions in corpus linguistics.* Berlin: Mouton, pp.35-60.

Firth, J.R. 1957. *Papers in linguistics 1934-1951.* London: Oxford University Press.

Fontenelle, T. 1998. Discovering significant lexical functions in dictionary entries. In: Cowie, A.P. ed. *Phraseology.* Oxford: Clarendon Press, pp.189-207.

Fox, G. 1987. The case for examples. In: Sinclair, J. ed. *Looking up.* London: Collins, pp.137-149.

Fox, G. 1998. Using corpus data in the classroom. In: Tomlinson, B. ed. *Materials development in language teaching.* Cambridge: Cambridge University Press, pp.25-43.

Francis, G. 1993. A corpus-driven approach to grammar. Principles, methods and examples. In: Baker, M., Francis, G. & Tognini-Bonelli, E. eds. *Text and technology. In honour of John Sinclair.* Amsterdam: John Benjamins, pp.137-156.

Francis, G. & Sinclair, J. 1994. 'I bet he drinks Carling Black Label': A riposte to Owen on corpus grammar. *Applied Linguistics*, 15(2), pp.190-200.

Francis, W.N. 1982. Problems of assembling and computerizing large corpora. In. Johansson, S. ed. *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities, pp.7-24.

Frey, E. 1981. Subjective word frequency estimates and their stylistic relevance in literature. *Poetics,* 10(4-5), pp.395-407.

Gavioli, L. 1997. Exploring texts through the concordancer: guiding the learner. In: Wichmann, A., Fligelstone, S., McEnery, T. & Knowles G. eds. *Teaching and language corpora.* Harlow: Longman, pp.83-99.

Gavioli, L. & Aston, G. 2001. Enriching reality: language corpora in language pedagogy. *English Language Teaching Journal,* 55(3), pp.238-246.

Gernsbacher, M.A. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness and polysemy. *Journal of Experimental Psychology: General*, 113(2), pp.256-281.

Gibbs, R.W. 1993. Why idioms are not dead metaphors. In: Cacciari, C. & Tabossi, P. eds. *Idioms: processing, structure and interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum, pp.57-77.

Gilquin, G. 2005a. What you think ain't what you get: Highly polysemous verbs in mind and language. Paper presented at "From Gram to mind: Grammar as Cognition". Bordeaux 19-21 May 2005.

Gilquin, G. 2005b. To take or not to take phraseology into account. Paper presented at "The many faces of phraseology. An interdisciplinary conference", Louvain-la-Neuve, 13-15 October 2005, pp.155-157.

Giora, R. 2002. Literal vs. figurative: Different or equal? *Journal of Pragmatics,* 34, pp. 487-506.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: collocations and formulae. In: Cowie, A.P. ed. *Phraseology.* Oxford: Clarendon Press, pp.145-160.

Greenbaum, S. 1988. *Good English and the grammarian.* London and New York: Longman.

Greidanus, T. & Nienhuis, L. 2001. Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal,* 85(4), pp.567-577.

Griffin, Z.M. & Bock, K. 1998. Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language,* 38, pp.313-338.

Groot, A.M.B. de. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 15(5), pp.824-845.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

Hamilton, M. & Rajaram, S. 2001. The concreteness effect in implicit and explicit memory tests. *Journal of Memory and Language,* 44, pp.96-117.

Hanks, P. 2004. The syntagmatics of metaphor and idiom. *International Journal of Lexicography,* 17(3), pp.245-274.

Hasher, L. & Chromiak, W. 1977. The processing of frequency information: an automatic mechanism? *Journal of Verbal Learning and Verbal Behavior,* 16, pp.173-184.

Hasher, L. & Zacks, R.T. 1979. Automatic and effortful processes in memory. *Journal of Experimental Psychology: General,* 108, pp.356-388.

Hasher, L. & Zacks, R.T. 1984. Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist,* 39, pp.1372-1388.

Herbst, T. 1996. What are collocations: sandy beaches or false teeth? *English Studies,* 4, pp.379-393.

Hintzman, D.L. 1978. Contextual variability and memory for frequency. *Journal of experimental psychology: Human Learning and Memory,* 4(5), pp.539-549.

Hirsh, K.W. & Tree, J.J. 2001. Word association norms for two cohorts of British adults. *Journal of Neurolinguistics*, 14, pp.1-44.

Hoey, M. 1991. *Patterns of lexis in text.* Oxford: Oxford University Press.

Hoey, M. 2000. A world beyond collocation: new perspectives on vocabulary teaching. In: Lewis, M. ed. *Teaching collocation.* Hove: Language Teaching Publications, pp.224-243.

Hofland, K. & Johansson, S. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities.

Howarth, P. 1998a. Phraseology and second language proficiency. *Applied Linguistics,* 19(1), pp.24-44.

Howarth, P. 1998b. The phraseology of learners' academic writing. In: Cowie, A.P. ed. *Phraseology.* Oxford: Clarendon Press, pp.161-186.

Howell, W.C. 1973. Representation of frequency in memory. *Psychological Bulletin,* 80(1), pp.44-53.

Howes, D.H. 1966. A word-count of Spoken English. *Journal of Verbal Learning and Verbal Behaviour,* 5, pp.572-604.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & Francis, G. 2000. *Pattern grammar*. Amsterdam: John Benjamins.

Jessen, F., Heun, R., Erb, M., Granath, D.-O., Klose, U., Papassotiropoulos, A. & Grodd, W. 2000. The concreteness effect: Evidence for dual coding and context availability. *Brain and Language,* 74, pp.103-122.

Jenkins, J.J. & Palermo, D.S. 1965. Further data on changes in word-association norms. *Journal of Personality and Social Psychology,* 1(4), pp.303-309.

Jonides, J. & Jones, C.M. 1992. Direct coding for frequency of occurrence. *Journal of Experimental Psychology: Learning Memory and Cognition,* 18(2), pp.368-378.

Johnson, D.M. 1956. Word-association and word-frequency. *American Journal of Psychology,* 69, pp.125-127.

Juilland, A, Brodin D. & Davidovitch C. 1970. *Frequency Dictionary of French words.* The Hague-Paris: Mouton.

Kahneman, D. & Tversky, A. 1982. On the study of statistical intuitions. In: Kahneman, D., Slovic, P & Tversky, A. eds. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, pp.493-508.

Keller, F. & Lapata, M. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics,* 29(3), pp.459-484.

Kennedy, G. 1991. Between and through: The company they keep and the functions they serve. In: Aijmer, K. & Altenberg, B. eds. *English corpus linguistics*. Essex: Longman, pp.95-110.

Kennedy, G. 2002. Variation in the distribution of modal verbs in the British National Corpus. In: Reppen, R. ed. *Using corpora to explore linguistic variation.* Philadelphia: John Benjamins, pp.73-90.

Kennedy, G. 2003. Amplifier collocations in the British National Corpus. *TESOL Quarterly,* 37(3), pp.467-485.

Kilgarriff, A. 1995. *BNC database and word frequency lists* [online]. Available at: <URL: http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html> [Accessed 22nd March 2005].

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), pp.1-37.

Kilgarriff, A. & Grefenstette, G. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics,* 29(3), pp.333-347.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. 2004. The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress. Lorient, France: Universite de Bretagne-Sud, pp.105-116.

Kita, K. & Ogata, H. 1997. Collocations in language learning: corpus-based automatic compilation. *Computer Assisted Language Learning*, 10(3), pp.229-238.

Kjellmer, G. 1982. Some problems relating to the study of collocations in the Brown corpus. In. Johansson, S. ed. *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities, pp.25-33.

Kjellmer, G. 1991. A mint of phrases. In: Aijmer, K. & Altenberg, B. eds. *English corpus linguistics*. Harlow: Longman, pp.111-127.

Kjellmer, G. 1994. *Dictionary of collocations*. Oxford: Clarendon Press.

Knowles, G. & Don, Z.M. 2004. The notion of a "lemma": headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1), pp.69-81.

Koriat, A. 2000. Control processes in memory. In: Tulving, E. & Craik, F.I.M. eds. *The Oxford handbook of memory.* Oxford: Oxford University Press, pp.333-345.

Krenn, B. & Evert, S. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In 39th Annual meeting and 10th conference of the European Chapter of the Association for Computational Linguistics (ACL39), pp.39-46.

Kruse, H., Pankhurst, J. & Sharwood Smith, M. 1987. A multiple word association probe in second language acquisition research. *Studies in Second language Acquisition,* 9, pp.141-154.

Kučera, H. & Francis, W. 1967. *Computational analysis of present-day American English.* Providence: Brown University Press.

Lawrence, S. & Giles, C.L. 2000. Accessibility of information on the Web. *Intelligence,* 11(1), pp.32-39.

