The development of a corpus-informed list of formulaic sequences for language pedagogy

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

Ron Martinez, BA, MSc.

December 2011

Dedicated to the memory of Marcia Trinidad Martinez

(October 14, 1939 – May 18, 2011)

Abstract

Discussion around the importance and prevalence of multiword expressions in the lexicon and the teaching of vocabulary has existed for a number of years in applied linguistics (e.g. Irujo, 1986; Pawley and Syder, 1983; Sinclair, 1987; Wray, 2002). While there seems to be a general agreement among scholars that formulaic language should feature in language learning and, perhaps to a lesser extent, language testing, there appears to be rather less agreement when it comes to how to select and/or prioritize specific items for inclusion. One criterion for selection which has been used often for vocabulary items of single words is frequency (i.e. how relatively common a word is), data for which can be consulted using various frequency lists that have long existed and are in the public domain, such as the General Service List (West, 1953). However, to date, no list of formulaic language that could be considered comparable to the General Service List in terms of intended use and relevance to language instruction has been attempted. The work presented in the present thesis aims to address this lack. The thesis first presents the need for such a list, and then describes the methodology employed by the researcher to ultimately produce a frequency-informed and pedagogically-relevant list of multiword expressions that can be used in conjunction with existing lists single orthographic words to help inform such instruments of L2 pedagogy as language textbooks and language tests, entitled the PHRASal Expressions List, or PHRASE List. To that end, two projects are also presented in the thesis which exemplify ways in which the list may be usefully employed. The first is a research validation exercise carried out in collaboration with the English Profile project in order to compare the phraseological component of the English Profile Wordlist to the expressions in the PHRASE List. The second project presents the development and validation of a kind of vocabulary test that samples from the PHRASE List, and which is intended to be used to supplement knowledge assessed in existing tests of single orthographic words, such as the Vocabulary Size Test (Nation & Beglar, 2007).

Acknowledgements

First and foremost, my deepest gratitude goes to my thesis supervisor, Norbert Schmitt. His tireless dedication and guidance should serve as an example to researchers and professors everywhere, and as an inspiration to me, personally. Michael McCarthy also provided invaluable supervisory support, especially during the early stages of the research, and I am very grateful for his immensely helpful feedback as well. I would also like to thank Bhatia Laufer, Nick Ellis, Robert Waring, Alison Wray, Barry O'Sullivan, John de Jong, Annette Capel and Tom Cobb for various comments and insights provided during the writing of the thesis which doubtlessly contributed to it becoming a better piece of research than it would have been otherwise.

Finally, a very special thanks to past and current colleagues at the University of Nottingham, especially those belonging to the Vocabulary Reading Group. I have been fortunate to have people like Irina Dahlmann, Anna Siyanova, Ana Pellicer Sanchez, Phoebe Lin, Suhad Sonbul, Phil Durrant, Hilde van Zeeland, Pawel Szudarski, and Kholood Saigh as my teachers as well as my peers during my time at Nottingham, and the ways in which the quality of the work produced herein has benefited from their knowledge and character cannot be quantified.

Contents

List of tables

List of figures

PAGE
NUMBERING
AS
ORIGINAL

Chapter 1 – Introduction

In addressing concerns about his ability to communicate effectively with his players, it was reported in *The Times* (UK)[1] that the manager of England's national football team, Italian Fabio Capello, claimed he only needed to know "100 words of English" to do his job. Unreasonable though that figure may sound to some, he perhaps was not as far off base as the press made his estimate appear. First of all, research would support his assertion that the question of 'how many words to I need to know' is not answerable by an absolute number but through the follow-up question of 'what do you need to do' – which is perhaps a particularly relevant question when it comes to applied linguistics and second language education. After all, L2 learners often only need to use the target language in a much narrower range of situations and contexts than native speakers, particularly if they are not living in the country in which the target language is spoken (Kirkpatrick, 2007; Pennycook, 2001; Seidlhofer, 2005). Furthermore, although surely 100 words would not suffice to do much in any language, there has been mounting evidence over the years that while there are hundreds of thousands of words in a language like English, even native speakers only really use and come into contact with a fraction of those words on a day to day basis (Nation, 2001; O'Keeffe, McCarthy & Carter, 2007). A large part of what the present thesis aims to address is the nature of that 'fraction' that is supposedly used more than the rest.

In Chapter 2, that issue will be explored in some depth, and will involve questioning the very concept of 'word', and some assumptions made about the words that are claimed to be the most important ones to know. Specifically, it will be argued that while it is probably true that we do not need to use all of the

[1] March 29th, 2011.

words we know (or at least recognize) in order to communicate, much of the existing research regarding frequent words needs to be revisited in the light of what is now known about words that recombine to form their own lexical items. For example, one could argue that a word like *take* is one of the most common and basic words in English, and therefore it is a word that should be taught and learned early on, but research will be presented in Chapter 2 that shows that many of the commonest words in English are actually merely tips of phraseological icebergs (Figure 1.1):

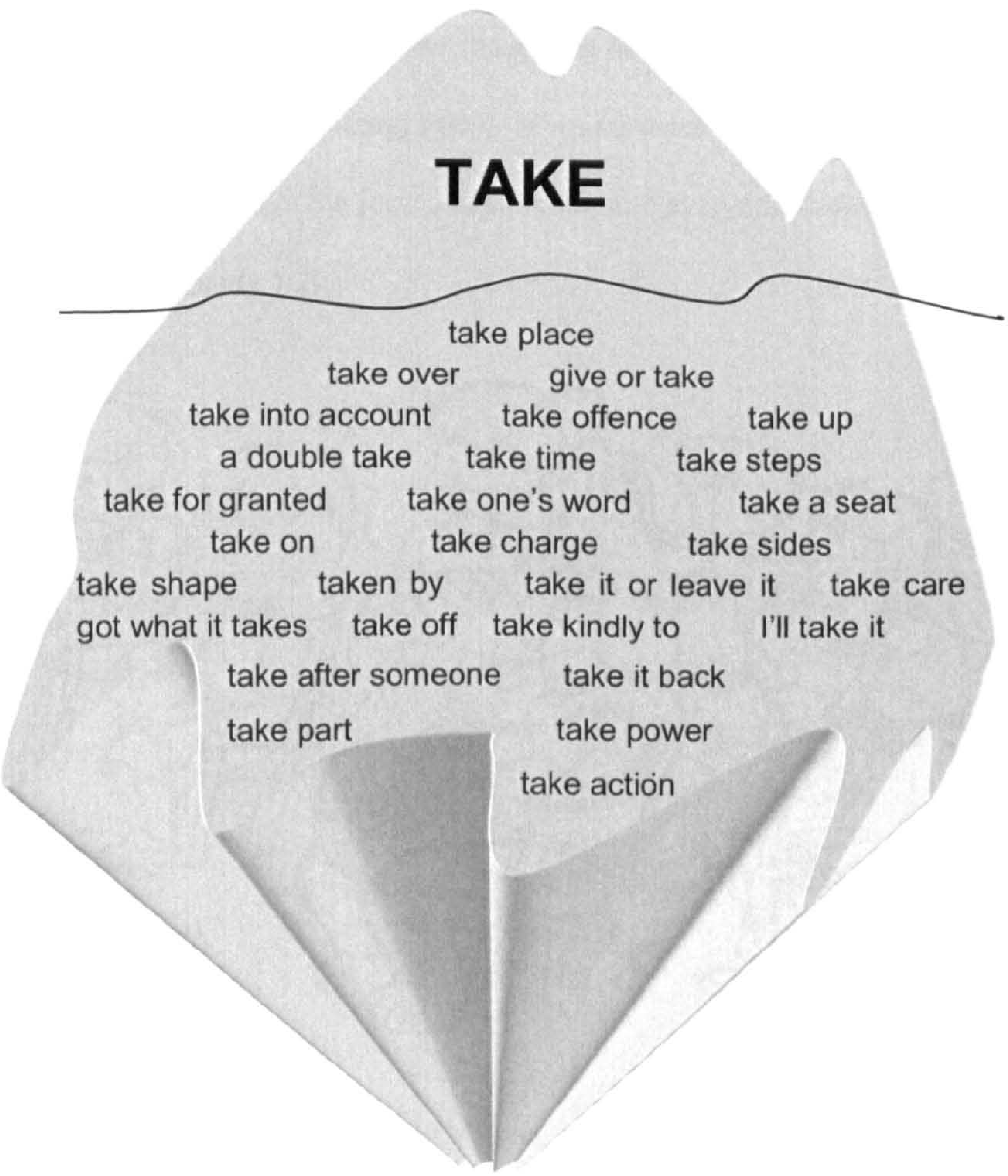


Figure 1.1. An underlying complexity of phraseology

In reality, what is presented conceptually in Figure 1.1 has been shown empirically by researchers working in corpus linguistics for some time (e.g. Sinclair, 1987), and has long been faced as a conundrum by lexicographers who are often unsure as to how such expressions should be listed in their dictionaries (Kilgarriff, 1994); inconsistencies regarding their classification abound (see also Chapter 5). There is even awareness of the importance of such multiword expressions, or what Wray (2002) has called 'formulaic sequences'[2], in language teaching circles (e.g. Lewis, 1997; O'Dell & McCarthy, 2008). However, the field of formulaic language is broad and not yet fully defined (Granger & Meunier, 2008; Wray, 2008), and teachers – and researchers – still often struggle to find ways to actually incorporate such language into their work (Boers & Lindstromberg, 2009). It would seem that formulaic language in pedagogy is still something of an elephant in the room – a phenomenon everyone knows is there, but is just unsure what to do with exactly (Figure 1.2).

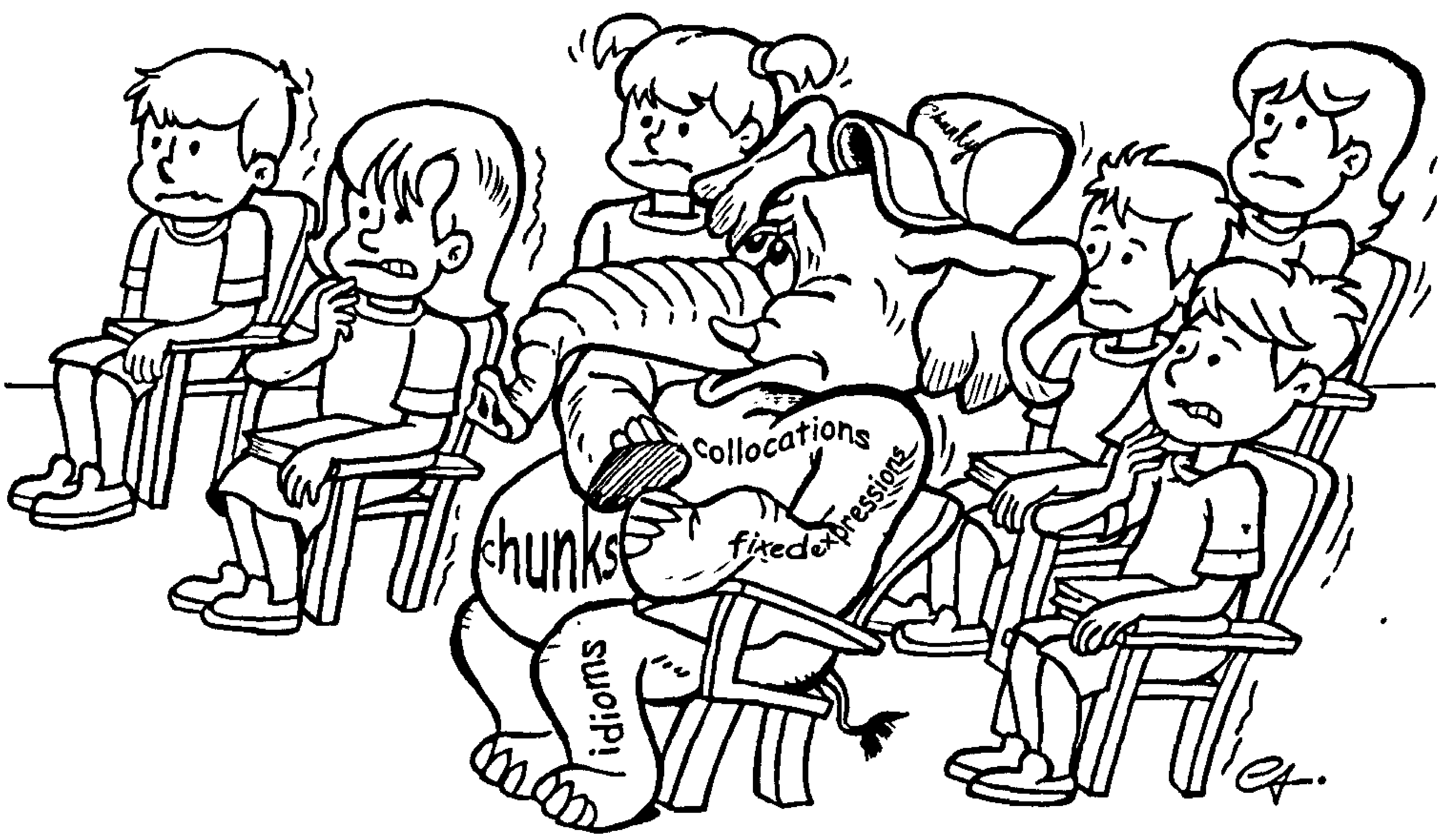


Figure 1.2. The phraseological 'elephant in the room' in pedagogy

[2] Following Schmitt (2010), the overarching term for the phenomenon of phraseology often will be referred to as 'formulaic language' throughout the present work, while 'formulaic sequences' employed as the term for the specific linguistic realizations of this phenomenon. The issue of terminology will be revisited in Chapters 2 and 3, but the author will also occasionally use the term 'multiword expression' to refer to the same phenomenon.