Lee, D. 2002. Genres, register, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language and Computers,* 42(1), pp.247-292.

Leech, G. 1991. The state of the art in corpus linguistics. In: Aijmer, K. & Altenberg, B. eds. *English corpus linguistics*. Essex: Longman, pp.8-29.

Leech, G., Rayson, P. & Wilson, A. 2001. *Word frequencies in written and spoken English*. Harlow: Longman.

Lewis, M. 1997. *Implementing the lexical approach.* Hove: Language Teaching Publications.

Lewis, M. 2000. Materials and resources for teaching collocation. In: Lewis, M. ed. *Teaching collocation.* Hove: Language Teaching Publications, pp.186-204.

Lichtenstein, S., Slovik, P., Fischhoff, B., Layman, M. & Combs, B. 1978. Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), pp.551-578.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Baker, M., Francis, G. & Tognini-Bonelli, E. eds. *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, pp.157-176.

Lyons, J. 1997. *Semantics 1*. Cambridge: Cambridge University Press.

MacAndrew, S.B.G. & Harley, T.A. 2000. Is lexical retrieval in speech production like recall or recognition? The effects of word frequency and neighbourhood size. In: Gleitman, L.R & Joshi, A.K. eds. *Proceedings of the 22nd Annual meeting of the cognitive science society*. Mahwah, NJ: Erlbaum, pp.328-333.

Mackin, R. 1978. On collocations: words shall be known by the company they keep. In: Strevens, P. ed. *In honour of A.S. Hornby*. Oxford: Oxford University Press, pp.149-165.

Malmkjær, K. 1993. Who can make *nice* a better word than *pretty*? Collocation, translation and psycholinguistics. In: Baker, M., Francis, G. & Tognini-Bonelli, E. eds. *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, pp.213-232.

May, R.B. & Tryk, H.E. 1970. Word sequence, word frequency, and free recall. *Canadian Journal of Psychology,* 24, pp.299-304.

McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meara, P. 1980. Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts,* 13(4), pp.221-246.

Meara, P. 1984. The study of lexis in interlanguage. In: Davies, A., Howart, A. & Criper, C. eds. *Interlanguage*. Edinburgh: Edinburgh University Press, pp 225-235.

Meijs, W. 1996. Linguistic corpora and lexicography. *Annual Review of Applied Linguistics,* 16, pp.99-114.

Mel'čuk, I. 1998. Collocations and lexical functions. In: Cowie, A.P. ed. *Phraseology.* Oxford: Clarendon Press, pp.23-53.

Meyer, C.F., Grabowski, R., Han H-Y., Mantzouranis, K. & Moses, S. 2003. The World Wide Web as linguistic corpus. *Language and Computers,* 46(1), pp.241-254.

Mindt, D. 1996. English corpus linguistics and the foreign language teaching syllabus. In: Thomas, J. & Short, M. eds. *Using corpora for language research.* London: Longman, pp. 232-259.

Moss, H. & Older L. 1996. *Birkbeck word association norms.* Hove, UK: Psychology Press.

Murison-Bowie, S. 1996. Linguistic corpora and language teaching. *Annual Review of Applied Linguistics,* 16, pp.182-199.

Nation, I.S.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.

Nelson, D.L. & Schreiber, T.A. 1992. Word concreteness and word structure as independent determinants of recall. *Journal of Memory and Language,* 31, pp.237-260.

Nelson, D.L., McEvoy, C.L. & Schreiber, T.A. 1998. *The University of South Florida word association, rhyme and word fragment norms* [online]. Available at: <URL: http://w3.usf.edu/FreeAssociation/Intro.html> [Accessed 10th February 2006].

Nelson, D.L. & McEvoy, C.L. 2000. What is this thing called frequency? *Memory and Cognition,* 28(4), pp.509-522.

Nelson, D.L., McEvoy, C.L. & Dennis, S. 2000. What is free association and what does it measure? *Memory and Cognition*, 28(6), pp.887-899.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), pp.223-242.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Nordquist, D. 2004. Comparing elicited data and corpora. In: Achard, M. & Kemmer, S. eds. *Language, culture and mind*. Leland Stanford University: CSLI publications, pp.211-223.

O'Keeffe, A. & Farr, F. 2003. Using language corpora in initial teacher education. *TESOL Quarterly*, 37(3), pp.389-417.

Owen, C. 1993. Corpus-based grammar and the Heineken effect: lexico-grammatical description for language learners. *Applied Linguistics,* 14(2), pp.167-187.

Owen, C. 1996. Do concordances require to be consulted? *English Language Teaching Journal,* 50(3), pp.219-224.

Partington, A. 1998. *Patterns and meanings: Using corpora for English Language research and teaching*. Amsterdam: John Benjamins.

Partington, A. 2004. "Utterly content in each other's company" Semantic prosody and semantic preference. *International Journal of Corpus Linguistics,* 9(1), pp.131-156.

Pawley, A. & Syder, F.H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J.C. & Schmidt, R.W. eds. *Language and communication*. New York. Longman, pp.191-226.

Peters, A.M. 1983. *Units of language acquisition.* Cambridge: Cambridge University Press.

Postman, L. 1970. The California norms: association as a function of word frequency. In: Postman, L. & Keppel, G. eds. *Norms of word association.* New York: Academic Press, pp.241-320.

Quirk, R. & Stein G. 1996. Sipping a cocktail of corpora. In: Thomas, J. & Short, M. eds. *Using corpora for language research.* London: Longman, pp.27-35.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1972. *A Grammar of contemporary English.* London: Longman.

Radden, G. 1992. The cognitive approach to natural language. In Pütz, W. ed. *Thirty years of linguistic evolution. Studies in honour of René Dirven on the occasion of his sixtieth birthday.* Amsterdam: John Benjamins, pp.513-541.

Rastall, P. 1997. Intuitions, associations and indeterminacy. *La Linguistique*, 33(2), pp. 79-94.

Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing,* 10, pp.355-371.

Read, J. 2000. *Assessing vocabulary.* Cambridge: Cambridge University Press.

Reber, A.S. 1993. *Implicit learning and tacit knowledge – An essay on the cognitive unconscious.* Oxford: Oxford University Press / Clarendon.

Renouf, A. 1987a. Moving on. In: Sinclair, J. ed. *Looking up.* London: Collins, pp.167-178.

Renouf, A. 1987b. Corpus development. In: Sinclair, J. ed. *Looking up*. London: Collins, pp.1-39.

Renouf, A. 1997. Teaching corpus linguistics to teachers of English. In: Wichmann, A., Fligelstone, S., McEnery, T. & Knowles G. eds. *Teaching and Language Corpora*. Harlow: Longman, pp.255-266.

Renouf, A. 2003. WebCorp: providing a renewable data source for corpus linguists. *Language and Computers,* 48(1), pp.39-58.

Renouf, A. & Sinclair, J.M. 1991. Collocational frameworks in English. In: Aijmer, K. & Altenberg, B. eds. *English corpus linguistics*. Harlow: Longman, pp.128-143.

Resnik, P. & Smith, N.A. 2003. The web as a parallel corpus. *Computational Linguistics,* 29(3), pp.349-380.

Richards, J.C. 1974. Word lists: problems and prospects. *RELC Journal,* 5(2), pp.69-84.

Richards, J.C. 1976. The role of vocabulary teaching. *TESOL Quarterly*, 10(1), pp.77-89.

Riegel, K. & Zivian, I.W.M. 1972. A study of inter- and intra-lingual associations in English and German. *Language Learning,* 22, pp.151-163.

Ringeling, T. 1984. Subjective estimations as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin,* 20(8), pp.59-69.

Rinsland, H.D. 1945. *A Basic vocabulary of Elementary School Children.* New York: Macmillan.

Roland, D. & Jurafsky, D. 2002. Verb sense and verb subcategorisation probabilities. In: Paola, M. ed. *The lexical basis of sentence processing. Formal, computational and experimental issues*. Philadelphia: John Benjamins, pp.333-353.

Rozenzweig, M.R. 1964. Word associations of French workmen: comparisons with associations of French students and American workmen and students. *Journal of Verbal Learning and Verbal Behavior*, 3, pp.57-69.

Rozin, P., Kurzer, N. & Cohen, A.B. 2002. Free associations to "food": the effects of gender, generation and culture. *Journal of Research in Personality,* 36, pp.419-441.

Rugg, M.D., Cox, C.J.C., Doyle, M.C. & Wells, T. 1995. Event-related potentials and the recollection of low and high frequency words. *Neuropsychologia,* 33(4), pp.471-484.

Schmid, H-J. 2000. *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin: Mouton de Gruyter.

Schmitt, N. 1998a. Measuring collocational knowledge: key issues and an experimental assessment procedure. *International Review of Applied Linguistics,* 119-120, pp.27-47.

Schmitt, N. 1998b. Quantifying word association responses: what is native like? *System,* 26, pp.389-401.

Schmitt, N. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. & Dunham, B. 1999. Exploring native and non-native intuitions of word frequency. *Second Language Research*, 15(4), pp.389-411.

Schmitt N., Grandage, S. & Adolphs, S. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In: Schmitt, N. ed. *Formulaic sequences. Acquisition, processing and use*. Amsterdam and Philadelphia: Benjamins, pp.127-147.

Schönefeld, D. 1999. Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics,* 4(1), pp.137-171.

Schonell, F.J., Meddleton, I.G. & Shaw, B.A. 1956. *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Segalowitz, S.J. & Lane, K.C. 2000. Lexical access of function versus content words. *Brain and Language,* 75, pp.376-389.

Shapiro, B.J. 1969. The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior,* 8, pp.248-251.

Sheskin, D.J. 2000. *Handbook of parametric and nonparametric statistical procedures*, Second Edition. Boca Raton: Chapman & Hall / CRC.

Sinclair, J. 1987. The nature of the evidence. In: Sinclair, J. ed. *Looking up*. London: Collins, pp.150-159.

Sinclair, J. 1991a. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. 1991b. Shared knowledge. In: Alatis, J.E. ed. *Georgetown University round table on languages and linguistics 1991 linguistics and language pedagogy: the state of the art*. Washington DC: Georgetown University Press, pp.489-500.

Sinclair, J. 1992. Trust the test: The implications are daunting. In Davies, M. & Ravelli, L. eds. *Recent Advances in Systemic linguistics*. London and New York: Pinter publishers, pp.5-19.

Sinclair, J. 1997. Corpus evidence in language description. In: Wichmann, A., Fligelstone, S., McEnery, T. & Knowles G. eds. *Teaching and Language Corpora*. Harlow: Longman, pp.27-50.

Sinclair, J. 2004. The lexical item. In: Sinclair, J. ed. *Trust the text. Language corpus and discourse*. London and New York: Routledge, pp.131-148.

Sinclair J. & Renouf, A. 1988. A lexical syllabus for language learning. In: Carter, R. & McCarthy, M. eds. *Vocabulary and language teaching*. London: Longman, pp.140-160.

Singleton, D. 2000. *Language and the lexicon*. London: Arnold.

Smadja, F. 1994. Retrieving collocations from text: Xtract. In: Armstrong, S. ed. *Using large corpora*. Cambridge, Mass: Bradford Book MIT Press, pp.143-177.

Söderman, T. 1993. Word associations of foreign language learners and native speakers-different response types and their relevance to lexical development. In Hammarberg, B. ed. *Problem, process and product in language learning*. Stockholm: Stockholm University, Department of Linguistics, pp.157-169.

Sosa, A.V. & MacFarlane, J. 2002. Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*. *Brain and Language,* 83, pp.227-236.

Stefanowitsch, A. & Gries, S.T. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), pp. 209-243.

Stemberger, J.P. & MacWhinney, B. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14(1), pp.17-26.

Stock, O., Slack, J. & Ortony, A. 1993. Building castles in the air. Some computational and theoretical issues in idiom comprehension. In: Cacciari, C. & Tabossi, P. eds. *Idioms: processing, structure and interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum, pp.229-247.

Stubbs, M. 1995a. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language,* 2(1), pp.23-55.

Stubbs, M. 1995b. Collocations and cultural connotations of common words. *Linguistics and Education,* 7, pp.379-390.

Stubbs, M. 2000. Using very large text collections to study semantic schemas: a research note. In: Heffer, C. & Saunston, H. eds. *Words in Context: A tribute to John Sinclair on his retirement. English language discourse analysis monograph 18. University of Birmingham,* CD ROM.

Stubbs, M. 2001. Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics,* 22(2), pp.149-172.

Stubbs, M. 2002a. On text and corpus analysis: A reply to Borsley and Ingham. *Lingua,* 112, pp.7-11.

Stubbs, M. 2002b. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7(2), pp.215-244.

Stubbs, M. 2002c. *Words and phrases. Corpus studies of lexical semantics.* Oxford: Blackwell.

Stubbs, M. & Barth, I. 2003. Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of language,* 10(1), pp.61-104.

Summers, D. 1988. The role of dictionaries in language learning. In Carter, R. & McCarthy, M. eds. *Vocabulary and language teaching*. London and New York: Longman, pp.111-125.

Summers, D. 1996. Computer lexicography: the importance of representativeness. In: Thomas, J. & Short, M. eds. *Using corpora for language research*. London: Longman, pp.260-266.

Sun, R., Merrill, E. & Peterson, T. 2001. From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science,* 25, pp.203-244.

Svartvik, J. 1993. Lexis in English language corpora. *Quarterly of Language, Literature and Culture,* 41(1), pp.15-30.

Svartvik, J. 1996. Corpora are becoming mainstream. In: Thomas, J. & Short, M. eds. *Using corpora for language research*. London: Longman, pp.3-13.

Swinney, D.A. & Cutler, A. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior,* 18, pp.523-534.

Takaie, H. 2002. A trap in corpus linguistics: The gap between corpus based analysis and intuition based analysis. *Language and computers,* 38, pp.111-130.

Taylor, S.E. 1982. The availability bias in perception and interaction. In: Kahneman, D., Slovic, P. & Tversky, A. eds. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, pp.190-200.

Thorndike, E.L. & Lorge, I. 1944. *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teacher's College, Columbia University.

Thorndike, E.L. & Lorge, I. 1952. The teacher's wordbook of 30,000 words. New York: Columbia University Press.

Tognini-Bonelli, E. 1993. Interpretive nodes in discourse. actual and actually. In: Baker, M., Francis, G. & Tognini-Bonelli, E. eds. *Text and technology. In honour of John Sinclair*. Amsterdam: John Benjamins, pp.193-212.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work.* Amsterdam: Benjamins.

Toth, J.P. 2000. Nonconscious forms of human memory. In: Tulving, E. & Craik, F.I.M. eds. *The Oxford handbook of memory.* Oxford: Oxford University Press, pp.245-261.

Toth, J.P. & Daniels, K.A. 2002. Effects of prior experience on judgments of normative word frequency: automatic bias and correction. *Journal of Memory and Language,* 46(4), pp.845-874.

Tryk, H.E. 1968. Subjective scaling of word frequency. *American Journal of Psychology,* 81, pp.170-177.

Tversky, A. & Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology,* 5, pp.207-232.

Tversky, A. & Kahneman, D. 1982. Judgment under uncertainty: heuristics and biases. In: Kahneman, D., Slovic, P. & Tversky, A. eds. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, pp.3-19.

Wänke, M., Schwaz, N. & Bless, H. 1995. The availability heuristic revisited: Experienced ease of retrieval in mundane frequency estimates. *Acta Psychologia,* 89, pp. 83-90.

West, M.P. 1953. *A General Service List of English Words*. London: Longmans.

Whitsitt, S. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics,* 10(3), pp.283-305.

Widdowson, H.G. 1989. Knowledge of language and ability for use. *Applied Linguistics,* 10(2), pp.128-137.

Widdowson, H.G. 1991. The description and prescription of language. In: Alatis, J.E. ed. *Georgetown University round table on languages and linguistics 1991 linguistics and language pedagogy: the state of the art*. Washington DC: Georgetown University Press, pp.11-24.

Widdowson, H.G. 2000. On the limitations of linguistics applied. *Applied Linguistics,* 21(1), pp.3-25.

Willis, D. 1990. *The Lexical syllabus*. London: Harper Collins.

Williams, G. 2002. In search of representativity in specialized corpora. *International Journal of Corpus Linguistics,* 7(1), pp.43-64.

Williams, J.N. 1992. Processing polysemous words in context: evidence for interrelated meanings. *Journal of Psycholinguistic Research*, 21(3), pp.193-218.

Winter, B & Reber, A.S. 1994. Implicit learning and the acquisition of natural languages. In Ellis, N. ed. *Implicit and explicit learning of languages*. London: Academic Press, pp.115-145.

Wolter, B. 2001. Comparing the L1 and the L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Learning*, 23, pp.41-69.

Wray, A. 1999. Formulaic Language in learners and native speakers. *Language Teaching,* 32, pp.213-231.

Wray, A. 2002. *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Zacks, R.T., Hasher, L. & Sanft, H. 1982. Automatic encoding of event frequency: further findings. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 8, pp.106-116.

Zughoul, M.R. & Abdul-Fattah, H. 2003. Translational collocational strategies of Arab learners of English: A study in lexical semantics. *Babel,* 49(1), pp.59-81.