As will be illustrated in Chapter 2, the field is so broad that it is not really practical to set about simply 'learning' or 'teaching' formulaic language in general and at random. When focusing on specific pedagogical applications, just like with nearly all elements of language when organized into a course syllabus (e.g. grammar structures, pronunciation focus, thematically-related vocabulary), it seems more useful to narrow down the phraseological field and find a way to prioritize which items to learn or teach. The formulaic sequences depicted in the iceberg in Figure 1.1 can actually be considered to have different properties, and therefore may not be equally relevant to the language needs of L2 learners. For example, some are more frequent than others; some may sound more informal; still others may be more easily understood than others which may not be decodable by a literal understanding of each word (i.e. semantically less transparent). Moreover, all of the aforementioned features may have differential implications for language teaching, research and testing. Therefore, the main hypothesis at the outset of the present thesis is that it is possible to identify which formulaic sequences are more likely to be more useful in than others in specific applied linguistic contexts, and there is a way to prioritize these items. Therefore, Chapter 3 presents the primary and pivotal part of the present work, describing the rationale for and development of a list designed to facilitate that prioritization, entitled the PHRASE List.

Chapter 3 lays the necessary groundwork for two case studies that follow it which deal with concrete applications of the list. The first, in Chapter 4, provides a detailed account of a research validation exercise in which the PHRASE List played a central role. The research described in the chapter serves in part as an example of the type of the practical ways in which the list can be used, but also reciprocally acted as a validation exercise for the PHRASE List. The second, in Chapter 5, details the development and piloting of a new test designed and constructed using the PHRASE List, including the challenges that were faced in the development process and how they were addressed.

Chapter 6 concludes the thesis, and does so by summarizing the research while also speaking to current and envisaged future implications of the work herein described. There of course has been a great deal of thought and time that has been invested in the research outcomes that will be presented here, but there is little doubt that it represents only a very small part of the work that is hoped will still be carried out in applying it to actual research and pedagogy.

Chapter 2 – Vocabulary and formulaic language

The focus of the present research rests primarily on multiword expressions, but 'vocabulary' is very often conceived of as consisting mostly of individual words (Hill, 2000), so 'words' will be the first point of analytical departure in this chapter. Section 2.1 deals with some of the key concepts and prevailing assumptions concerning vocabulary; Section 2.2 revisits those concepts from a more critical perspective; Section 2.3 deals with the operationalizaton of formulaic sequences, and Section 2.4 concludes the chapter by presenting the central research questions.

2.1 Key concepts regarding vocabulary knowledge

What superficially would seem a fairly straightforward concept – *word* – is actually, as most lexicographers, language teachers and L2 learners are well aware, not very often straightforward at all (Miller, 1999). The very word *word*, for instance, can obviously be a noun, but it can also be a verb (*I don't know how to word this email*). Further, with affixation, the meaning can change (*I don't like the wording of this email*) and even change word class (*This email is too wordy*). Further still, when other words are added to collocate with *word*, new meanings often form, such as *word* meaning 'news' (*Have you received any word from the mechanic?*), 'promise' (*I give you my word*), and even 'authority' (*He's the last word on the subject*). Nation (2001) provides a guideline to what it means to 'know a word' (Table 2.1), categorized into 'form', 'meaning' and 'use', which in turn are subdivided into 'receptive' knowledge ('R') and 'productive' knowledge ('P').

Table 2.1 Elements involved in knowing a word (Nation, 2001: 27)

FORM	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express meaning?
MEANING	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this word make us think of?
		P	What other words could we use instead of this one?
USE	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency ...)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?

R = receptive, P= productive

Hence, if vocabulary itself is often full of complexity, it follows that vocabulary knowledge – as illustrated in Table 2.1 – must also be complex, and therefore the practical implications related to the learning, teaching and testing of vocabulary knowledge equally multifaceted. The current section provides an overview of four key concepts in vocabulary as it pertains to pedagogy: the idea of **a lexical space**, the construct of **word**, the role of word **frequency** measures, and the notion that vocabulary knowledge can be divided into **thresholds**.

2.1.1 *The 'lexical space': breadth, depth and fluency of vocabulary knowledge*

According to Zechmeister, Chromis, Cull, D'Anna and Healy (1995), the average educated adult native speaker of English possesses a receptive knowledge of around 20,000 word families[3]. To arrive at their estimate, the researchers tested students and older adults (n = 112) on 200 'functionally important' words (i.e. entries generally considered 'vocabulary words' (p. 202)) randomly selected from a dictionary. Participants were asked to rate each word from 1-5, with a value of 5 indicating that the participant knew the word well enough to give its definition (p. 204). To verify reported word knowledge, the researchers then administered a multiple-choice test, and found that around 80% of older adults could demonstrate some knowledge of over 20,000 word families.

It is clear, however, that participants merely showed, on a relatively superficial level, the extent of their word knowledge (i.e. how many words), otherwise known as 'breadth' (Figure 2.1).

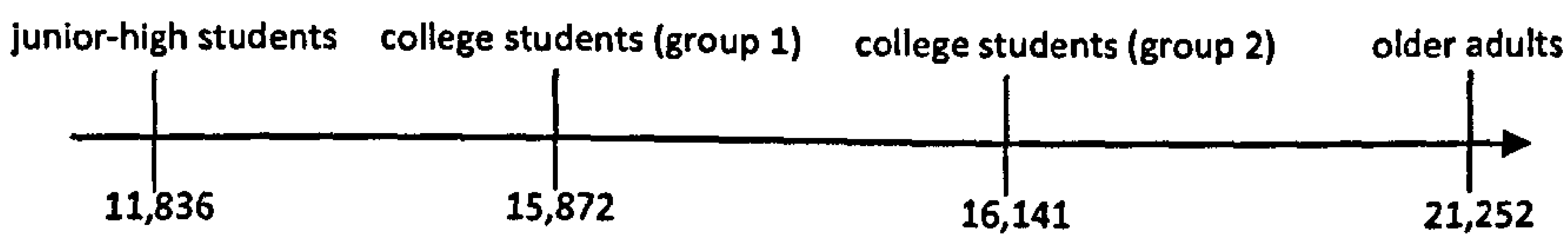


Figure 2.1. Breadth of word knowledge in the Zechmeister et al. (1995) study

Although participants in the Zechmeister et al. study did demonstrate they knew what the words meant that they had claimed to know, what they in fact demonstrated was knowledge of *a* meaning of each word on some receptive level[4]. For example, the word *chivalry* appeared with the following alternatives on the multiple-choice test (p. 205):

a. warfare b. herb c. bravery d. lewdness e. courtesy

---

[3] A word family consists of a headword, its inflections and other derived forms that still retain most of the original semantic meaning of the headword. (See also Section 2.1.2.)

[4] Rating knowledge of a word as '5' in the Zechmeister et al. (1995) meant the participant claimed to be able to actually write a definition for the word, which would indicate a level of productive knowledge. However, this assertion was never followed up with actual testing of that ability.

However, nothing in that test item assesses whether or not participants knew that it is usually men, who show chivalry towards women (by cultural convention), and that it can also denote the set of qualities 'expected of a knight'[5]. Nor does the above test item check if participants know the pronunciation of the word. We cannot know if the participant could use it appropriately in a sentence, or even spell it correctly. And we cannot extrapolate other important knowledge, such as if the candidate knows that *chivalry* is an uncountable noun (i.e. its grammar), or that it can become an adjective by dropping the *–y* ending and adding *–ous* (i.e. its morphology). In other words, the Zechmeister et al. (1995) research does tell us something about how many words the participants tested knew (breadth), but provides little insight into the completeness of that knowledge, or *depth* (Figure 2.2).

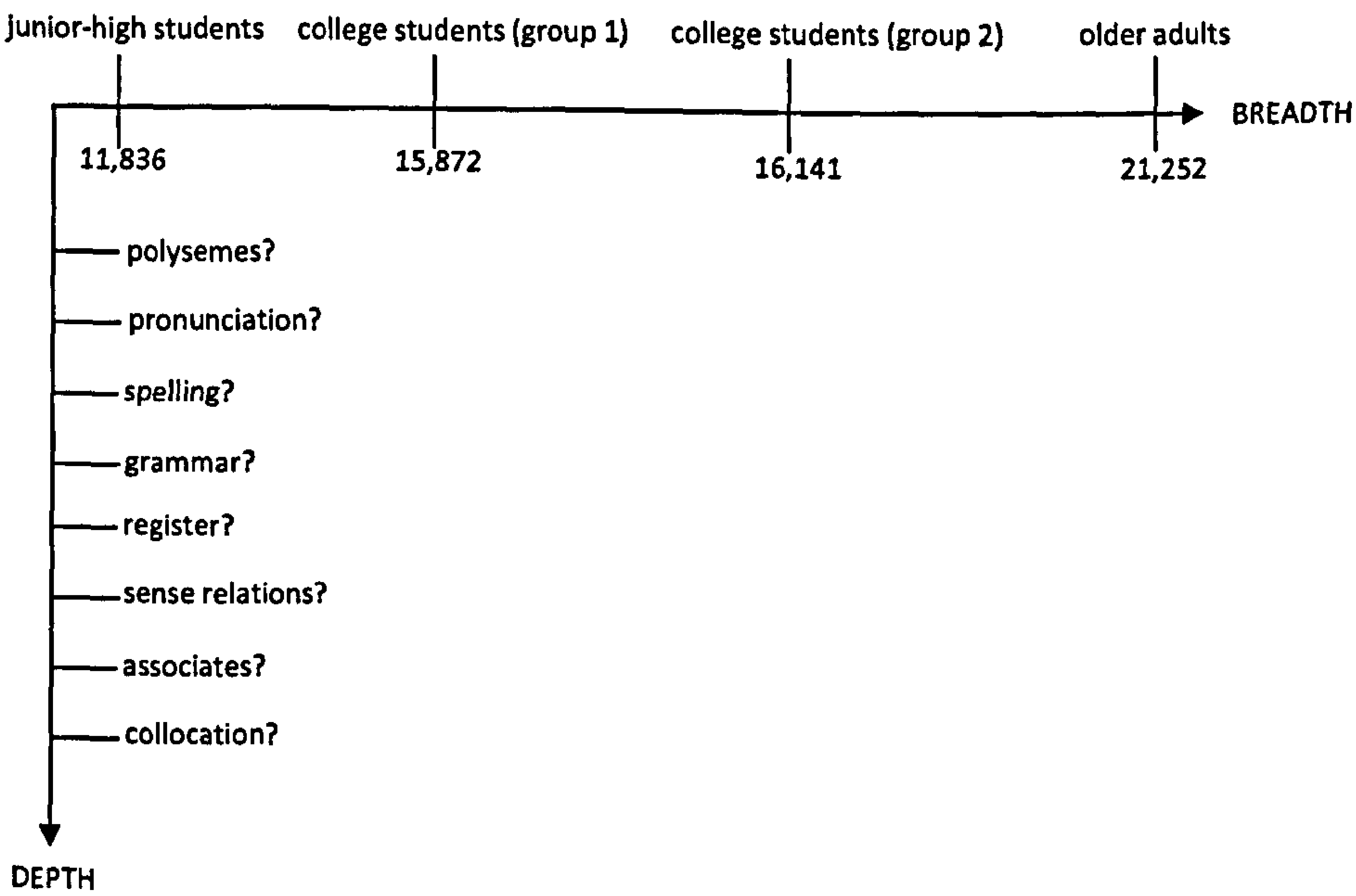


Figure 2.2. Breadth of vocabulary knowledge in Zechmeister et al. (1995) study and unknown knowledge depth

[5] As defined in the Macmillan English Dictionary (Rundell & Fox, 2007: 250).

Beyond the breadth of how many words one knows, and in addition to the depth of one's knowledge of those words, another important dimension of the mental lexicon is how readily one is able to recall and use a word, and the relative ease with which it is used. This notion, which typically involves the speed and accuracy with which a given word can be used by an individual, is known variously as 'automaticity' (e.g. Segalowitz, 2003) and 'fluency' (e.g. Meara, 2002). Altogether, *breadth, depth* and *fluency* form a kind of three-dimensional 'lexical space' (Daller, Milton and Treffers-Daller, 2007), as illustrated in Figure 2.3.