Appendix 1 – The Formulas of MI and z-score

The BNC provides a z-score analysis and an MI score as well as providing raw frequency co-occurrence data for collocations, as described in chapter 5, section 2. Below a brief explanation of how the MI score and z-score are calculated is provided.

MI score (I)

$$I(x, y) = \log_2 \frac{P(x,y)}{P(x)\,P(y)}$$

Church & Hanks (1989) explain this calculation in the following way:

> Mutual information compares the probability of observing x and y together (the joint probability) with the probabilities of observing x and y independently (chance). If there is a genuine association between x and y, then the joint probability P (x,y) will be much larger than chance P (x) P (y), and consequently I (x,y) >> 0 (1989, p.77).

P(x) and P (y) are the observations of the words in the corpus. P (x,y) is calculated by counting the number of times that the two words co-occur, or co-occur within a certain window. Clear (1993, p.278) notes that MI is non-directional – i.e. it averages the association: there is only one score for the association no matter which is the node word. As Stubbs (1995a, p.35) comments, as a consequence of this, such a measure of association strength may be misleading, because it fails to show that it may be the case that one word *always* predicts the occurrence of another, e.g. *kith* predicts *kin*, but *not* the reverse, i.e. *kin* is not in an exclusive combination with *kith*.

z score (z)

The equation for the calculation of the z score is as follows:

$$z = \frac{O-E}{\sigma}$$

In this equation O = observed frequency of the word within the span, E = expected frequency of the same word (calculated by dividing the number of occurrences of the word with the number of words in the corpus) and σ = the standard deviation of the word – the square root of the number of tokens, multiplied by the probability of occurrence (p) (occurrences divided by words in the corpus) multiplied by 1- p (Barnbrook 1996, p.95). The statistic z is not uni-directional – the z-score for *blonde* as a node word and *hair* as a collocate is not the same as the z-score for *hair* as the node and *blonde* as a collocate.

Appendix 2 – Task Sheet for Experiment 1

Judging Frequency

People are generally good at noticing how often things happen; for example, they can estimate quite accurately how often they watch films. They are also good at ranking the frequency of words in the English language; for example, they can judge that *house* is a more common word than *ambulance*. I am interested in whether people are able to judge the relative frequency of 'pairs of words'.

Large collections of English language data (taken from newspapers, magazines, books, conversation etc.) are available to us, and these enable us to conduct language research. The BNC (British National Corpus) is an example of such a database. We can search this resource to see how often pairs of words (collocations) occur. Similarly, we can conduct 'exact phrase' searches using internet search engines, e.g. AltaVista.

In the task below I am interested in whether you can guess the relative frequencies of the various sets of collocations as they are given in BNC and AltaVista. Put a number 1 next to the pairing you think is the most frequent, 2 by the next most frequent, down to 8 for the least frequent pairing. Do each set separately.

Set 1	Order?	Set 2	Order?
...personal corruption...	___	...bad sport...	___
...personal number...	___	...bad news...	___
...personal problems...	___	...bad habit...	___
...personal initials...	___	...bad form...	___
...personal experience...	___	...bad miscalculation...	___
...personal disaster...	___	...bad headache...	___
...personal minute...	___	...bad egg...	___
...personal letters...	___	...bad luck...	___

Please Turn over

Set 3		Set 4	
	Order?		Order?
...personal quality...	___	...bad decision...	___
...personal computer...	___	...bad injury...	___
...personal life...	___	...bad reputation...	___
...personal liberty...	___	...bad idea...	___
...personal setback...	___	...bad figure...	___
...personal belongings...	___	...bad danger...	___
...personal impression...	___	...bad revelation...	___
...personal sorrow...	___	...bad terms...	___

Thank you for your help

Appendix 3 – Distributions in BNC and Brown Corpus for Adjective Stimulus Words, Experiment 2

Table A3, below, indicates the distribution of the stimulus words used in experiment 2 in the BNC (complete) and the Brown Corpus (Kučera and Francis 1967). Comments about range and Juilland D are made in chapter 2, section 2.1. It can be seen that all of the words are well distributed across the British and American corpora.

Table A3. Stimulus words, Experiment 2: Distributions in the BNC and the Brown corpora

	BNC		Kučera & Francis	
	Range (Max100)	Juilland D (Max 100)	Genre types (Max 15)	Samples (Max 500)
A. Different	100	95	15	181
B. Difficult	100	96	15	127
C. Full	100	98	15	166
D. Good	100	97	15	319
E. Great	100	97	15	291
F. Important	100	95	15	211
G. Large	100	95	15	214
H. Main	100	96	14	98
I. Old	100	95	15	256
J. Particular	100	93	14	112
K. Personal	100	94	15	109
L. Possible	100	95	15	226
M. Real	100	97	15	156
N. Recent	100	94	13	120
O. Similar	100	94	14	113
P. Small	100	97	15	243
Q. Special	100	96	14	155
R. Strong	100	97	14	133
S. Various	100	95	15	131
T. Young	100	96	15	190

BNC noun collocate raw frequency rankings and details on adjective stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

A. Different	B. Difficult	C. Full	D. Good	E. Great	F. Important	G. Large
1. ways (1213)	1. task (204)	1. time (592)	1. idea (1861)	1. deal (2673)	1. part (1048	1. number (1868)
2. types (1134)	2. time (139)	2. employment (481)	2. news (1194)	2. majority (389)	2. thing (632)	2. numbers (1251)
3. kinds (653)	3. question (118)	3. details (464)	3. time (878	3. success (370)	3. role (625)	3. part (560)
4. parts (524)	4. times (115)	4. range (417)	4. thing (831)	4. interest (322)	4. point (427)	4. scale (500)
5. way (505)	5. situation (101)	5. length (246)	5. deal (786)	5. importance (317)	5. factor (411)	5. proportion (423)
6. levels (432)	6. questions (88)	6. year (230)	6. job (745)	6. care (290)	6. aspect (263)	6. amounts (379)
7. people (417)	7. problem (77)	7. moon (189)	7. reason (738)	7. difficulty (260)	7. feature (225)	7. quantities (355)
8. things (407)	8. job (71)	8. potential (185)	8. evening (697)	8. war (257)	8.5. issue (212	8. amount (346)
9. kind (375)	9.5. problems (69)	10. advantage (166)	9. morning (657)	9. fun (238)	8.5. things (212)	9. extent (342)
10. groups (370)	9.5. thing (69)	10. board (166)	10. example (591)	10. man (189)	10. element (209)	10. areas (329)
11. forms (369)	11.5 decisions (68)	10. stop (166)	11. luck (511)	11. pleasure (187)	11. issues (205)	11. companies (239)
12. times (332)	11.5. decision (68)	12. day (163)	12. quality (476)	12. number (183)	12. question (166)	12. sums (229)
13. countries (300)	13.5. circs (66)	13.5. support (147)	13. practice (461)	13. hall (182)	13. source (165)	13. firms (182)
14. areas (283)	13.5. year (66)	13.5. use (147)	14. night (398)	14. help (168)	14. contribution (160)	14. area (169)
15. species (238)	15. position (64)	15. circle (143)	15. condition (382)	15. variety (167)	15. factors (145)	15. family (144)
16.5. aspects (212)	16. part (56)	16. force (138)	16. girl (358)	16. value (160)	16. implications (135)	16. group (137)
16.5. matter (212)	17. situations (53)	17. extent (135)	17. health (357)	17. power (158)	17. step (132)	17. measure (136)
18. places (211)	18. period (46)	18. back (131)	18. faith (346)	18. advantage (156)	18. questions (131)	18. majority (116)
19. approach (190)	19. cases (45)	19. colour (129)	19. way (345)	19. extent (141)	19. differences (126)	19. house (109)
20. colours (181)	20. conditions (42)	20. report (124)	20. chance (343)	20. powers (135)	20. aspects (122)	20. room (104)

BNC noun collocate raw frequency rankings and details on adjective stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

H. Main	I. Old	J. Particular	K. Personal	L. Possible	M. Real	N. Recent
1. road (680)	1. man (2358)	1. interest (377)	1. computer (654)	1. way (162)	1. world (679)	1. years (2777)
2. reason (379)	2. age (1261)	2. case (359)	2. computers (470)	2. exception (117)	2. life (568)	2. times (311)
3. line (375)	3. people (1164)	3. attention (329)	3. injury (334)	3. explanation (90)	3. terms (512)	3. months (304)
4. problem (301)	4. woman (785)	4. time (271)	4. experience (262)	4. ways (89)	4. thing (366)	4. research (244)
5. reasons (240)	5. lady (719)	5. area (227	5. life (171)	5. explanations (78)	5. problem (287)	5. work (241)
6. street (219)	6. days (567)	6. problem (215)	6. development (164)	6. use (72)	6. wages (211)	6. developments (226)
7. concern (218)	7. testament (515)	7. type (214)	7. relationships (163)	7.5. reasons (71)	7. reason (209)	7. weeks (220)
8. thing (217)	8. friend (457)	8. way (202	8. knowledge (145)	7.5. effects (71)	8. time (200)	8. study (209)
9. areas (214)	9. boy (397)	9. problems (183)	9. communication (130)	9. causes (64)	9. name (189)	9. studies (191)
10. aim (212)	10. friends (319)	10.5. circs (179)	10.5.responsibility (126)	10. alternative (60)	10. value (187)	10. survey (182)
11. point (208)	11. men (293)	10.5. form (179)	10.5. interest (126)	11. solutions (58)	11. sense (157)	11. past (179)
12.5. source (206)	12. school (284)	12. point (176)	12. service (118)	12.5. sources (57)	12. danger (152)	12. history (139)
12.5. points (206)	13. town (267)	13. importance (166)	13. qualities (111)	12.5. time (57)	13. problems (149)	13. changes (131)
14. purpose (180)	14. world (256)	14. kind (148	14. pension (109)	14. solution (56)	14. people (148)	14. report (127)
15.5. entrance (166)	15. house (252	15. reference (140)	15. care (103)	15. reason (55)	15. ale (132)	15.5. decades (114)
15.5. body (166)	16. master (199)	16. emphasis (133)	16.5. injuries (101)	16. role (54)	16.5. interest (129)	15.5. meeting (114)
17. types (164)	17. girl (187)	17. group (131)	16.5. assistant (101)	17. consequences (53)	16.5. power (129)	17. events (101)
18. features (155)	18. person (177)	18. needs (126)	18. contact (97)	18. means (48)	18. estate (116)	18. visit (91)
19. building (153)	19. ladies (170)	19. areas (123)	19. problems (88)	19. moment (47)	19. Work (97)	19. book (86)
20. parties (152)	20. system (132)	20. concern (118)	20. accident (84)	20. changes (46)	20. threat (92)	20. reports (72)

BNC noun collocate raw frequency rankings and details on adjective stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

O. Similar	P. Small	Q. Special	R. Strong	S. Various	T. Young
1. way (294)	1. number (925)	1. needs (630)	1. sense (209)	1. ways (432)	1. people (3613)
2. problems (146)	2. group (698)	2. interest (326)	2. support (152)	2. forms (334)	2. man (2667)
3. pattern (126)	3. businesses (438)	3. attention (220)	3. winds (144)	3. kinds (317)	3. children (1177)
4. results (125)	4. amount (428)	4. relationship (216)	4. position (124)	4. parts (308)	4. men (1162)
5. situation (102)	5. groups (392)	5. case (202)	5. case (118)	5. types (293)	5. woman (971)
6.5. position (89)	6. firms (367)	6. schools (173)	6. feeling (113)	6. aspects (216)	6. women (523)
6.5. effect (89)	7. proportion (326)	7. circumstances (167)	7. evidence (102)	7. stages (181)	7. girl (506)
8.5. fashion (81)	8. part (310)	8. offer (157)	8. wind (85)	8. reasons (164)	8. lady (411)
8.5. manner (81)	9. business (300)	9. branch (147)	9.5. opposition (83)	9. times (146)	9. girls (272)
10.5. lines (80)	10. children (294)	10. care (135)	9.5. views (83)	10. groups (127)	10. person (264)
10.5. circumstances (80)	11. town (290)	11. school (134)	11. emphasis (82)	11. things (119)	11. boy (195)
12. vein (75)	12. amounts (242)	12.5. events (133)	12. feelings (79)	12. points (117)	12. offenders (191)
13. reasons (69)	13. scale (240)	12.5 report (133)	13. links (70)	13. levels (101)	13. child (181)
14. age (64)	14.5. towns (218)	14. place (122)	14. man (65)	14. methods (90)	14. adults (143)
15.5. size (63)	14.5. boy (218)	15. meeting (119)	15. commitment (64)	15. people (89)	15. players (139)
15.5. approach (63)	16. area (197)	16. effects (117)	16.5. interest (61)	16. sources (88)	16. king (132)
17. terms (62)	17.5. minority (193)	17. education (113)	16.5. point (61)	17. factors (83)	17. couple (121)
18. sort (59)	17.5. numbers (193)	18. treatment (112)	18. influence (57)	18. elements (75)	18. boys (118)
19.5. cases (58)	19. companies (184)	19. occasions (110)	19. hand (56)	19. sizes (72)	19. ladies (114)
19.5. problem (58)	20. room (166)	20. features (100)	20. argument (55)	20. countries (68)	20. lad (105)

Appendix 5 – Task Sheet for Experiment 2

Words and their partners

Words often have partners, i.e. certain words often go together. A computer program (called the BNC) has been invented that can look at thousands of books, newspapers and conversations, and count up how often different words occur together. Now we can take a word of English and see which word most often follows it. I want to see if you can guess what the computer program has discovered.

Below is a list of 20 common adjectives. Can you guess, for each adjective, which noun the computer has found to be the one that most frequently occurs straight after that adjective? Please write only **ONE** answer on each line.

Try to complete the task quickly

ADJECTIVE	**Most frequent NOUN partner?**	*ADJECTIVE*	**Most frequent NOUN partner?**
Example		*Example*	
i) low	*level*	*ii) hard*	*work*
1. different	_____	**11. possible**	_____
2. strong	_____	**12. small**	_____
3. full	_____	**13. real**	_____
4. great	_____	**14. recent**	_____
5. important	_____	**15. similar**	_____
6. large	_____	**16. special**	_____
7. main	_____	**17. difficult**	_____
8. old	_____	**18. good**	_____
9. particular	_____	**19. various**	_____
10. personal	_____	**20. young**	_____

Appendix 6 – NS and NNS Responses for Experiment 2

In the tables that follow, all of the responses of the subjects to the adjective stimuli of experiment 2 are listed. The central column identifies the responses which were common to the two groups; the number in brackets in that column indicates first the number of native speaker responses and then the number of non-native speakers who provided that particular response. It should be noted that the table differentiates the responses which were within the 20 most frequent collocates of the stimulus word in the BNC, and those which were >20.

A. Different

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
	In top 20	
-	Way (1+1), People (5 + 1)	Ways (2), Things, Matter
	Outside top 20	
Views, Causes, Reason (2) Dates, Style, Jobs, Answers, Thing, Notions, Ideas, Circumstances, Perspective	Opinion (1+1)	Type (2), Issues, Locations, Idea, Subjects, Opinions, Issue, Reasons, Part, Person, Question, Area

B. Difficult

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
	In top 20	
Time (3), Problems, Times (3)	Task (3 + 7), Decision (1+1), Job (1 + 3), Problem (5 + 3), Situation (1+2)	-
	Outside top 20	
Proposition, Choice	-	Class, Subject, Question, Work

C. Full

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Details, Range (2)	Time (3 + 5), Day (3+1)	-
Outside top 20		
Glass, House (3), Head, Impact, Complement (2), Stomach (2), Story	-	Amount (2), Load, Paper, Fledged, Tank (3), Mark, Fledge, Proof, Capacity (2), Grade

D. Good

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Time, News (2)	Idea (6+2), Luck (2+2)	Morning (2), Example, Job
Outside top 20		
Father, Food, Times, Meal, Understanding	Person (1+2), Boy (1+3), Work (1+2), Day (1+1)	Behaviour (2), Nature, Design

E. Great

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Deal	-	Man (2), Pleasure, Help, Value
Outside top 20		
Amount, Nation, Men (3), Time (6), Expectation, Spectacle, Looks, Britain, News	Person (1+1), Idea (2+4)	Job (4), Work (2), Things, Adventure, Achievement, Day

F. Important

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
-	Issue (3+1), Thing (1+2)	Part
Outside top 20		
Information (3), Facts, Decision, Work, Points, Date, Meeting	Person (4+1), Idea (1+2), News (2+1)	Task (6), Concept (2), Matter (2), Things, Subject

G. Large

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Sums, Amount (2), Quantities, Areas, Scale	Area (1+2), Room (1+1)	House, Part
Outside top 20		
Towns, Quantity (2), Lady, Bonus, Markets, Cities, Truck, Ears, Meal	Building (1+1), Size (1+3)	Thing, Reservoir, Section, Car, Class room, Income, Task, Structure, Office, Tree, Image

H. Main

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Problem, Body, Purpose	Street (3+4), Point (1+1)	Reason, Road
Outside top 20		
Frame (2), Component, Argument, Theme, Event (3)	Idea (5+4)	Difference, Subject, Purpose, Objective (2), Task (2), Part, Issue

I. Old

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
System, Woman (2) People (2), Friend, Men	Man (4+10), Person (1+1), Age (1+1)	House
Outside top 20		
News, Furniture, Machine, Clothes, Hat, Model, Women	-	Fashion (2), Method, Idea, Past, Subject, Factory