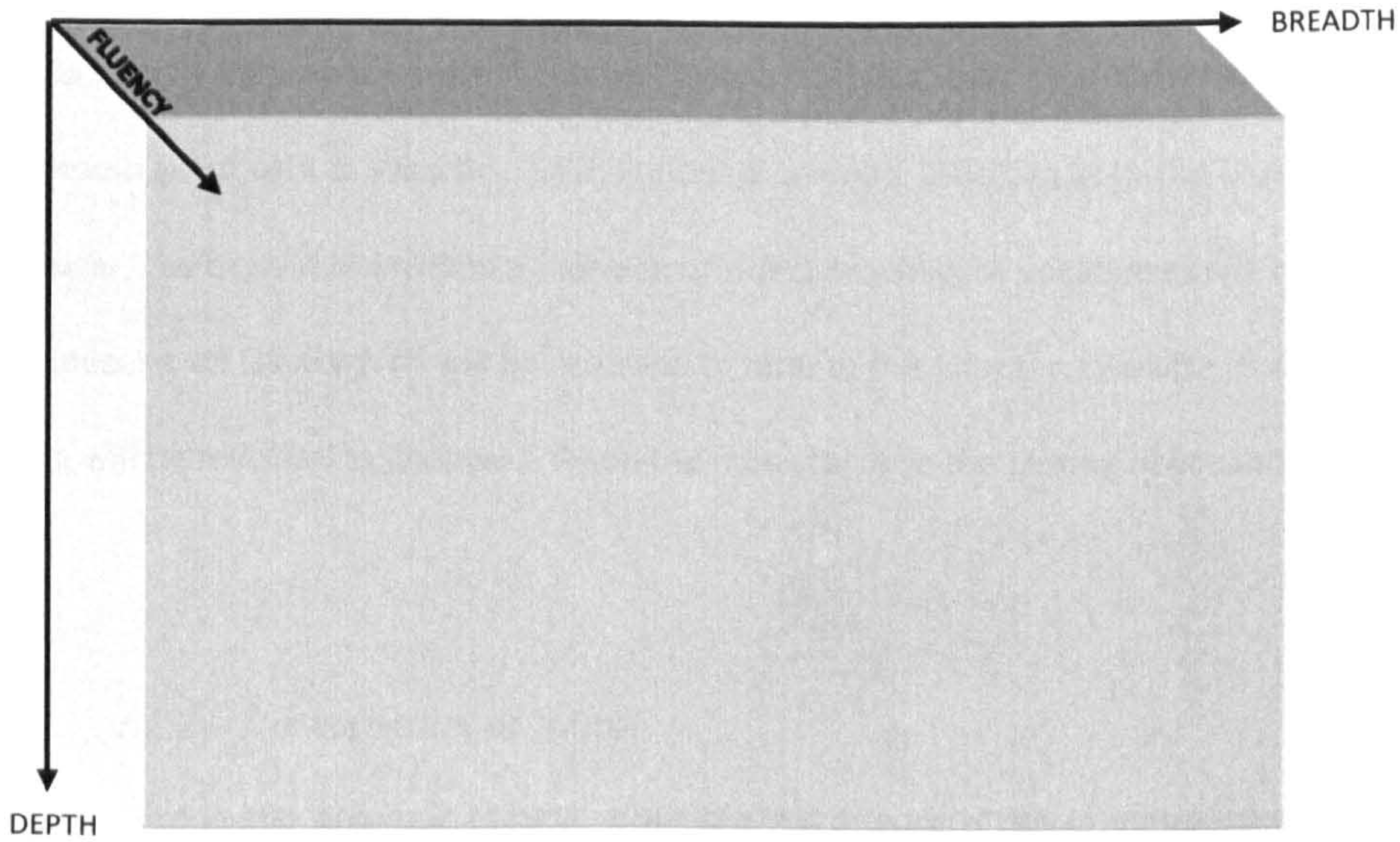


Figure 2.3. The lexical space: dimensions of word knowledge and ability (Daller et al., 2007: 8)

The concept of a lexical space is central to nearly any discussion of vocabulary and its instruction and assessment[6]. For example, students and teachers should be made aware of the dimensions in Figure 2.3 because, generally speaking, the ability to recognize a word and its meaning is an element of word

[6] It should be noted that although the notions of breadth and depth are convenient in terms of pedagogy, there is a growing body of research to suggest that the mental lexicon is not organized in this way but rather as a complex interconnected network (Aitchison, 1987). The breadth/depth model, however, is presented here because of its influence in vocabulary testing, for example, and also because the network that has been theorized has yet to be operationalized (Meara, 2009).

knowledge that can occur relatively quickly, as when children 'fast map' the meaning of new worlds they encounter (Heibeck and Markman, 1987), however that is not usually the case in the other dimensions of the lexical space (Nation, 2001). The acquisition of vocabulary depth and fluency tends to occur incrementally (Schmitt, 2010: 19), through, for example, multiple exposures to the same word in different contexts (Elley, 1989; Waring & Takaki, 2003; Webb, 2007). Although this incremental process seems to occur largely naturally and subconsciously in the L1 (Hoey, 2005; Nagy, Herman & Anderson, 1985), what research suggests is that the same holds for L2 vocabulary acquisition, but to a lesser degree (Laufer, 1998)[7]. What seems to be of greater benefit to L2 learners is some kind of form focused instruction to complement the incidental acquisition that may take place outside (or even inside) the classroom (Paribakht & Wesche, 1997; Sonbul & Schmitt, 2010). In both the L1 and L2 learning contexts, however, the benefit of incidental learning or direct teaching of vocabulary will be affected by frequency of exposure, an issue which will be returned to later in this chapter. Likewise, issues related to the lexical space will be revisited in Chapter 5 when the focus turns to the testing of vocabulary.

2.1.2 The construct of 'word'

As discussed in the previous section, even the word *word* is full of complexity, but not all words are created equal. Although most words (in the English language, at least) also contain varying degrees of semantic complexity, there is a class of words which is characterized by a lack of propositional content. Words like *word* are often called 'content' or 'lexical' words, whereas words like *the*, *of* and *who* are called 'function' or 'grammar' words (Lado, 1955; Segalowitz & Lane, 2000). Although function words are among the commonest in the English language (Leech, Rayson & Wilson, 2001), it is content words

[7] There can be many reasons for this difference (e.g. age, working memory limitations), but one that is widely accepted is the fact that one is exposed to the L1 usually on a daily basis when growing up, and therefore hears and reads many of the same words again and again. Since language is considered to be an interconnected network (Hudson, 2007; Meara, 2009), the connections are often made stronger and at a faster rate in the L1 because the activation of one word also strengthens the network to which it is connected in the mental lexicon.

which are typically the focus of vocabulary instruction and tests assessing knowledge of vocabulary (e.g. Flower & Martinez, 1995; Meara & Jones, 1990) – for reasons that seem obvious. (For example, how would one go about teaching/testing the word *of* alone?) Nevertheless, it is germane to the present discussion to highlight the fact that when estimates are made of 'how many words' one knows, the reported numbers do not generally include function words. As will be seen in Section 2.2, it is likely such words play a much more important part in vocabulary knowledge than is reflected in current vocabulary pedagogy.

Estimates of vocabulary size also do not generally count how many individual word forms one knows. Returning to the example of *word*, it is reasonable to assume that when an adult native speaker of English knows the word *word*, she or he also knows its plural – *words*. Words and their grammatical inflections (e.g. *do, does, did, doing, didn't*) are called lemmas (Nation, 2001: 7), and a lemma and all its inflections are generally considered to be stored together in the mental lexicon (Aitchison, 1987). Perhaps more controversial, in terms of mental representation, are morphological derivations from a lemma, such as *word* → *wordless*. Such shifts in form also generally change – to varying degrees – what the word means (Moon, 1987: 89).

Bauer and Nation (1993) claimed that some lemmas have forms created by derivational morphology (i.e. prefixes and suffixes) that should really be grouped together. For example, according to the authors, since words such as *wordy* and *wordless* derive from a known word, those three words – *word, wordy* and *wordless* – are actually part of 'word family':

> From the point of view of reading, a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately. [...] The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort. (Bauer & Nation, 1993:253)

On the basis of the above assertion, Bauer and Nation (ibid.) devised a 7-level model of affixation to systematize what words should be included in a word family, using the following criteria (Nation, 2001:267):

- frequency (how commonly used the affix is);
- regularity (how much the word changes as a result of the affixation);
- productivity (how usable the affix is on other words);
- predictability (how transparent/narrow in meaning the affix is).

In Bauer and Nation's model, the lower the level, the more reasonable it is to include a given word in a word family. Therefore, since *–y* and *–less* belong to easiest stage in the Bauer and Nation classification, it would be safe to assume that *wordy* and *wordless* should be considered part of a one word family, as their recognition requires, reiterating Bauer and Nation's claim, "little or no extra effort'" from a person already familiar with the lemma *word.*

Indeed, at least in the case of native speakers of English, the word family does appear to have some psycholinguistic validity (Zareva, 2007). (See, however, discussion of Schmitt and Meara (1997) in Section 2.2.1 of this chapter.) What is important to bear in mind for the present is that the validity of the word family only holds as long as the meanings of all the word forms contained in it are related, linked to the core meaning of the base word. Otherwise, as Sinclair (1991) advises, a sentence like "*give* occurs 50 times in this text" loses meaning (p. 173).

2.1.3 *Frequency*

The word 'frequency' as it pertains to vocabulary can actually mean different things, and each one of those elements of meaning are relevant to the whole of the current thesis. In fact, of all the points thus far made in the present chapter, the notion of 'frequency' probably has the strongest influence on the research that forms the bulk of the outcomes of the research that will be reported on starting in the next chapter.

'Frequency' from the perspective of the language learner, for example, can refer to frequency of exposure, i.e. how often that learner reads or hears a word (whether consciously or otherwise). It is a variable that is controlled for in vocabulary-related studies to try to better understand, for example, how many exposures a learner needs on average in order for a word to begin to enter the mental lexicon (e.g. Pellicer-Sanchez & Schmitt, 2010; Rott, 1999). As alluded to in the previous section, frequency of exposure has also been shown to have influence on depth of vocabulary knowledge, at least among native speakers (Lessard-Clouston, 2006).

On the other hand, a lexicographer compiling a learner dictionary might understand 'frequency' to refer to how common, relatively speaking, a given word is – usually (in recent years) according to corpus[8] data. Such information is seen as important in learner dictionaries, and indeed an indication of frequency is often attached to the most common words in those works of reference. Frequency is also one of the way some lists of the most common words are ranked in English, such as those in the *Word Frequencies in Written and Spoken English* (Leech at al., 2001) and the *Frequency Dictionary of Contemporary English* (Davies & Gardner, 2010), in which the latter authors state plainly that "corpus-derived frequencies are still the best current estimate of a word's importance that a learner will come across" (Davies & Gardner, 2010: v111).

[8] A corpus in the present thesis will be taken to mean a collection of texts that are analyzable by computer.

Even before computerized corpora were in existence, there was an awareness of this 'word frequency=importance' relationship (e.g. Thorndike, 1921). However, a problem that has also been recognized is that estimates of word frequency made by individuals are not very reliable ways of judging how relatively important a word is (Alderson, 2007), and this realization is in part what touched off what later came to be referred to as the 'Vocabulary Control Movement' (Gilner, 2011; Schmitt, 2000), a kind of imperative to systematize vocabulary learning in schools. In practical terms, what famously emerged from the movement were some of the earliest empirically-derived frequency lists, some of which are still in use today.

One of the earliest of such lists was the Thorndike and Lorge (1944) *The Teacher's Word Book of 30,000 Words,* based on a corpus of around 5 million words (West, 1954: 121). The alphabetical list was constructed following an exhaustive manual word tally of several different texts of various genres used in American elementary schools of the early 1920s (Thorndike, 1921). Interestingly, nearly 100 years have passed, but the same basic methods Thorndike and his colleagues used to extract frequency information prevail to this day: by counting how many times a given word occurs in a corpus. Indeed, the source corpus sizes of today are generally larger, and the computer does most of the work, but the method and end product are essentially the same (cf. Garside, Leech & McEnery, 1997).

In terms of the present discussion, what is particularly relevant is the natural tendency for shorter, common words to account for the bulk of most texts, with highest-frequency words (e.g. the top 2000) comprising nearly 80 percent (Nation, 2001: 11; O'Keeffe, McCarthy & Carter 2007: 32). This tendency for the most common words to also generally represent the most word coverage is also a consequence of Zipf's Law. Zipf's Law (Zipf, 1949) states that the frequency of any given word is usually inversely proportional to its rank in a corpus-derived frequency list (Matthews, 1997: 409). Therefore, the second-most-common word will usually be twice as common as the fourth ranked word, and that word twice as common as the word ranked eighth, which in turn will be twice as frequent as the sixteenth, and so on. What Zipf's Law reflects is the tendency towards a kind of lexical Pareto principle[9] in language, in other words, that a relatively small number of the most common words in a language are doing the bulk of the work. It is also worth pointing out that the most common words also tend to be shorter (Leech et al., 2001) – a fact that will reveal itself as important in the later chapters of the present thesis. As will be discussed further in Sections 2.2.2 and 2.2.3, the natural Zipfian distribution of the lexicon has led many researchers to assume that text is largely based on those most frequent words as individual words (e.g. Nation & Waring, 1997; Read, 2004), *ipso facto*. This notion and its possible negative consequences will be explored further later in this chapter.

2.1.4 *Vocabulary thresholds*

As mentioned in the previous section, Thorndike's lists (e.g. Thorndike, 1921; Thorndike, 1931) were aimed at helping educators prioritize vocabulary instruction. Thorndike believed, as is still often held today (Alderson, 2007; Schmitt & Dunham, 1999), that "even expert teachers have very inadequate and inaccurate notions of the relative frequency and importance of words" (Thorndike, 1921: 360). Therefore, teachers of English as a Second Language, for example, could use a frequency list to help an L2 learner of English start reading faster. Thorndike estimated that a mastery of the top 500 words in English represented a threshold level for that type of learner:

> By the elimination of certain specially childish or "literary" words from the first 500 of our list and the addition of certain words of special importance to the newcomer to America, such as *danger, poison, cent, dollar, entrance, exit,* we shall have a basic list of great value in teaching foreign adults to read English. (Thorndike, 1921: 364)

---

[9] That in many phenomena, 80% of the effects come from only 20% of the causes (Pareto, 1935).

Some years later, West (1953) revisited this idea of 'threshold' and focused on how to build a "minimum adequate vocabulary" (West, 1954: 121) to function in English. West studied Thorndike and Lorge's (1944) list of the most common 30,000 words and narrowed it down to just the top 2000, which he called 'A General Service List of Words' (West, ibid.)[10].

Research into West's cut-off point of 2000 words as a pedagogically meaningful threshold would not occur until many years later, in more recent research (e.g. Hirsh & Nation, 1992; Nation & Waring, 1997) endeavoring to answer the question of *How many words are really necessary in order to comprehend most texts?* After all, to be able to put a concrete number on the amount of words one needs to know to function in the target language is to be able to set teaching goals, divide proficiency levels and see a proverbial light at the end of the L2 learning tunnel (Schmitt, 2010).

However, the answer to that question has also proved somewhat complex, requiring indentifying not so much how many words one needs to know in absolute terms, but rather how many words a learner needs to know in order to understand a text in spite of unknown vocabulary.

One of the first empirical studies to examine this issue was Laufer (1989a). Laufer sought to explore how much vocabulary was needed in order to score at least 56 percent (the minimum required to pass) on a Haifa University reading test. To this end, Laufer asked learners to underline all unfamiliar words in the texts, counted those items, and then adjusted this number by comparing those words with a test that checked their L1 knowledge of the words in the texts. Using this methodology, Laufer found that knowing at least 95 percent of all running words was what was required in order to reach that 56

[10] It is interesting to note here that the General Service List was originally intended for language production (i.e. writing and speaking), and not the receptive skills (p. 126). West therefore chose not to base the list on frequency alone; rather, he decided to prioritize the words that were most common in English but that were also 'essential' to speech and writing (p. 122).

percent mark on the Haifa University test. Laufer then used earlier research conducted by Ostyn and Godin (1985) to calculate that around 5000 words would provide that 95 percent word coverage. However, since Ostyn and Godin's research was based on Dutch and not English, such an estimate may be misleading. Moreover, Ostyn and Godin do not specify if their methodology for counting words was based on word families or individual words (e.g. lemmas), and so Laufer's estimate of 5000 may be somewhat dubious (cf. Hirsh & Nation, 1992). Further, it should be noted that Laufer (1989a) does not justify the 56 percent figure as being necessarily 'good' comprehension, rather simply what is minimally required in the Haifa University system.

To further elaborate on her (1989a) research, Laufer (1991) tested 92 adult students of English as a Foreign Language for their vocabulary knowledge and divided them into bands of 2000, 3000 and 5000 word families and compared their scores on the Haifa University reading test. According to the author, the difference between the means was only significant in the transition between the 2000 and 3000 word-family level (p. 23), which, according to Laufer, suggested that the most significant threshold for L2 readers of English is 3,000 word families. Since 3,000 word families actually translates into about 4800 lexical items (p. 24), Laufer asserts that her earlier (1989a) estimate of 5000 words in order to score at least 56 percent on the Haifa University test was accurate. The author further claims that although reading comprehension does increase with greater lexical knowledge (i.e. above 3000 word families), it does not do so significantly, and therefore to teach 3000 word families is a defensible pedagogical aim. It should be reiterated, however, that Laufer's (1991) estimates are still merely based on achieving a passing mark in her university's system, and that 56 percent figure does not necessarily mean adequate or inadequate comprehension by other, non-institutionally established criteria.

Taking a slightly different approach from the Laufer (1989a, 1991) studies, later research considered what percentage of words a reader needs to know in order to read for more general purposes. Basing their assertions mostly on the assumption that 'pleasurable' reading occurs only when a reader knows almost all words in a text, Hirsh and Nation (1992) stipulated the ideal percentage of words known in an unsimplified text at around 98%. To further test this hypothesis, Hu and Nation (2000) gave sixty-six adult learners studying in New Zealand a relatively short (673 words) fiction text to calculate the relationship between the density of unknown words and reading comprehension. The researchers replaced varying amounts of low frequency words in the text with nonsense words in order to establish a minimum level of reading comprehension. For example, in the 95 percent coverage version, 5 percent of the running words were replaced with nonsense words (32 words), and in the 90 percent coverage version, 10 percent of the original words were replaced (about 67 words), and so on (p. 410). Following comprehension measures which included multiple choice tests and cued written recall, the authors concluded that when participants read the version in which 20 percent of the text consisted of nonsense words, no instance of 'adequate comprehension' occurred (p. 415). ('Adequate comprehension', according to Hu and Nation's criteria, meant accurately answering at least 12 out of 14 questions correctly on the multiple choice measure and 70 out of 124 on the written recall questions – arbitrary numbers by the authors' own admission.) When 5 percent of the running words were nonsense words, less than half of all participants achieved adequate comprehension. Since it was only at 100 percent text coverage (i.e. all words understood) that the vast majority of participants showed minimally acceptable reading comprehension, the authors postulate that 'around 98 percent coverage' (p. 419) is what is probably needed for learners to gain adequate comprehension.

Hirsh and Nation's (1992) study, at least according to Hu and Nation (2000), had accurately identified the ideal word coverage threshold at around 98%. In the same study, Hirsh and Nation (1992) also estimated that a learner would need to know at least 5000 word families in order to reach that threshold. However, Hirsh and Nation used novels written for teenagers and adolescents in their study to arrive at their estimates. To determine whether the same word-family figure would apply to authentic texts designed for general (i.e. adult native-speaker) consumption, Nation (2006) conducted a new analysis of fiction and non-fiction text (e.g. novels and newspapers). The trialing showed that if 98 percent coverage of a text is needed for unassisted comprehension – as determined in the Hu and Nation (2000) study – then an 8000 to 9000 word-family vocabulary is needed. Therefore, assuming Hu and Nation's (2000) assertions are mostly accurate (as most recently supported by Schmitt, Jiang and Grabe, 2011), a learner requires a knowledge of at least 8000 word families in order to adequately comprehend most unsimplified fiction and non-fiction text.

2.2 Key concepts: a more critical look

Thus far in the present chapter, the focus has been on presenting a core of the key concepts and assumptions which underlie vocabulary teaching and testing as it largely exists today. However, there are a few weaknesses in each of the concepts so far discussed, and those shortcomings in large part form the rationale for the research which will be presented in detail in the chapters which follow this one. These concepts will therefore now be revisited with a more critical eye.

2.2.1 A critical look at 'word family'

As mentioned in section 2.1, current research suggests that 8000-9000 words can provide around 98 percent coverage of most texts (Nation, 2006). However, Nation's recommendations are for a "8000-9000 word-family vocabulary" (p. 79), which does not necessarily mean knowing 8000 *words*. In the lists Nation and other researchers have used to calculate word knowledge, a word can include a base form and over 80 derivational affixes (Nation, 2006: 66), resulting in "some large word families, especially among the high-frequency words" (Nation, ibid.). A word like *nation*[11], for example, may include a word

[11] ***nation**: national, nationally, nationwide, nations, nationalism, nationalisms, internationalism, internationalisms, internationalization, nationalist, nationalistic, nationalistically, internationalist, internationalists, nationalize, nationalized, nationalizing, nationalization, nationalizations, nationalize, nationalized, nationalizing, nationhoods.* (Nation, 2006: 67)

family of over 20 separate words, which means that 8000 words in the lists Nation and others refer to may actually translate to well over 30,000 separate words (Schmitt, 2008: 332). Although some research suggests that native speakers of English often group morphologically related words together in the mental lexicon (Bertram, Baayen & Schreuder, 2000; Nagy, Anderson, Schommer, Scott & Stallman, 1989), the same may not apply to non-native learners.

Schmitt and Meara (1997) have demonstrated that students of English do not often readily know the derived forms of words or their associates in the same way an L1 user typically does. The researchers gave a word association and affixation test to three groups of Japanese learners of English, and received some surprising results. Although the learners performed adequately on the inflection suffixes (59 percent correct), they only managed to get 15 percent of derivation suffixes right (p. 26). In a separate study, Hay (2001) has shown that when a derived form is more frequent than its base form (e.g. *difference = differ +ence*), the affixation becomes less transparent, and the semantic meaning more distant. In other words, it is possible that many of the derived words currently included in word families should really be listed as separate words for pedagogical purposes.

Consider, for example, the semantic distance between the following pairs of words: *name → namely; price → priceless, fish → fishy; puzzle → puzzling*. Each of the preceding pairs would be grouped into the same respective word family, but it is unlikely that a learner of English would require 'little or no extra effort' to derive the meaning of a word like *fishy* from *fish*[12]. It is therefore conceivable that a large number of those 8000-9000 word families do not have the psycholinguistic validity their grouping would seem to imply, and a considerable portion of the 30,000 (or so) separate word forms subsumed in those families may in fact need to be learned as separate items.

[12] According to the *Cambridge Advanced Learner's Dictionary* (2008), which is informed by the 1-billion-word Cambridge International Corpus, the first sense of *fishy* is 'dishonest or false' (p. 537), and not 'smelling of fish'.

2.2.2 A critical look at 'word'

Current word lists have another potential limitation beyond what constitutes a word family: word frequency lists only list individual words. Similar to the semantic distance between *fish* and *fishy*, there is often an equal or greater disparity of meaning when a word is juxtaposed with another or more words and a new expression forms (Moon, 1997; Wray, 2002). For example, the words *fine, good* and *perfect* each have meaning, however those meanings do not remain in the expressions *finely tuned, gone for good* and *perfect stranger*. Sinclair (1991) posited the notion that most texts are not, in fact, composed of entirely of individual words, but also of formulaic sequences that should not be broken down into separate parts. He called this theory the 'idiom principle':

> The principle of idiom is 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. (p. 110)

However, such "semi-preconstructed phrases" have generally not been included in wordlists or the research into vocabulary thresholds. On the basis of the idiom principle, uncertainty can even be cast on what a 'word' is exactly. To some extent, perhaps the division between word and multiword expression is not as psycholinguistically cut-and-dried as conventionally believed. Consider, for example, the following sets of lexical items:

1. *underway, awhile, insomuch as, notwithstanding, straightaway, alright, onto*
2. *under way , a while, in so much as, not withstanding, straight away, all right, on to*

The lexical items in lines 1 and 2 above share the same meaning, but vary in form only by spaces between the words. There is no question that the 'words' in line 2 are written separately, but there is a question as to the extent to which that orthography actually represents the way those items are stored in the mental lexicon. In other words, is *under way* a formulaic sequence, but *underway* a word? There is evidence to suggest that the mind does not always 'separate' or 'break down' formulaic sequences into words, and this evidence will be discussed in the following section. For now, perhaps the following questions merit reflection[13]:

- Why is it that even highly educated native speakers still do not agree on the orthography of hundreds of words like 'on line', 'on-line' and 'online', for example?
- Why are spelling mistakes like 'alot', 'eachother' and 'nextdoor' so common?
- Why are so many native speakers unsure of the grammar rules regarding forms like 'altogether/all together', 'everyday/every day' and 'anymore/any more'?

Wray (2008) has suggested that "[s]eeing the single word as a consistent unit of language, and thus viewing a two-word string as two units joined by a grammatical rule, gives a special status to what, in some cases, is a rather arbitrary result of historical practices" (p. 10). Similarly, Himmelmann (2006: 255) notes, regarding individual words, that "there is a fuzzy middle-ground, particularly relating to compounds, particle constructions, and formulaic sequences."

These observations of 'arbitrariness' and 'fuzziness' when it comes to words and formulaic language deserve closer attention, and leads to the question of why formulaic language is not included in wordlists if the meaning of some multiword expressions differs from the meaning of the words of which it is composed, and if, in any case, there may be no meaningful difference in terms of the way such lexical items are stored and/or processed (Conklin & Schmitt, 2008; Underwood, Schmitt & Galpin,

[13] The data that informed these questions can be found in the Wikipedia webpage on 'Lists of common misspellings' (http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings)

2004). Interestingly, Nation (2006) recognizes this limitation of current wordlists, however does not consider it a problem. Nation bases this assertion on the assumption that most learners will be able to guess the meaning of multiword expressions that have some element of transparency, and since the number of 'truly opaque' phrases in English is relatively small, for the purposes of reading they are 'not a major issue' (p. 66). However, it is debatable just how 'small' in number those opaque expressions are and, much like the previously discussed derived word-forms that are actually semantically dissimilar, just how easy it is for a learner of English to accurately guess the meaning of more 'transparent' expressions. In support of his claims, Nation (2006) cites Grant and Nation's (2006) estimates of how many truly opaque phrases exist in English, which they term 'core idioms'. For example, the phrase *by and large* is considered a core idiom by the authors because the meaning of the whole expression cannot be related to the component words (i.e. it is non-compositional), and one cannot use the words to arrive at some kind of figurative meaning either (p. 2). Therefore, according to the authors, a core idiom is both non-compositional and non-figurative. Following an exhaustive search using 10 dictionaries and 5 EFL textbooks, Grant and Nation came up with a list of only 103 core idioms in English.

There are some potential issues with Grant and Nation's methodology, however. It should be noted that the primary source of data for the research was specialized idiom dictionaries, not a corpus of language or even dictionaries of general English. The authors were therefore relying on the selection criteria of the editors of those published works, and what those editors considered an idiom – which research has shown to be often unreliable (Liu, 2003).

Another methodological concern regarding the Grant and Nation study is the exclusion – for no explained reason – of phrasal verbs (e.g. *get away with*, *take after*, *put up with*, etc.) from their data. Phrasal verbs (verb + adverbial particle) are extremely common in English and many would also meet the non-compositional/non-figurative 'core idiom' criteria (Biber et al., 1999; Cowie, 1998; Gardner & Davies, 2007).

Gardner and Davies (2007) used the BNC to explore just how common phrasal verbs are in English. The authors found 518,923 total phrasal verb occurrences in the corpus, and then focused only on the 100 most frequent of those. Of particular interest is the fact that all of the 100 most common phrasal verbs would be frequent enough to be included in the list of the 8000 most common words in English.

Returning to the concept of 'lexical space' (Section 2.1), therefore, there seems to be a prevalent assumption in the literature that phraseological elements in the lexicon pertain to the dimension of vocabulary knowledge depth, which would support Nation's assertion that the sole representation of single orthographic words in wordlists is probably of no real negative consequence. Certainly, this is likely the case for more 'transparent' formulaic sequences, such as 'freer' or less 'restricted' (Cowie, 1994; Howarth, 1998a) two-word collocations (e.g. *highly unlikely, torrential rain*). However, as mentioned in Chapter 1, the field of phraseology is very broad, and as shown in the iceberg in that chapter (Figure 1.1), there may be formulaic sequences whose co-occurrence is not a function of one word tending to prefer another, but actually, as in the examples of *straight away/straightaway,* are really more like 'words' –and it would be dubious to suggest that through many exposures *straight* that one would eventually learn that it collocates with *away.* In other words, there may be a sub-set of formulaic sequences that really can be included in any discussion of vocabulary breadth, as for all intents and purposes, they constitute single lexical items, constituting, as Sinclair put it, 'single choices'. This idea will be explored further in the following sections.