J. Particular

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Group, Area, Time	Case (2+1), Problem (1+2), Point (1+1)	Interest (2), Way
Outside top 20		
Thing, Types, Detail, Instances, Items, Colour, Person, Idea	Situation (2+1), Information (1+1), Reason (2+1)	Issue (2), Name, Subject (2), Thing (3), One, Matter

K. Personal

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
-	-	Problems
Outside top 20		
Mail, Freedom, Details, Idea, Remarks, Differences, Services, Habits, Opinion, Lives, Feelings, Goals, Data, Trainer	Matter (1+3), Feeling (2+1) Problem (1+1), Information (2+1)	Issue (3), Preference, Interests, Relationship, Affair, Relations, Items, Affairs (2), Property, Space

L. Possible

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Explanation (2), Effects	Solution (1+4), Reason (1+2), Way (1+2)	Solutions
Outside top 20		
Chance, Advantages, Meanings, Answers (3), Path	Outcomes (1+1), Outcome (2+1), Cause (1+3), Answer (3+1)	Remark, Things, Job, Choice, Opening

M. Real

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Ale	Life (2+1), Time (7+1), World (1+4), Estate (1+1), Problem (1+1), Thing (1+2)	Work (3)
Outside top 20		
Situation (2), Events, Livewire, Man, Food		Image, Case, Number, Evidence, Trouble, Things, Effect

N. Recent

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Times, History	Events (6+1)	-
Outside top 20		
Development (2), Letter, Happening (2), Data, Memories	News (1+2), Event (4+3)	Time (2), Occasion, Day Advances, Finding, Advancements, Edition, Encounter, Job(2), Days, Information, Accident

O. Similar

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Situation (2) Circumstances	-	Cases
Outside top 20		
Remarks, Types, Case, Views, Instances, Occurrence, Ideas, States, Type, Features, Answer Meaning, experience	Ideas (3+2), Thing (1+4)	Job, Example, Properties, Things (3), System, Matter, Solution, Behaviour, Proof, Comparison, Topic

P. Small

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Children, Amount, Amounts, Town (2)	Boy (1+1)	Part
Outside top 20		
Tree, Point, Outcome, Creatures, Fry, Child, Time, Difference, Feet, Details, Person, Car	Matter (1+1), Mind (1+2)	Chance, Size (2), Thing (3) Item (2), Class room, Box, Things, World, Portion, Quantity, Toy

Q. Special

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Effects, Circumstances	Treatment (1+1), Case (1+3), Offer (1+1)	-
Outside top 20		
Problem, Status, Purpose, Occasion (4), Event (5), Considerations, Privileges	Person (1+1)	Gift, Group, Area, Task (3) Assignment, Issue, Thing, Relation, Variety, Subject, Job, Day

R. Strong

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Winds (2), Argument	Man (2+7), Feeling (1+3)	Evidence
Outside top 20		
Language, Wind, Body, Smell (2), Arm (2), Character, Taste, Odour, Influence, Coffee	Personality (2+3)	Relation, Relevance, Background, Boy, Person Level

S. Various

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Factors, Kinds, People (2)	Ways (1+2), Things (2+7), Aspects (1+1), Reasons (1+1)	-
Outside top 20		
Issues, Places, Answers, Outcomes, Opinions, Duties	Items (3+1), Possibilities (2+1)	Topics, Activities, Courses, Subjects, Issues (2), Ideas

T. Young

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Girl (2)	Boy (1+3), Man (5+7), People (7+1), Person (1+1)	Lady, Woman
Outside top 20		
Generation, Kid, Adult, Men		Faculty, Age (2), Tree, Engineers, Talent

Appendix 7 – Task Sheet for Experiment 3

You versus the computer!

The British National Corpus (BNC) is a large bank of language data collected from books, newspapers, conversations etc. When a computer searches the BNC data it can work out the most frequent collocate of a word. I would like to compare your judgements (as a language teacher) with the computer's findings.

Below is a list of twenty common adjectives, together with three collocation patterns for each word. Put a tick in the box next to the collocation which you think represents the most frequent collocation in the BNC data.

Example

1. new
...new year... ☑
...new generation... ☐
...new lease... ☐

A. Similar
...similar lines... ☐
...similar problem... ☐
...similar way... ☐

B. Great
...great man... ☐
...great deal... ☐
...great powers... ☐

C. Important
...important part... ☐
...important aspects... ☐
...important element... ☐

D. Particular
..particular circumstances ☐
...particular interest... ☐
...particular concern... ☐

E. Possible
...possible alternative... ☐
...possible way... ☐
...possible changes... ☐

F. Old
...old friends... ☐
...old man... ☐
...old system... ☐

G. Large
...large areas... ☐
...large number... ☐
...large room... ☐

H. Real
...real name... ☐
...real world... ☐
...real estate... ☐

I. Various
...various ways... ☐
...various groups... ☐
...various countries... ☐

J. Young
...young person... ☐
...young people... ☐
...young lad... ☐

K. Strong
...strong views... ☐
...strong sense... ☐
...strong argument... ☐

L. Good
...good example... ☐
...good way... ☐
...good idea... ☐

M. Special
...special care... ☐
...special needs... ☐
...special features... ☐

N. Full
...full time... ☐
...full board... ☐
...full report... ☐

O. Different
...different ways... ☐
...different groups... ☐
...different colours... ☐

P. Personal
...personal accident... ☐
...personal responsibility ☐
...personal computer... ☐

Q. Difficult
...difficult thing... ☐
...difficult task... ☐
...difficult conditions... ☐

R. Recent
...recent survey... ☐
...recent years... ☐
...recent report... ☐

S. Small
...small children... ☐
...small number... ☐
...small companies... ☐

T. Main
...main aim... ☐
...main road... ☐
...main features... ☐

Appendix 8 – Distributions in BNC and Brown Corpus for Noun Stimulus Words, Experiment 4

Table A8, below, indicates the distribution of the stimulus words used in experiment 4 in the BNC (complete) and the Brown Corpus (Kučera and Francis 1967). Comments about range and Juilland D are made in chapter 2, section 2.1. It can be seen that all of the words are well distributed across the British and American corpora.

Table A8. Stimulus words, Experiment 4: Distributions in the BNC and the Brown corpora

	BNC		Kučera & Francis	
	Range (Max 100)	Juilland D (Max 100)	Genre types (Max 15)	Samples (Max 500)
A. Amount	100	87	15	99
B. Approach	100	93	15	85
C. Basis	100	93	14	112
D. Chance	100	95	15	99
E. Details	100	93	12	49
F. Evidence	100	93	14	121
G. Fact	100	96	15	233
H. Future	100	94	15	134
I. Importance	100	93	13	79
J. Information	100	92	14	155
K. Kind	100	96	15	186
L. Matter	100	95	15	196
M. Moment	100	91	14	151
N. Part	100	97	15	301
O. Problem	100	96	15	154
P. Purpose	100	92	14	107
Q. Range	100	93	15	89
R. Role	100	93	13	64
S. Sense	100	95	15	163
T. Word	100	94	15	153

Note that the check is on the singular noun, not the lemma figures. Note also that Kučera & Francis do not provide a part of speech measure – so their figures will include the other forms of the word (i.e. verbs and adjectives).

BNC adjective collocate raw frequency rankings and details on noun stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