2.2.3 *A critical look at 'frequency'*

There is a fairly pervasive claim in the existing literature on vocabulary learning that it is pedagogically worthwhile to focus on the 2000 to 3000 most common words in English (e.g. Coady, 1997; Laufer, 1991; Nation, 2004; Nation & Waring, 1997; Staehr, 2008), the implication generally being that encouraging students to learn these words will somehow facilitate accessing meaning in texts. Although knowledge of 2000 words will only provide around 80 percent text coverage (Nation, 2001: 11; O'Keeffe et al., 2007: 32), the purported advantage is that there is a greater pedagogical 'payoff' in learning those very frequent words:

> There is an obvious payoff for learners of English in concentrating initially on the 2000 most frequent words, since they have been repeatedly shown to account for at least 80% of the running words in any written or spoken text. (Read, 2004: 148)

Indeed, Read is correct: that figure does generally reflect the research. However, corpus-based research also tells us that those most common words that "do most of the work" (O'Keeffe et al., 2007: 32) also tend to have "less of a clear and independent meaning" (Sinclair, 1991: 114). Hence, the suggestion that there is a pedagogical benefit in focusing on the commonest words in English in order to allow for maximal comprehension of a text, for example, may in fact be somewhat misleading. Those frequency figures, and the lexical complexity that may underlie them, should be given more careful consideration. This is also the position of Gardner (2007):

> I would argue that suggested applications of corpus research based on frequency of word forms, without considerations of word meanings, will invariably suffer from one of three problems – or combinations of the three: (a) they will overestimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate the word forms; and/or (c) they will underestimate the actual number of meanings inherent in the word forms. (p. 253)

However, the limited validity of using frequency data alone to determine which words are most useful may be even more complex than what Gardner points out. According to Sinclair, not only are the most frequent words more prone to polysemy (i.e. multiple meanings for the same word form), as seems to be Gardner's principal caution regarding frequency counts, but they also are prone to lose even their 'basic meaning' because they tend to feature in formulaic sequences:

> For example, we think of verbs like *see, give, keep,* as having each a basic meaning: we would probably expect those meanings to be commonest. However, the database tells us that *see* is commonest in uses like *I see, you see, give* in uses like *give a talk* and *keep* in uses like *keep warm*. (Sinclair, 1987: vii)

A good example of very frequent individual words combining to form independent, complex meanings is the phrasal verb in English. Consider the verb *pick,* which occurs 14,274 times in the BNC, but 9,997 of those occurrences are as a phrasal verb, or precisely 70 percent of all instances (Gardner & Davies, 2007: 348). Add to that number a whole host of other idiomatic expressions in which *pick* occurs, such as *take your pick, pick a lock* and *pick and choose*, and it is likely that *pick* recombines with other very common words to form separate multiword units of meaning over 80 percent of the time. The verb *pick* is clearly the tip of its own phraseological iceberg, but it is obviously not alone in exhibiting this behavior, and is improbable that such a phenomenon does not somehow affect reading comprehension.

2.2.4 A critical look at 'thresholds'

As will be discussed in greater depth in Chapter 5, there are a number of vocabulary tests that have been designed to provide a general estimate of the test-taker's general vocabulary profile, operationalized by demonstration of knowledge of words sampled from wordlists at decreasing levels of frequency. In theory, the data that result from the measurement of these various vocabulary thresholds should be meaningful, telling the student, teacher or researcher something about how much vocabulary the candidate knows, and even what proportion of the running words in a text the student might be expected to comprehend. However, as has been argued over the past two sections, any data that are based on frequency counts of single orthographic words without attention to meaning and, in particular, phraseology, will only be painting a picture which is only partially complete. Indeed, if Sinclair's idiom principle is correct, then many of the words in texts should not be decoded individually. In the case of native speakers who seem to regularly be able to recognize and process formulaic sequences in discourse as they read (Underwood, Schmitt & Galpin, 2004), the idiom principle should pose no major barrier to comprehension. However, there is evidence that the same cannot be said of individuals reading in the L2.

Martinez and Murphy (2011) developed a two-part test of reading comprehension, with the texts in each part having perfect lexical symmetry (i.e. constructed using the exact same words), sampled from the 2000 most frequent words in the BNC – the same word threshold mentioned in the previous section as widely perceived as useful for language learners to focus on. The main difference between the two parts of the test was that, although both parts were written using the exact same high-frequency words, the arrangement of those words in one part was such that they often formed formulaic expressions of varying degrees of compositionality. The learners were required to answer true-false questions about each text, as well as assess how well they believed they had understood the text on a self-reported rating scale. The results show that the participants not only vastly underperformed on the measure of reading comprehension when the text contained formulaic sequences, they also tended to significantly overestimate how well they comprehended the reading passage as a function of the very common words contained in the multiword expressions which they often did not recognize. Instead, they tended to focus on the individual words – words they thought they should know because of how 'common' they are.

Formulaic language in text, irrespective of compositionality, might most usefully be classified as what Laufer (1989b) has callled 'deceptively transparent' lexis (p. 11). Laufer studied high school graduates learning English in Israel and discovered that many of them mis-analyze words like *infallible* as *in+fall+ible* (i.e. 'cannot fall') and *nevertheless* as never+less (i.e. 'always more') (p. 12). Likewise – although they were not part of her study – she found that formulaic sequences like *hit and miss* were being read and interpreted word-for-word. These lexical items that 'learners think they know but they do not' (Laufer, 1989b: 11) can impede reading comprehension in ways not accounted for in lists of common word families and research on 'thresholds' of comprehension. Nation (2006), in his dismissal of the limitations of the wordlists that informed his research, seems to assume that multiword expressions that have some element of transparency, however small, will be interpretable through guessing, and indeed there is research that supports that assumption (e.g. Cooper, 1999; Liontas 2002). However, a key caveat with respect to that research is that the participants being studied were aware that the focus of the exercises was formulaic language (and presented in isolation) – which does not mean that the same participants would have been as successful when encountering formulaic sequences *in natura*, which is what the Marinez and Murphy (2011) findings show, which in turn are consistent with the findings in Laufer 1989b):

> But an attempt to guess (regardless of whether it is successful or not) presupposes awareness, on the part of the learner, that he is facing an unknown word. If such an awareness is not there, no attempt is made to infer the missing meaning. This is precisely the case with deceptively transparent words. The learner thinks he knows and then assigns the wrong meaning to them [...]. (p. 16)

Substitute 'formulaic sequence' for 'word' above, which as has been discussed is not an unreasonable conceptual stretch, and it becomes clear that multiword expressions just may present a larger problem for reading comprehension than accounted for in the current literature. In fact, such 'deception' is even more likely to occur with multiword expressions, since such a large number of them are composed of very common words a learner would assume he or she knows.

In summary, the notions of 'frequency', 'threshold' and even 'word' have been shown in this section as potentially posing a threat to the validity of, for example, tests of vocabulary knowledge that are based on frequency counts of single words, and research on reading comprehension that does not take formulaic sequences into account.

In the section that follows, the field of formulaic language will be explored in greater depth.

2.3 Operationalizing formulaicity

The previous section in this chapter reviewed some of the key concepts in vocabulary in second language teaching, acquisition, testing and research, and how a lack of attention to formulaic sequences in those areas can pose a threat to validity and undermine results. However, as Read and Nation (2004) caution, formulaic language itself can have validity issues in its own right:

> Validity issues are particularly problematic with formulaic strings, as the essential criterion – storage as a whole unit – is a difficult one to operationalize. For internal validity, there is a need for a clear definition of what a formulaic string is, both at the conceptual level and in operational terms. (p. 35)

While the problems associated with a focus on single orthographic words to the exclusion of formulaic sequences are fairly clear, few have successfully been able to address that issue, in large part for the reason referred to by Read and Nation: pinning down the construct. That much of spoken and written communication consists of recurring sequences of two or more words is now very well established in the literature (e.g. Ellis, Simpson-Vlach, & Maynard, 2008; Howarth, 1998b; Meunier & Granger, 2008; O'Keefe, McCarthy & Carter, 2007; Pawley & Syder, 1983; Schmitt, 2004). What has eluded researchers since the beginning of formal exploration of phraseology in English is a uniform operationalization of formulaicity (e.g. an accounting of its properties, measurement of how it is represented psycholinguistically), and even a consistent nomenclature (cf. Alexander, 1987; Cowie, Mackin & McCaig, 1983; Moon, 1997; Nattinger & Decarrico, 1992). Nonetheless, as Wray (2008: 93) incisively observes, "[y]ou cannot reliably identify something unless you can define it, and the relationship between definition and identification is almost circular[.]"

After all, the pool of potential candidates for 'multiword expression' can be vast, depending on one's definition of what constitutes a formulaic sequence (cf. Erman & Warren, 2000). If strict *a priori* criteria for the identification of formulaic sequences can be arrived at, this in theory should be conducive to reducing that candidate pool. Leech et al. (2001), for example, observed very strict criteria for the inclusion multiword lexical items in their lists of the most frequent words in English based on the BNC, which they called 'multiword units':

> Multiword units are items which are treated as a single word token, even though they are spelt as a sequence of orthographic words. Because they function grammatically as single words (e.g. the conjunction *so that*, the preposition *in spite of*, *at least* as an adverb), they are treated as entries in their own right. (p. 8)

Hence, Leech et al.'s conceptualization of formulaicity is mainly based on fixedness and stable grammatical formal properties. The centrality of such criteria in their definition is actually largely the result of the *Constituent-Likelihood Automatic Word-Tagging System* (CLAWS) (Garside, 1987), which was used to analyze the corpus. CLAWS was devised to automatically identify the word class of orthographic words (i.e. noun, adjective, adverb, and so on) in a corpus, but it was found that certain words, such as those in the expression *to and fro*, deviated from their normal tags elsewhere in the corpus (Blackwell, 1987) when the tagging system was still in its development and training phase. Therefore, a mechanism called the 'ditto-tag'[14] (Blackwell, 1987: 111) was created, allowing CLAWS to search a corpus for "specific sequences of words ... whose syntactic role in combination differs from the syntactic role played by the same words in other contexts" (Blackwell, ibid.). To some extent, this "gestalt" (Blackwell, ibid.) analysis can be seen as an important objective and data-driven recognition that lexemes such as *to and fro* should not be analyzed word-for-word, but as one morpheme – in this case, 'adverb' – the criterion of functioning "grammatically as single words" also actually excludes a great number of potential multiword candidates. For example, since all phrasal verbs can be inflected and (some) separated (e.g. *sum up → sums it up*), Leech et al. do not consider them multiword units for the purposes of their research. (Or, rather, the tagger did not.) Moreover, the limited criteria also apparently caused the research to exclude forms which contain any verbs at all and could therefore theoretically undergo grammatical transformation, but in reality do not because of the expression's entrenched fixedness and/or 'grammaticization' (e.g. *followed by, going to, speaking of, having said that, mind you,* etc.) – an aspect of formulaicity which will be looked at again further on in this chapter.

By contrast, Wray proposes a more flexible definition:

> (A formulaic sequence is) a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored, retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray, 2002: 9)

[14] So called because after CLAWS tagged each word in the corpus, a specially-designed complement to CLAWS (IDIOMTAG) would identify a sequence of two or more words that are serving the function of one grammatical word, and then re-tag just the first word of that sequence and simply add the same tag with 'dittos' (") to the subsequent constituent words (Blackwell, 1987).

Wray's (2002) definition allows for far more inclusiveness with respect to multiword expressions. As opposed to the Leech et al. restriction of complete fixity of form, Wray allows for non-contiguous sequences, so that items such as *sum up* and *not only.../but...* can be also qualify as formulaic despite not having frozen forms. However, the subjectivity of some of the concepts inserted in that definition also makes it less principled. Consider, for example, the following excerpts:

"...which is, or appears to be, prefabricated..."	'Appears' to whom? Pawley and Syder (1983), for example, famously cite *'I want to marry you'* (p. 196) as being formulaic. However, this author has never heard that expression before in any other context (that he can recall). In fact, it does not seem, again, to this author at least, a stretch to say that the utterance as formulated sounds almost desperate. The more formulaic sequence, as recognized by the author, is *Will you marry me?* to express the same basic function. However, that is what 'appears to be' more formulaic to one author's idiolect.
"...that is, stored, retrieved whole from memory..."	There is currently no research available that can definitively prove that an item <u>is</u> prefabricated. There is plenty of evidence which suggests at least some degree of holistic storage and retrieval of certain formulaic sequences, using, for example, evidence from pause occurrences (Dahlmann & Adolphs, 2007; Erman, 2007), response times (Conklin & Schmitt, 2008; Sosa & MacFarlane, 2002), and even neurological imaging (Fogliata, Rizzo, Reati, Miniussi, Oliveri & Papagno, 2007; Tremblay, Baayen, Derwing & Libben, 2008) – but it is still only evidence. Little is known, for example, about whether individual differences in

	working memory or even literacy have any effect on the holistic storage or retrieval of multiword units, and it seems plausible that the holistic processing and production of an item is more likely to vary from individual to individual (and even utterance to utterance) than be an objective property of it. (This issue is revisited later in this section.)
"...rather than being subject to generation or analysis by the language grammar."	The jury also seems to be out with respect to the (lack of) analysis that occurs when encoding and decoding formulaic language. Giora (1997, 1999), for example, proposes the 'graded salience hypothesis' which posits that the most salient meaning of a word in an idiomatic expression is always active, and accessed first. We already know this to most likely be the case with L2 learners of a language, who tend to focus on 'the strongest' lexical word in an expression to unpack its meaning (Spöttl & McCarthy, 2003). Even in native speakers, a few studies have shown that the parts of the brain which activate when generating or processing language analytically (i.e. accessing grammar) are also active when decoding idiomatic expressions. Zempleni, Haverkort, Renken & Stowe (2007), for example, used fMRI to show that the bilateral inferior frontal gyri and left middle temporal gyrus are always involved in the processing of decontextualized ambiguous idioms (e.g. *down to earth* – with literal and figurative meanings both plausible). More research is needed, however, to investigate whether the same would be true for the types of multiword expressions Leech et al. included in their study (e.g. *in spite of, rather than*).

As seen in both the Leech et al. and the Wray definitions, no matter the definition, there will be potential problems associated with it, either because it excludes certain formulaic sequences that some might consider important, or in its attempt to allow for a wide variety of them, runs into the issue of including items that some do not even recognize as formulaic sequences. This is the risk with a definition such as Wray's, and in the case of so-called 'lexical bundles', described below.

There are those who prefer to allow the data to drive the selection, focusing solely on the token frequency of a string of words to help determine whether it is formulaic or not, or what has been termed the 'frequency-based' tradition (Barfield & Gyllstad, 2009; Nesselhauf, 2004). Biber, Johansson, Leech, Conrad & Finegan (1999), for instance, use the term 'lexical bundle' to describe "sequences of word forms that commonly go together" to form formulaic lexemes, with the only requirement being that the items 'recur' a certain number of times in a corpus (p. 990). An example of a lexical bundle is *I don't know*, which the corpus Biber et al. used also shows can be expanded to *I don't know if*, *I don't know if I*, and so on. The multiword items, therefore, are required to have a degree of fixedness as judged by the computer software (hence making the criteria for their selection more objective than the Wray (2002) definition), but they also can be 'lemmatized' in the same fashion Bauer and Nation (1993) established for individual orthographic words.

Wray (2002) argues, however, that just because a string is recurrent does not mean that it is necessarily formulaic, showing her preferences towards what has been called the 'phraseological' tradition of defining formulaicity (Erman, 2009), which weighs semantics and functional properties of sequences. Indeed, to the average native speaker a string like *I don't know if I* might hardly seem like a coherent unit stored anywhere in the mental lexicon. To this Biber and Barbieri (2007) respond that no one has yet defined what formulaic language is – further exemplifying the circularity that often characterizes the argumentation in the operationalization of multiword units.

An attempt to operationalized formulaicity may also involve identifying different types of formulaic sequences and describing their individual properties, such as their pragmatic or sociolinguistic function. Tannen and Öztek, for instance, define formulaic expressions as set formulas that "afford (speakers) the tranquility of knowing that what they say will be interpreted by the addressee as in the same way that it is intended" (Tannen & Öztek, 1981: 46). The type of formulaic utterance Tannen and Öztek describe has elsewhere been variously called "conversational routines" (Aijmer, 1996; Coulmas, 1991), which in other studies have even been further broken down into categories such as discourse markers (*I see, I mean*), hedges (*kind of, you know*) and vague language (*or whatever, something like that*), to cite a few examples (Channel, 1994; McCarthy, 1998; McCarthy and Carter 1997). Such pragmatic approaches have proven to be useful in the area of English as a Second/Foreign Language, where students are taught to use formulaic expressions to 'do' certain things in conversational English, such as manage discourse (*by the way, speaking of*) and sound polite (*Would you like to, Do you mind*), and so on (cf. Wilkins, 1976).

There are some problems that can surface when one attempts to provide a comprehensive functional classification of multiword expressions in English, however. Nattinger and DeCarrico (1992), for example, identified a type of multiword expression – the 'polyword', which is similar to the class described earlier by Leech et al. (2001): "Polywords are short phrases which function very much like individual lexical items", "allow no variability" and "are continuous" (Nattinger & DeCarrico. 1992: 38). However, instead of attempting to assign a grammatical classification to each polyword, the authors venture a functional one:

Form	Function
for the most part	(qualifier)
in a nutshell	(summarizer)
by the way	(topic shifter)
I'll say	(agreement marker)
hold your horses	(disagreement marker)
at any rate	(fluency device)
what on earth?	(marker of surprise)
so long	(parting)
for that matter	(relator)
so to speak	(fluency device)

The authors continue to name other functions, including 'evaluator', 'clarifier', 'exemplifier', 'approval marker', and many others. What the Nattinger and DeCarrico list perhaps illustrates is the fact that while it is certainly possible to assign a function to just about any formulaic sequence, it is questionable how useful and helpful it is do so. Attempts such as the more recent Granger and Paquot (2008) functional classification (Table 2.2) simultaneously illustrate the sheer breadth of the typology of phraseology that exists while providing insight into its inherent complexity; unfortunately, each new attempt at a functional description also usually entails the addition of one or more new names for different formulaic sequences, thereby potentially contributing to the existing and ever-growing terminological disarray.

Table 2.2 A functional categorization of multiword lexical items (Granger & Paquot, 2008)

Referential	Textual	Communicative
(lexical) collocations	complex prepositions	speech act formulae
idioms	complex conjunctions	attitudinal formulae
binomials and trinomials	linking adverbials	proverbs
similes	textual sentence stems	commonplaces
compounds		slogans
phrasal verbs		idiomatic sentences
grammatical collocations		quotations

Other researchers have focused on the word-for-word decodability of an expression (i.e. its compositionality) to define formulaicity. Fernando and Flavell (1981:17), for example, proposed the following criteria for 'idiomaticity', with compositionality at the top of the list:

- the meaning of the idiom is not the result of the compositional function of its constituents;
- an idiom is a unit that either has a homonymous literal counterpart or at least individual constituents that are literal, though the expression as a whole is not interpreted literally;
- idioms are transformationally deficient one way or another;
- idioms constitute set expressions in a given language;
- idioms are institutionalized

Likewise, observing that the number of sequences to be learned in English is potentially daunting for one attempting to learn the language as an L2, Grant and Bauer (2004) used compositionality to narrow the field down to the ones most likely to cause interpretation problems, which they called 'core idioms', also mentioned in Section 2.2.2. Core idioms are defined by <u>not</u> fitting into any of the following three sub-types compositionality:

'non-compositional' (the entire expression)	the meaning does not change if any given word in the expression is interpreted literally Ex. *by and large* → **by and big*
'figurative'	the meaning can be derived from a word or words in the expression that have a metaphorical mapping Ex. *on top of things*
'ONCE'	contains only One Non-Compositional Element Ex. *<u>perfect</u> stranger*

However, both the Fernando and Flavell and Grant and Bauer criteria can be considered problematic because compositionality, as will be discussed below, is a function of one's interpretation of an expression, and not an attribute inherent to the item itself. Moreover, like the Leech et al. definition of multiword units for their study, it is possible that the Grant and Bauer classification in particular is excessively restrictive. For example, in their study the authors classify *have designs on sb/sth* as being a 'ONCE', since adverb *on* has one of the "dictionary" meanings (p. 57). Grant and Bauer would therefore claim that the expression is not a core idiom, and would further recommend excluding such an expression from a list of idioms worth learning in an L2 context.

By contrast, other authors advocate a less radical view of compositionality in the identification of multiword expressions. Lewis (1993: 98) proposed that formulaic language can be variously distributed on what he termed a "spectrum of idiomaticity" (Figure 2.4), a kind of continuum of compositionality:

> Technically, an idiom is a particular kind of lexical item, defined by the fact that the meaning of the whole is not immediately apparent from the meanings of the constituent parts. It is unsurprising that the transparency of idiomatic expressions is a matter of degree. (Lewis, ibid.)

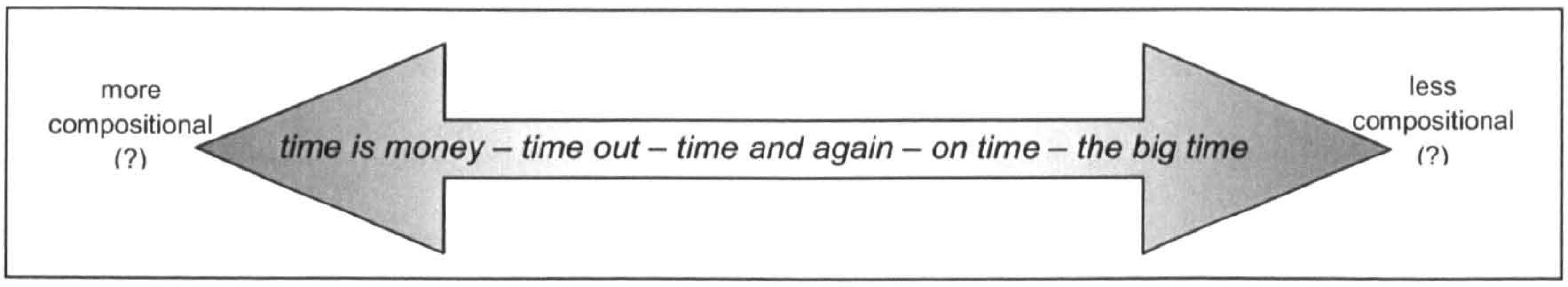


Figure 2.4. A spectrum of idiomaticity (compositionality) (Lewis, 1993)

This view is also supported by more recent corpus-informed research conducted by Wulff (2008), who following detailed analyses correlating native-speaker and computerized compositionality rankings of multiword expressions, concludes that "the difference between constructions qualifying as core idioms and other more or less idiomatic constructions is a matter of degree" (p. 2). Wulff further emphasizes that "idiomaticity in its entirety is a purely psychological construct, which is only real in the head of a speaker" (p. 3). Therefore, even when an expression does not meet the criteria of 'core' idiom, the relative ease or difficulty with which a learner will unpack its meaning is less inherent to the item itself and more a learner-dependent variable. As discussed earlier in the critical examination of the construct of word family in Section 2.2.1, just as knowing *fish* may or may not translate into understanding *fishy*, neither does knowing *perfect* necessarily mean understanding *perfect stranger*.

This could be seen as potentially good news in terms of addressing the original concerns raised by Read and Nation at the beginning of the present section (regarding the "essential" criterion of a sequence constituting a "whole unit"), as it is therefore impossible to assert with absolute certainty that any given item does not 'decompose'. On the other hand, that would also be suggestive of a lack of theoretical falsifiability with respect to the validity of formulaic sequences as a construct, which could then undermine any work involving them as a central construct. Therefore, while direct evidence cannot be obtained, a critical mass of indirect evidence of formulaic sequencing not 'decomposing' may at least help.

At least in the case of very common multiword expressions, there is some theoretical support in the literature (e.g. Arnon & Snider, 2010; Ellis & Frey, 2009). Bybee (2006), for example, has noted that some of the most common formulaic sequences have actually gone through a process in which their compositional meaning has been lost due to the frequency of repetition of those sequences, a process Bybee calls 'grammaticization':

> [A]s a particular string grows more frequent, it comes to be processed as a unit rather than through its individual parts. As it is accessed more and more as a unit, it grows autonomous from the construction that originally gave rise to it. (Bybee, 2006: 720)

Bybee cites examples such as *going to* (e.g. *I'm going to think about it*), which now is taught as a grammatical marker to refer to the future in instructed English language contexts, but which originally only literally meant *go + to*. Bybee finds evidence for the phenomenon in such diverse areas as speed of articulation and even the documented etymology of the expression, and suggests that grammaticization may be more ubiquitous in language than is currently accounted for in the literature on vocabulary and grammar. Indeed, considering Bybee's notion that when a formulaic sequence begins to "grow autonomous from the construction that originally gave rise to it" that it can be an indication of the word combination being "processed as a unit", an argument can be made that common 'mistakes' made in speaking or writing formulaic sequences may be evidence of a lack of compositionality, as in the case of corpus-attested variants like **once and a while*, **all of the sudden*, and **for all intensive purposes* (Brians, 2008). The point is, as Read and Nation themselves suggest, what is needed is an "eclectic approach" to help operationalize such an elusive construct (2004: 33).

Regardless, even though it may not be feasible to directly and empirically address Read and Nation's main concern regarding "storage as a whole unit" – which they assert may affect the validity of such instruments as vocabulary tests that include formulaic sequences – the point may be addressable from a theory-driven standpoint. Indeed, this is largely the purpose of what Wray (2008) calls a 'Morpheme Equivalent Unit' (MEU), defined as

> a word or word string, whether incomplete or including gaps for inserted variable items, that is **processed like a morpheme**, that is, without recourse to any form-meaning matching of any sub-parts it may have. (p. 12, boldface added for emphasis).

Wray emphasizes that, unlike some of the criteria by other authors covered in this section, the MEU cannot be used to help sort out what qualifies as a formulaic sequence and what does not, but is instead a theoretical position on the holistic nature of multiword expressions. She points out the evidence that exists for the psycholinguistic and cognitive validity of multiword expressions, and on the basis of that evidence, posits that while it is not possible to declare with absolute certainty that any given formulaic sequence definitely is not decomposable, research (at least among native speakers) suggests that when it comes to multiword expressions, they are indeed processed and stored as whole units by default[15].

In the end, what can be surmised from the body of literature aiming to operationalize formulaic language is that no one definition or conceptualization can suffice. Phraseology is clearly an area of the lexicon fraught with complexity and overlap, and it seems to be the case that – at least for academic research purposes – the operationalization of a formulaic sequence will depend on the nature of the research (Wray, 2008: 99). Here, therefore, it will be assumed that 1) formulaic language is a real phenomenon; and 2) there exists an identifiable subset of formulaic language that can be used to form a list whose purposes are similar to those of the wordlists discussed in this chapter.

[15] Or what Wray (2002, 2008) calls the 'Needs Only Hypothesis'.