A. Amount	B. Approach	C. Basis	D. Chance	E. Details	F. Evidence	G. Fact
1. certain (769)	1. new (219)	1. regular (436)	1. good (343)	1. further (742)	1. further (276)	1. actual (226)
2. small (428)	2. different (190)	2. daily (166)	2. better (230)	2. full (464)	2. empirical (167)	2. very (156)
3. large (346)	3. alternative (106)	3.5. day-to-day (85)	3. last (169)	3. other (79)	3. other (161)	3. mere (105)
4. considerable (342)	4. general (83)	3.5. part-time (85)	4. best (157)	4. personal (67)	4. new (160)	4. simple (89)
5. total (245)	5. positive (76)	5. permanent (76)	5. only (127)	5. precise (46)	5. clear (159)	5. important (58)
6. fair (234)	6. similar (63)	6. annual (71)	6. second (96)	6. technical (41)	6. sufficient (140)	6. sad (52)
7. enormous (190)	7. systematic (60)	7. temporary (63)	7. real (62)	7. specific (38)	7.5. medical (126)	7. plain (45)
8. substantial (135)	8. flexible (58)	8. legal (61)	8. fair (61)	8. brief (35)	7.5. direct (126)	8. historical (44)
9. vast (120)	9. second (57)	9. individual (60)	9. great (57)	9. fine (33)	9. scientific (114)	9. obvious (33)
10. maximum (111)	10. traditional (56)	10. regional (55)	10. greater (55)	10. financial (33)	10. available (108)	10. well-known (33)
11. full (105)	12. pragmatic (47)	11.5. rational (53)	11. first (54)	11. small (30)	11. strong (102)	11. interesting (28)
12. tremendous (103)	12. scientific (47)	11.5. national (53)	12. big (45)	12. final (28)	12. conclusive (93)	12. hard (21)
13. limited (99)	12. final (47)	13. theoretical (52)	14. realistic (35)	13.5. biographical (26)	13. good (90)	13. curious (20)
14. significant (87)	14. cautious (46)	14. commercial (52)	14. pure (35)	13.5.exact (26)	14.5. ample (85)	14. known (19)
15. right (85)	15. first (45)	15.5. weekly (51)	14. reasonable (35)	15. important (25)	14.5. historical (85)	15.5. undeniable (18)
16. huge (78)	16. direct (44)	15.5. sound (51)	16. fat (34)	16. certain (24)	16.5. circumstantial (83)	15.5. basic (18)
17. minimum (75)	17.5. fresh (43)	17. monthly (45)	17.5. fighting (29)	17.5. minor (21)	16.5. hard (83)	17.5. significant (17)
18. reasonable (65)	17.5. best (43)	18.5. ad hoc (44)	17.5. equal (29)	17.5.relevant (21)	18. experimental (80)	17.5. central (17)
19. fixed (60)	19. integrated (42)	18.5. one-to-one (44)	19. outside (27)	19. intimate (20)	19. archeological (75)	19.5 scientific (16)
20. increasing (50)	20. whole (41)	20. full-time (39)	20. main (22)	20. written (19)	20. oral (75)	19.5. established (16)

BNC adjective collocate raw frequency rankings and details on noun stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

H. Future	I. Importance	J. Information	K. Kind	L. Matter	M. Moment	N. Part
1. near (585)	1. great (317)	1. further (1131)	1. different (375)	1. different (207)	1. last (270)	1. important (1047)
2. forseeable (299)	2. relative (188)	2. new (304)	2. new (216)	2. organic (156)	2. very (162)	2. integral (619)
3. immediate (126)	3. particular (166)	3. detailed (274)	3. other (153)	3. simple (145)	3. right (136)	3. first (603)
4. long-term (91)	4. crucial (135)	4.5. additional (257)	4. particular (148)	4. serious (95)	4. long (106)	4. large (560)
5. political (74)	5. paramount (122)	4.5. relevant (257)	5. right (137)	5. important (83)	5. brief (98)	5. major (438)
6. distant (71)	6. considerable (119)	6. useful (226)	6. certain (95)	6. other (71)	6. particular (89)	6. second (380)
7. bright (65)	7. greater (119)	7. confidential (181)	7. special (92)	7. whole (53)	7. next (87)	7. essential (332)
8. uncertain (62)	8. vital (111)	8. financial (144)	8. only (43)	8. easy (44)	8. present (77)	8. only (326)
9. not-too-distant (57)	9. central (83)	9. inside (138)	9. wrong (42)	9. laughing (43)	9. first (71)	9. early (319)
10. better (49)	10. utmost (78)	10. general (117)	10. second (36)	10. small (39)	10. given (60)	10. small (310)
11. distant (36)	11. prime (76)	11. following (111)	11. worst (34)	11.5. personal (35)	11. precise (55)	11. greater (254)
12. whole (28)	12. fundamental(74)	12. available (106)	12. third (31)	11.5. particular (35)	12. possible (47)	12. latter (244)
13. great (26)	13. greatest (67)	13. private (100)	13.5. best (20)	13. private (26)	13. crucial (41)	13. vital (222)
14. secure (25)	14. major (64)	14. factual (99)	13.5. general (20)	14. solid (25)	14. critical (38)	14. best (207)
15. new (23)	15. growing (61)	15. basic (97)	15. first (20)	15. grey (23)	15. fleeting (37)	15. significant (184)
16. economic (20)	16.5. strategic (57)	16. up-to-date (96)	16.5. strange (17)	16. complex (22)	16. appropriate (32)	16. active (181)
17. brighter (19)	16.5. secondary (57)	17. full (95)	16.5. human (17)	17. living (17)	17. single (28)	17. main (152)
18. bleak (16)	18. increasing (56)	18. technical (93)	18. usual (15)	18.5. delicate (15)	18. wrong (26)	18. upper (145)
19. possible (15)	19. national (55)	19. electronic (91)	19. traditional (13)	18.5. printed (15)	19. worst (25)	19. substantial (139)
20. indefinite (11)	20. practical (53)	20. valuable (85)	20. funny (12)	20. straightforward (14)	20. opportune (23)	20. lower (134)

BNC adjective collocate raw frequency rankings and details on noun stimuli (rank no. is to the left of the word; no. of collocation instances is to the right, in brackets)

O. Problem	P. Purpose	Q. Range	R. Role	S. Sense	T. Word
1. major (401)	1. general (183)	1. wide (2743)	1. important (625)	1. common (1125)	2. last (308)
2. main (301)	2. main (180)	2. whole (658)	2. major (274)	2. good (288)	2. single (150)
3. real (287)	3.5. multi (91)	3. full (422)	3. key (267)	3. strong (209)	3. right (109)
4. only (248)	3.5. primary (91)	4. long (300)	4. leading (239)	4. real (157)	4. written (105)
5. serious (217)	5. particular (86)	5. new (281)	5. vital (218)	5. general (96)	5. spoken (101)
6. particular (215)	6. common (83)	6. close (214)	6. new (208)	6. great (95)	6.5. good (87)
7. big (125)	7. specific (73)	7. broad (159)	7. central (194)	7. economic (72)	6.5. key (87)
8. biggest (112)	8. real (70)	8. free (136)	8. active (174)	8. literal (66)	8. first (79)
9. social (85)	9. sole (67)	9. short (123)	9. crucial (169)	9.5. strict (64)	9. final (77)
10. second (83)	10. useful (65)	10. limited (117)	10. significant (134)	9.5. broad (64)	10. new (70)
11. further (82)	11. special (60)	11. vast (95)	11. political (76)	11. new (61)	11. English (69)
12.5. common (77)	12. original (55)	12. narrow (93)	12. traditional (70)	12. clear (59)	12. Greek (50)
12.5. difficult (77)	13. whole (53)	13. extensive (84)	13. dominant (66)	13. deep (58)	13. better (47)
14. fundamental (71)	14. dual (47)	14.5. normal (83)	14. social (65)	14. sixth (57)	14. wrong (45)
15. first (70)	15. other (46)	14.5. large (83)	15. future (63)	15. greater (56)	15. printed (41)
16. basic (67)	16. special (42)	16. comprehensive (81)	16. dual (61)	16. 5. false (55)	16. dirty (40)
17. great (66)	17. social (37)	17. complete (65)	17.5. prominent (59)	16.5 true (55)	17. very (37)
18.5. central (65)	18. present (32)	18. greater (50)	17.5. greater (59)	18. widest (52)	18. quick (33)
18.5. other (65)	19. different (27)	19. diverse (48)	19. possible (54)	19. narrow (49)	19. particular (32)
20. growing (60)	20. true (26)	20. huge (45)	20. supporting (53)	20. broadest (48)	20. quiet (31)

Appendix 10 – Task Sheet for Experiment 4

Words and their partners

Words often have partners, i.e. certain words often go together. We can see what these common partners are by referring to large collections of English language data (taken from newspapers, magazines, books, conversation etc.). The BNC (British National Corpus) is a computer database of language data. We can search this resource to see how often pairs of words occur together. I want to see if you can guess what the computer program has discovered.

Below is a list of 20 common nouns. Can you guess, for each noun, which **adjective** the BNC has found to be the one that **most frequently occurs** immediately before that noun? Please write only **ONE** answer on each line. For example, for the noun '**age**' you might guess that the most frequent adjective to occur in front of it is '**old**'. For '**quality**' you might guess '**high**'. These adjectives are the most frequent adjective partners of these nouns according to the BNC.

Note: Please remember that your response should be an **adjective** (e.g. **old**). Your response can also be a comparative form of an adjective, e.g. **older**, or a superlative form e.g. **oldest**. Sometimes you may want to provide an **adjective phrase** if you think that it is very frequent, e.g. '**knock-on**' (as in 'knock-on effect') or '**half-hearted**' (as in 'half-hearted attempt'). **This is acceptable.** If you want to use an adjective / adjective phrase more than once you may do so.