2.4 Summary and research questions

To conclude, the following has been argued thus far:

- In part due to a lack of availability of a list of the most common multiword expressions, vocabulary research, teaching and testing has tended to focus on individual orthographic words to the exclusion of formulaic sequences.
- Formulaic sequences are "as important as individual words" (Schmitt, 2010: 8), and should be given equal priority in language research and pedagogy that focuses on vocabulary.
- Current frequency-informed wordlists that have been used in research and pedagogy (e.g. to estimate vocabulary size) can be misleading as the most common words in such lists are often merely tips of phraseological icebergs (Figure 1.1).
- Estimates of how many words one needs to know in order to comprehend most texts may be inaccurate due to over-inclusion of derived word forms and a near total exclusion of formulaic sequences from the research, largely due to the sole availability of frequency data on single orthographic words.
- There is evidence that formulaic language is pervasive in naturally-occurring discourse, and while formulaic sequences have been shown to offer processing advantages to native speakers, they have also been shown to negatively affect comprehension among L2 learners; there needs to be a way to systematically incorporate formulaic sequences into L2 pedagogy.
- Formulaic language for pedagogical purposes should not be operationalized by way of one unitary construct, but rather by considering a confluence of factors, such as non-compositionality, fixedness and function – in short, giving priority to a criterion of pedagogical relevance and 'usefulness' from the perspective of the L2 learner.

It is therefore clear that a pedagogically-relevant list of formulaic sequences is needed, ideally one which provides frequency information so that it can be used in the same or similar contexts and instruments in which wordlists are used today. This in turn raises the following questions:

RQ1: From the perspective of L2 comprehension, which type of formulaic sequence should be given priority?

RQ2: How can sequences of the type defined in RQ1 then be identified and put into a list?

RQ3: How many items should a list of pedagogically-relevant formulaic sequences contain?

RQ4: How should a pedagogically-relevant list of formulaic sequences be presented?

These questions will form the basis of the discussion in the next chapter.

Chapter 3 – Development of a list of formulaic sequences: the PHRASE List

As discussed in Chapter 2, it is now widely accepted that vocabulary knowledge comprises both individual words and multiword expressions of various types, and that a deficit of knowledge and/or awareness of such expressions can negatively affect important aspects of second language development, such as reading comprehension. If this is the case, then it is important to ensure the inclusion of multiword expressions in language teaching, and to revisit existing vehicles of vocabulary teaching and assessment to integrate multiword expressions where they may currently be lacking. As discussed in the previous chapter, one tool that has been influential in both the teaching and assessment of vocabulary is the wordlist, such as the GSL. One limitation of such lists is that the focus has invariably been on single orthographic words, and so the teaching materials that have used them to systematically introduce vocabulary (e.g. graded readers[16]) have tended to not integrate multiword expressions in any principled way, and the tests that have drawn on them (e.g. the *Vocabulary Levels Test*, discussed in greater detail in Chapter 5) only assess single words, and this in turn may have the cyclical washback effect[17] of a continued trend of second language vocabulary instruction that provides only a partial representation of the lexicon.

[16] Books with controlled or otherwise simplified vocabulary, often used in childhood education in the L1, and in second language contexts, used by both adults and children to make reading in the L2 more accessible.
[17] The effect of testing on teaching.

The present chapter, therefore, describes the development of a list of phrasal expressions, entitled the *PHRASE List*, which is aimed at addressing the aforementioned gap in vocabulary lists. Section 3.1 provides an overview of previous attempts at creating lists of formulaic sequences; Section 3.2 reviews issues of conceptualization of a list of formulaic sequences; Section 3.3 reports on the development of the list itself; Section 3.4 offers a discussion of the results of the research, and Section 3.5 concludes the chapter.

3.1 Lists of formulaic sequences

In reality, lists of phrases in various forms have existed for hundreds of years. William Caxton, for example, published his *Dialogues in French and English* in 1483, and included phrases such as the following (cited in Watkin, 1996), in a section on 'Greetings':

- *Ye be welcome.*
- *Where haue ye ben so longe?*
- *I have not seen you in longe tyme.*

The above were situated in dialogues that were meant to help the traveler be able to interact when abroad, much like modern travel guides. Likewise, lists of phrases in the form of functional exponents[18] have existed in modern language teaching at least since Hymes' theory of Communicative Competence (Hymes, 1971) and Wilkins' theory of notional syllabus design (Wilkins, 1976) found widespread adoption in L2 pedagogy circles (and, perhaps more importantly, among textbook publishers).

However, what all the aforementioned works have in common is that they were not informed by large amounts of empirical data, but instead mostly by what an individual or a small group of individuals 'felt' was useful and/or important to know.

[18] The language needed to realize certain functions or speech acts (e.g. agreeing, greeting, inviting, etc.).

It was not until the advent of computers and the ensuing ability to analyze large amounts of language data for patterns that researchers began to produce lists of phrases that were based on more than their own intuitions. Some of the earliest and arguably most influential of such research was carried out by John Sinclair and his team at University of Birmingham in the 1980s, particularly on the Collins Birmingham University International Database (COBUILD) project. What was perhaps unique about the COBUILD project was that, although computers, concordancers[19], and learner dictionaries all existed prior to 1980, it was not until Sinclair and colleagues began work on COBUILD that all three endeavors converged (Sinclair, 1987). Although the initial purpose of the project was to produce a corpus-informed dictionary that English language learners could use, in the process the COBUILD team also began to realize that the commonest words in English tend to recombine to form phrases with unique meanings far more than had previously been accounted for in lexicography (Sinclair, 1987: vii). Using concordancers, the COBUILD team produced corpus-based wordlists which they interrogated further, word by word, for collocational behavior. The product was a data-driven list of words that also included phraseology and went beyond mere intuition, and which eventually was published in the form of a dictionary[20].

Nonetheless, although inclusive of phrases, works of lexicography like those produced in the early years of COBUILD are not fully comparable with wordlists like the GSL. For one, although it can be argued that both are types of lists, and that both can inform pedagogy, one is used primarily to know more about words, while the other is primarily about which words to know. In other words, it would be difficult, or nearly impossible, to extrapolate a prioritized list of which words and phrases to know from a dictionary, even one that is corpus-informed like the ones which resulted from the COBUILD work (or even from

[19] Computer software that can automatically extract all instances of a given word or phrase in a corpus along with the words with which it most often occurs.
[20] *The Collins Cobuild English Language Dictionary* (1988).

most learner dictionaries today[21]). Without frequency data, such as those provided in most corpus-informed modern wordlists, even computer-generated lists of phrases are not much more useful to students and teachers than the one published by Caxton in 1483.

In the sections that follow, a few of the existing phrase lists that are both corpus-informed and inclusive of frequency data will be discussed.

3.1.1 'Word Frequencies in Written and Spoken English' (Leech et al., 2001)

Although corpora grew in their use by applied linguists over the years that followed the COBUILD work (e.g. Biber, Conrad & Reppen, 1994), and indeed work on computer-generated lists of 'multiwords' had already begun in the mid 1990s during the compilation of the BNC (Leech, Garside & Bryant, 1994), it is not until *Word Frequencies in Written and Spoken English* (henceforth, WFWSE) 2001 that the first real attempt at a comprehensive, corpus-based, computer-generated, frequency-ranked list of words <u>and</u> phrases is published[22] (Figure 3.1).

[21] Many learner dictionaries today do indicate frequency (e.g. via a symbolic rating system adjacent to headwords, or by setting the most frequent words in bold typeface), but do not report numerical frequency data, thus precluding any possibility of relative frequency comparison between words or phrases.
[22] There were attempts at including phrases in previously published lists, such as West's original GSL (1953) and Hindmarsh's *Cambridge English Lexicon* (1980), but they were based on manual counts of relatively small corpora.

Word	POS	Derivations	Frequency (p/million)
A / a	Lett	:	268
A	NoP		38
		A	10
		A.	28
a bit	Adv	:	119
a great deal	Adv	:	14
a little	Adv	:	104
a lot	Adv	:	40
abandon	Verb		44
		abandon	12
		abandoned	26
		abandoning	5
		abandons	1
abbey	NoC		20
		abbey	19
		abbeys	1
Aberdeen	NoP		14
		Aberdeen	14
ability	NoC		105
		abilities	13
		ability	91
able	Adj	:	304

Figure 3.1. Sample from the alphabetical list portion of WFWSE (Leech et al., 2001)

Leech, Rayson and Wilson (2001) were able to generate their list by running the entire 100 million word BNC through an automatic part-of-speech (POS) tagger, in this case CLAWS (also discussed previously in Chapter 2). Along with being able to identify and label the word classes of the individual words in a corpus, the CLAWS tagger also allows for 'ditto' tagging (Blackwell, 1987), meaning applying the same tag to a series of words when that series is recognized as one grammatical unit:

> For example, *so that* is made up of two word strings but functions in the same way as a one-word conjunction: it simply does not make sense to analyze it (say) as an adverb preceding a conjunction. (Leech et al., 2001: 14)

However, as discussed in Chapter 2, this automated method of identification has a number of limitations, including being limited solely to phrases with immutable forms (e.g. *in order to, in accordance with, with respect to*) for which CLAWS could reliably assign a grammatical function (e.g.

'preposition'). This means that important lexical items that do vary (e.g. by inflection, separation), such as phrasal verbs (e.g. *set up* → *setting up*, *set* something *up*), could not receive ditto tags automatically and are therefore not represented at all in WFWSE. There also was a complete reliance on statistical data, and unlike the lists that were produced before the era of widespread use of computerized corpora, there was no attention in WFSWSE to the individual senses of the words or phrases listed. Further, even when items would qualify as multiword expressions under the automatic ditto-tag CLAWS system, the mechanism does not have the same degree of accuracy as its other, single-word tagging system[23]. On the other hand, the lists in WFWSE do provide some valuable information, such as frequencies for both written and spoken corpora within the BNC, and the differences between them for each word.

Limitations regarding the type of methodology employed by Leech et al. in order to indentify phrases will be taken up again later in this chapter.

3.1.2 The Academic Formulas List (Simpson-Vlach & Ellis, 2010)

Simpson-Vlach and Ellis (2010) sought to compile a list of the most useful formulaic sequences used in Academic English. In order to do so, the authors used a combination of corpora, including the Michigan Corpus of Academic Spoken English (MICASE) and the BNC files that contained academic spoken English, then went about choosing a method of identifying the phrases for their list. The researchers decided to avoid a pure "lexical bundle approach of Biber and colleagues" (Simpson-Vlach & Ellis, 2010: 4), an approach discussed in Chapter 2 of this thesis, as lexical bundles present sequences as *"at the same time"* and *"to do with the"* as having equal psycholinguistic salience even though they instinctively do

[23] In fact, there are not even any data available for how accurate the ditto-tags are, only that the items tagged as multiwords ('<mw>') "should not be included in any assessment of the CLAWS error rate." (http://www.natcorp.ox.ac.uk/XMLedition/URG/codes.html#defrobs)

not (Simpson-Vlach and Ellis, ibid.). On the other hand, the authors continue, methods of phrase identification that prioritize pure intuition can be too open to subjectivity. Therefore, Simpson-Vlach and Ellis attempted to arrive at a metric to reconcile both approaches.

The authors first extracted n-grams[24] from their corpora and then calculated mutual information[25] (MI) – a measure of strength of association between words – for each of the phrase candidates. However, as pointed out by the authors, although MI can help separate more meaningful n-grams from pure lexical bundles, one reason for this is that MI also tends to identify relatively infrequent words that co-occur.

Therefore, in order to determine which quantitative information (e.g. frequency, n-gram length, MI score, or combination) would help them best inform their ultimate metric, Simpson-Vlach and Ellis recruited twenty native-speaker judges (with language testing and teaching experience) to rate a stratified random sample of the formulas on the basis of the following criteria (p. 10):

> A. whether or not they thought the phrase constituted 'a formulaic expression, or fixed phrase, or chunk' [...];
>
> B. whether or not they thought the phrase has 'a cohesive meaning or functions, as a phrase' [...];
>
> C. whether or not they thought the phrase was 'worth teaching, as a bona fide phrase or expression' [...].

Simpson-Vlach and Ellis were then able to correlate the qualitative judgment data with the quantitative statistics and, through multiple regression, arrive at a metric could be applied to all quantitatively-

---

[24] Word combiations that recur in a corpus, identified by specialized software. N-grams can be of any length, usually determined by the researcher.

[25] According to Manning and Schutze (1999), MI is a kind of "measure of how much one word tells us about another" (p. 178). In practical terms, a high MI score indicates that when one word appears it is likely that it will also appear with the other. An example is the word 'torrential' which is strongly associated with 'rain', and therefore 'torrential rain' would usually be assigned a high MI score in most corpora.

derived formulas and predict which ones would be worth teaching (or 'formula teaching worth' – FTW), which ended up mostly being MI, with some influence from frequency (β 0.56 MI + β 0.31 frequency). Therefore, the items in the Academic Formulas List (AFL) are in theory prioritized by this FTW metric (Table 3.1), with formulaic sequences most likely to be deemed useful listed first.

It is worth highlighting that the native-speaker judges, who only examined a subset of the formulaic sequences, were used to help inform the multiple regression alone – their judgments did not directly influence the selection of items that ultimately made it to the AFL. As the authors point out, such strict adherence to pure statistical selection criteria virtually eliminates possible "claims of subjectivity" (p. 4); however, as the criteria (A, B, C above) did not actually guide the selection, many items in the AFL – particularly those with lower FTW ratings, might be seen as only marginally having "cohesive meaning" as a "bona fide phrase" (see sample Table 3.1). Moreover, it is important to note that the items are not ranked by how commonly they occur in discourse, which is also a departure from most current wordlists.