Please note that the following words are not permissible: **my, your** etc.; **this, that, any, another, one, more, every, much, such, few, all, little, own, some, enough, each, same.**

Try to complete the task quickly

Most frequent ADJECTIVE partner?	NOUN	Most frequent ADJECTIVE partner?	NOUN
E.g. *old*	*age*	*E.g.* *high*	*quality*
1. _____	**matter**	11. _____	**purpose**
2. _____	**range**	12. _____	**future**
3. _____	**sense**	13. _____	**evidence**
4. _____	**chance**	14. _____	**kind**
5. _____	**problem**	15. _____	**moment**
6. _____	**word**	16. _____	**basis**
7. _____	**approach**	17. _____	**details**
8. _____	**fact**	18. _____	**information**
9. _____	**importance**	19. _____	**role**
10. _____	**part**	20. _____	**amount**

Appendix 11 – NS and NNS Responses for Experiment 4

In the tables that follow, all of the responses of the subjects to the noun stimuli of experiment 4 are listed. The central column identifies the responses which were common to the two groups; the number in brackets in that column indicates first the number of native speaker responses and then the number of non-native speakers who provided that particular response. It should be noted that the table differentiates the responses which were within the 20 most frequent collocates of the stimulus word in the BNC, and those which were >20.

A. Amount

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Small, Total, Significant	Large (11+6), Right (1+1), Huge (1+2)	Fair, Full
Outside top 20		
Correct (2), Low, known	-	Good (3), Great (2), Big, Appreciable, High

B. Approach

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
New (6), Scientific, Final Cautious, Fresh	-	Different, Direct, Best (2)
Outside top 20		
Standard, Common-sense, Simple, Basic, Slow, Rational	Close, Practical, Right (1 +3), Wrong	Good (4), Correct (3), Clear (2), Quick

C. Basis

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Regular (2), Theoretical	Sound (5+1)	-
Outside top 20		
Firm (4), Frequent, First, Broad, Formal	Solid (3+2), Strong (1+4)	Clear (2), Good (2), Important, Common, Large, Extra, Similar, General, Correct, Fair, Real

D. Chance

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Second (2), Fat (2), Fair, Great, Big	Good (1+11) , Big (1+1) Last (7+2)	Better, Best
Outside top 20		
Fresh, Half, Slim	Lucky (1+1)	High, Small, Single

E. Details

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Precise, Small (3), Final, Relevant	Full (2 +1), Fine (1+3), Important (3+1)	-
Outside top 20		
Telling, Tiny, Insignificant, Unpleasant, Intricate, Essential	Boring	Minute (3), Clear (2), Complete (2), Right, Correct, Available, Good, Great, Accurate

F. Evidence

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
New, Conclusive, Circumstantial	Clear (4 +7), Strong (2+5), Hard (2+2)	Good
Outside top 20		
Corroborating, Vital, Hidden, Trustworthy, Admissible, Fresh, False Sound	Real	Great, Solid, Best, Meaningful

G. Fact

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Simple, Plain, Well-known (4), Interesting, Basic	Actual (1+1), Important (2 +1), Hard (1+1), Known (1+1)	Very, Mere
Outside top 20		
Valid, Bare, Morbid, Salient, Little-known, Indisputable	Real (1+2)	Clear (4), True (2), Strong, Bitter, Main, New, Accurate, Sound

H. Future

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Foreseeable, Great	Near (12 +13), Bright (5+4)	-
Outside top 20		
Rosy	-	Good, Brilliant, Nice

I. Importance

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Major (2), Vital (2), Prime (2), Utmost	Great (8+2)	Relative, Particular
Outside top 20		
Extreme, Limited	Real (1+1)	High (5), Significant (3), Large, Wide, Proven, True, Minor, Striking

J. Information

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Relevant	New (1+2), General (1+1), Useful (4+2)	Detailed
Outside top 20		
Secret, Common, Right, False, Vital, Specific	Important (5+3), True (1+1), Correct (1+1)	Good (3), Excessive, Accurate, Clear, Recent, Large, Necessary

K. Kind

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Different (4), Particular, Wrong, Human	Right (3+2), Best (2+2), Special (1+2)	Second
Outside top 20		
Usual (2), Unusual, Small	Rare (2+1), Good (1+3)	Common (2), Unique (2), Main, Bad, Weak, Fine, Durable

L. Matter

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Simple	Different (1+1), Serious (4+2), Important (6+5) Grey (3+1)	Easy
Outside top 20		
Atomic, Sensitive, Vital, Prime, Dark		Real (2), Clear, Hanging, Good, True, Complicated, Relevant, Low, New

M. Moment

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Long, Present, Opportune (2)	Last (2+3), Right (2+1), Brief (1+1)	Very
Outside top 20		
Decisive (2), Frightening, Magic, Big, Tender, Golden, Defining, Special, Small	Difficult (1+1)	Happy (3), Short (3), Lovely, Intimate, Nice, Meaningful, Precious, Fast Current

N. Part

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Integral, Vital	Important (2+2), Large (3+1), Small (5+3), Significant (1+1), Main (3+1)	First (3), Major, Best Essential
Outside top 20		
Final, Expensive	Spare (2+1)	Easy, Big, Great, Separate, Difficult

O. Problem

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Main, Particular	Big (3+4), Serious (4+3), Difficult (5+6), Common (2+1)	Major
Outside top 20		
Tricky, Unsolvable, Bad	-	Simple, Hard, Small, Substantial, Easy

P. Purpose

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Multi (3), Specific, Sole Useful	Main (5+2), Real (3+1)	General (2), Special True
Outside top 20		
Chief, Supposed	Clear (3+1), Right (1+1)	Good (4), Important (3), Strange, Basic, Noble Unique

Q. Range

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Free (2), Large	Wide (13+9), Long (3+4)	Full (2), Broad
Outside top 20		
-	High	Specific, Common

R. Role

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Key (2), Leading (3) Significant (2), Prominent	Important (1+8), Major (1+1)	Vital
Outside top 20		
Starring, Right, Huge, Assigned, Changing, Decisive	Main (3+2), Primary (1+1)	Big (2), Great, Best, Certain, Good, Common

S. Sense

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Real, Sixth	Common (13+13) Good (3+2)	Clear
Outside top 20		
Quick, Complete	-	Legitimate, Sharp, Meaningful

T. Word

NATIVE SPEAKERS	NS and NNS	NON-NATIVE SPEAKERS
In top 20		
Last (5), Final, New, Dirty	Good (1+1), Right (5+1)	Single, Key, First, Better
Outside top 20		
Precise, Appropriate, Important, Strong, Perfect, Common		Common (2), Different (2) Long (2), Clear, Lovely, Nice, Correct, Simple, Short, Easy, Small

Appendix 12 – Task Sheet for Experiment 5

Words and their partners

Words often have partners, i.e. certain words often go together. We can see what these common partners are by referring to large collections of English language data (taken from newspapers, magazines, books, conversation etc.). The BNC (British National Corpus) is a computer database of language data. We can search this resource to see how often pairs of words occur together. I want to see if you can guess what the computer program has discovered.

Below is a set of twenty common **nouns**, together with three adjective partners for each noun. Put a tick in the box next to the **most frequent combination** *in your opinion*.

Example

1. result

...satisfactory result ... ☐
...direct result ... ☑
...overall result ... ☐

A. role

...important role... ☐
...significant role... ☐
...supporting role... ☐

B. matter

...small matter... ☐
...different matter... ☐
...straightforward matter.. ☐

C. evidence

...further evidence ... ☐
...available evidence... ☐
...oral evidence... ☐

D. word

...new word... ☐
...particular word... ☐
...last word... ☐

E. part

...important part... ☐
...lower part... ☐
...small part... ☐

F. future

...better future ... ☐
...indefinite future... ☐
...near future... ☐

G. approach

...new approach... ☐
...traditional approach... ☐
...whole approach... ☐

H. purpose

...true purpose... ☐
...special purpose... ☐
...general purpose... ☐

I. information

...valuable information... ☐
...further information... ☐
...general information... ☐

J. chance

...good chance... ☐
...greater chance... ☐
...main chance... ☐

K. problem

...growing problem... ☐
...further problem... ☐
...major problem... ☐

L. importance

...practical importance... ☐
...great importance... ☐
...utmost importance ... ☐

M. basis

...casual basis... ☐
...regional basis... ☐
...regular basis... ☐

N. range

...huge range... ☐
...wide range... ☐
...limited range... ☐

O. kind

...funny kind... ☐
...different kind... ☐
...second kind... ☐

P. details

...written details... ☐
...financial details... ☐
...further details... ☐

Q. sense

...broad sense... ☐
...common sense... ☐
...broadest sense... ☐

R. moment

...last moment... ☐
...given moment... ☐
...opportune moment... ☐

S. fact

...actual fact... ☐
...established fact... ☐
...well-known fact ... ☐

T. amount

...certain amount... ☐
...maximum amount... ☐
...increasing amount... ☐