Table 3.1 Spoken AFL Top 10

		Speech		Writing		
		Raw freq	Freq per million	Raw freq	Freq per million	FTW
1	be able to	551	256	209	99	2.96
2	blah blah blah	62	29	0	0	2.92
3	this is the	732	340	127	60	2.77
4	you know what I mean	137	64	4	2	2.27
5	you can see	449	209	2	1	2.12
6	trying to figure out	41	19	2	1	2.05
7	a little bit about	101	47	0	0	2.00
8	does that make sense	63	29	0	0	1.99
9	you know what	491	228	4	2	1.99
10	the university of michigan	76	35	1	0	1.98

3.1.3 *'The Most Frequent Collocations in Spoken English' (Shin & Nation, 2008)*

In another formulaic-list related study, Shin and Nation (2008) sought to identify the most frequent collocations in spoken English. What the authors call a collocation for their list is any "group of two or more words that occur frequently together" (p.341) – a definition seemingly open enough to include a wide variety of phrase types. The researchers used a concordancing package to look for all the words that most commonly co-occurred with the highest-frequency content words (called "pivot words" in their study) in the 10 million word spoken portion of the BNC, and established six identification criteria for a manual checking of the collocations, involving such aspects as frequency (e.g. all pivot words belonged to the first 1000 most frequent content words, all collocation candidates had to occur at least three times per million words) and grammatical well-formedness (collocations could not cross immediate constituent boundaries[26]). At the end of the analysis, the researchers report identifying 4,698 collocations, with each criterion always met (p. 343). Uniquely among the phrase lists evaluated so far in the present section, Shin and Nation also considered semantics, particularly individual senses of collocations with the same form (e.g. *looking up* meaning 'improvement' and *looking up* as in 'find a word in a dictionary'). Although semantics were considered, and while the list is intended to be used as a tool in L2 pedagogy, the authors do not consider 'usefulness' in their selection criteria (an element present in criterion 'C' ["worth teaching"] in the AFL native speaker judgment exercise). A possible consequence of this is evident in Table 3.2, where 'collocations' like numbers 7 and 9 would not normally seem to present comprehension or learning difficulties to learners, relative to a collocation like *as well* and *used to* in that same list, which have meanings that are not as easily derived from a literal reading of those words. On the other hand, the list is presented in descending frequency order, thus

[26] An immediate constituent is a component that makes up a larger part of a sentence, such as the words in a verb phrase, the last word of which would constitute a constituent boundary.

providing a tool that in theory allows for prioritization of items that reflects their relative importance in the language.

Table 3.2 Top 10 'collocations' – Shin and Nation (2008)

	Collocation
1	you know
2	I think (that)
3	a bit
4	used to {INF}
5	as well
6	a lot of {N}
7	{No.} pounds
8	thank you
9	{No.} years
10	in fact

3.1.4 Interim summary

Section 3.1 has reviewed some of the more prominent lists of formulaic sequences that are in existence today. One lesson they seem to offer collectively is that simply getting a computer to extract words and even co-occurring combinations of words is not difficult. Depending on the method used (e.g. n-grams, MI score, ditto tags) the type of output may vary, but specialized corpus software packages allow computers to carry out the same task that researchers also attempted without computers – like West for his GSL – in a fraction of the time. What the lists presented in Section 3.1 all seem to indicate is that the more challenging element when compiling a list of phrases is coming up with criteria for inclusion. Single orthographic words are easily identifiable by computer and there will be little debate by anyone that a string of letters bordered by a space constitutes a word; however, as has been seen in the lists presented here, the same consensus does not hold for multiword expressions. Furthermore, in all cases there seems to be a quantitative/qualitative tradeoff decision to be made. Using the CLAWS tagger, for example, the WFWSE was able to reliably and automatically include certain grammatical phrases – but by no means would anyone suggest that these are representative of the majority of formulaic sequence types in English. Unfortunately, in order to include a wider variety of phrase types a much greater degree of human intervention, and therefore time and labor, would be needed. Likewise, although the metric that was employed in the n-gram list extracted for the Simpson-Vlach and Ellis AFL enabled an efficient and uniform application of their selection criteria without subjective judgment, one can see how if the same native speaker judges that informed the metric were to scrutinize the list that ultimately emerged from it, applying the same criteria they were asked to follow in the rating exercise, many of the items currently on the list would probably be excluded. Finally, the Shin and Nation list of collocations illustrates the value of manual selection of items, as in their list all collocations reflect their true relative, sense-sensitive frequencies. However, once again, it is also seen how without criteria to account for usefulness (beyond frequency) – which would seem to invariably necessitate actual human judgment – items can end up on a list that perhaps would not be deemed by practicing teachers, or even students, as particularly deserving of special attention. A list of phrases that is also pedagogically relevant, therefore, would appear to need to incorporate both a machine-driven automated component, and a human-judgment-informed set of criteria for ultimate inclusion. To reiterate the research questions presented in Chapter 2:

RQ1: From the perspective of L2 comprehension, which type of formulaic sequence should be given priority?

RQ2: How can sequences of the type defined in RQ1 then be identified and put into a list?

RQ3: How many items should a list of pedagogically-relevant formulaic sequences contain?

RQ4: How should a pedagogically-relevant list of formulaic sequences be presented?

3.2 Conceptualization of a list of phrasal expressions

To some extent, the increasing ease with which copious amounts of texts can be both obtained and analyzed has been a simultaneous boon and bust in phraseology. On the one hand, the ability to comb through thousands of pages of text for words and word patterns at a keystroke has made it possible to extract information about words (e.g. frequency, dispersion) and co-occurring words (i.e. formulaic sequences) that would not be realistic without the use of a computer. On the other hand, what has tended to emerge is something of an obsession with 'unsupervised' automated corpus extraction of words and phrases (e.g. Van de Cruys & Villada Moiron, 2007; Fazly, Cook & Stevenson, 2009[27]), sometimes resulting in lists of multiword sequences that are frequency-driven and useful if employed in certain descriptive or experimental research (e.g. Biber, Conrad & Cortes, 2004; Ellis, Simpson-Vlach & Maynard, 2008; Hyland, 2008), but not clearly usable as teachable or testable items per se. Unlike words and word families (Chapter 2) which a software package can identify by simply counting continuous strings of letters bordered by blank spaces or punctuation, meaningful and pedagogically useful formulaic sequences cannot be found without some consideration of semantics, and as was also discussed in Chapter 2, the sheer variety of multiword expressions that might fall under the superordinate term 'formulaic language' is vast. There simply is no clear-cut way of automatically detecting such diversity through computers without consideration of meaning (Sag, Baldwin, Bond, Copestake & Flickinger, 2002). Although inroads have been made in recent years towards their automated extraction using a combination of semantic and grammatical tagging, for example (Katz & Giesbrecht, 2006; Korkontzelos & Manandhar, 2009; Piao, Rayson, Archer & McEnery, 2005), no computer application yet designed can replicate the qualitative judgments regarding individual

[27] Most of the research on developing computer applications to indentify formulaic sequences has emerged from the Natural Language Processing (NLP) field, which is concerned with developing 'intelligent machines' that can understand and produce language in a way that approximates human language processing.

multiword expressions that most native speakers apparently make unconsciously, intuitively, consistently and instantaneously (e.g. Deignan, 2009; Wulff, 2008).

Nonetheless, it is clear that the use of computers and concordancing packages will enhance the compilation of lists of formulaic sequences. This is not particularly surprising given that formulaic sequences are often difficult to intuit (Fox, 1987). While some formulaic sequences are quite salient (e.g. idioms like *raining cats and dogs*), others like *take place* (i.e. 'occur') are perhaps not. An easy illustration of this is attempting to determine the most frequent formulaic sequences in English by intuition. While it is probably possible to think of a number of these sequences, it is unlikely that the list would be very comprehensive, or that the relative frequency of occurrence could be stated with any degree of confidence (cf. Alderson, 2007). The limitations of intuition mean that language teachers, textbook writers, and test developers require a more principled manner of identifying and ranking formulaic sequences. The obvious solution is a list of frequent <u>and</u> useful formulaic sequences to which they can refer.

As reviewed in Chapter 2, wordlists have a long history as useful pedagogical tools, but provide an incomplete picture of the lexicon. Lists of formulaic sequences have been relatively slow to emerge in part because the software and computing power necessary to identify formulaic sequences has only relatively recently become available, but it is also partly because of a general lack of agreement on issues as fundamental as how to classify formulaic sequences, their nomenclature, and even what should be considered a multiword lexical item in the first place (see Chapter 2). Therefore, although some advances have been made in the automated extraction of formulaic sequences of different types, asking a computer to produce a list of expressions is still somewhat akin to writing 'buy fruit' on one's shopping list – and sending someone else to do the shopping.

The following sections will therefore report on the construction of a list multiword expressions deemed particularly useful for pedagogy – a process that necessarily involved both automated and manual selection of items. The list itself is provided in Appendix 1 in full.

3.2.1 Issues of frequency

The number of formulaic sequences is large in language, both in terms of frequency of occurrence and variety. So it is impractical (and probably not very useful) to attempt to compile a comprehensive list of all occurring sequences. As seen in the review of existing lists of formulaic sequences (Section 3.1), it is invariably necessary to delimit the field by some means, often by applying selection criteria to a list of candidate expressions.

Perhaps the starting point for any lexical list should be a determination of its purpose(s). In the case of the research endeavor under discussion here, the main objective was to create a list which would have pedagogic utility, mirroring purposes similar to the GSL and AWL lists, but for formulaic sequences. These purposes include, but are not limited to, the following:

- a guide for language learners and educators to include formulaic sequences in their learning and teaching, particularly for receptive purposes;
- a means of including formulaic sequences in tests that assess receptive L2 knowledge and receptive skills;
- an aid in monitoring the vocabulary acquisition progress.

Pedagogic purposes like the above dictate that the list needs to at least focus on the most frequent formulaic sequences in English. It is widely accepted that frequency of occurrence is one of the best indicators of usefulness of individual words in general English (e.g. Nation, 2001; O'Keeffe et al., 2007). For example, the GSL, used since the 1950s as a model of a pedagogically-based word list, drew on a number of selection criteria, but the essential one was frequency. This is true to the extent that it was often used as an indicator of the most frequent 2000 word families in English before more modern word counts came along. There is no reason to believe that this frequency-usefulness relationship does not also apply to formulaic language:

> Some items larger than a word behave like high frequency words. That is, they occur frequently as multiword units (*good morning, never mind*), and their meaning is often not clear from the meaning of the parts (*at once, set out*). If the frequency of such items is high enough to get them into a general service list in direct competition with single words, then perhaps they should be included (Nation & Waring, 1997: 18).

However, while frequency can be considered a valid indicator of usefulness, the list must stop at some point, so the issue of extent must also be taken into account, and where a list ends should be based on some sort of rationale. For example, the extent of the GSL is about 2000 entries. As discussed in Chapter 2, this threshold seemed adequate given the information available around the time that list was created. (For example, Schonell et al. (1956) found that 2,000 word families covered 99% of the spoken discourse of the Australian workers they studied.) However, more recent research indicates that a much wider vocabulary is necessary to communicate in English (e.g. Nation, 2006 – see discussion in Chapter 2). If a list of formulaic sequences stopping at the same frequency as the 2000 word family frequency level is obviously too small, surely a list extending to the 9000 word level (as suggested in Nation, 2006) would become too unwieldy for practical use. Nonetheless, there is some evidence that even students of English who never progress to such a large vocabulary size are sometimes still able to pass tests of general proficiency in English at more advanced levels. For example, Hindmarsh (1980) in his *Cambridge English Lexicon* found that 4500 words would provide coverage to FCE (Cambridge First Certificate in English) level, which itself has been aligned with the B2 band of the Common European Framework of Reference (CEFR)[28]. Hindmarsh's lexicon, in turn, was used in conjunction with a number of other corpora by the *English Profile Wordlists* project in 2009 (discussed in detail in Chapter 4) to compile a wordlist with levels aligned with the CEFR A1-B2 – ultimately arriving at a list totaling 4667 items (Capel, 2010). This is also consistent with Milton (2009), who affirms that "[s]tudents who take advanced level examinations would probably be expected to recognize over 4500, or 90% or more, of this corpus (of 5000 words)" (p.180). Using American corpora, Davies and Gardner (2010) in the *Frequency Dictionary of Contemporary English* also set the limit of their frequency lists at 5000 words as representing the most useful words of general English vocabulary. It would therefore seem plausible that, while knowledge of 9000 word families will allow comprehension of authentic written material that approaches native-like levels, 5000 word families may vaguely represent a ceiling of general, high-frequency vocabulary.

Frequency also has implications for the length of formulaic sequence that will be included on the list. There is an inverse relationship between length of formulaic sequence and frequency of occurrence, that is, the longer the sequence, the less likely it is to occur frequently. For example, O'Keefe, McCarthy and Carter (2007) found over 21,000 two-word sequences in the CANCODE corpus (with a frequency cut-off of four per million words), but 13,514 three-word sequences, and only 2,819 four-word sequences (p. 65). (Occurrences of five-word and six-word sequences were practically negligible: 262

[28] There is still little agreement with regard to the appropriateness of the alignment of the CEFR to examinations such as those of Cambridge ESOL. As stated in the ALTE *Manual for Language Test Development and Examining* (2011), "...it is important to remember that the CEFR is not intended to be used prescriptively and that there can be no single 'best' way to account for the alignment of an examination within its own context and purpose of use" (p. 8).

and a mere 18, respectively.) It was therefore established that a search for n-grams of no longer than four contiguous words would prove most fruitful.

3.2.2 *Issues of identification*

Broadly speaking, there are two general approaches to identifying formulaic sequences: one which uses frequency as the main criterion, the other which primarily considers semantics/grammar, or what Nesselhauf (2004) has called the 'frequency-based approach' and 'phraseological approach', respectively. As a list of phrases that is meant to help inform L2 pedagogy must take meaning and usefulness into account, the compilation should not be completely driven by frequency as one could end up including sequences like *is the* or *is of a* which encode very little meaning in themselves (cf. De Cock, 2000). A pedagogically-relevant list of phrases, it seems, should include only formulaic sequences which realize meanings or functions, in order to be of maximum utility. Hence the meaning/function aspect should surely be targeted as a selection criterion when going through any initial n-gram (i.e. frequency-based) corpus extraction. However, as was seen in the Shin and Nation list, items can convey discrete meanings and yet seem almost too obvious or transparent in meaning to really merit inclusion. Therefore, the relative semantic opacity of formulaic sequences should also inform selection criteria. Consider, for example, the following three expressions:

- *at all*
- *at all costs*
- *at all times*

Although all three expressions tend to occur as phrases according to the BNC, they differ in *compositionality*, as shown in Figure 3.2.

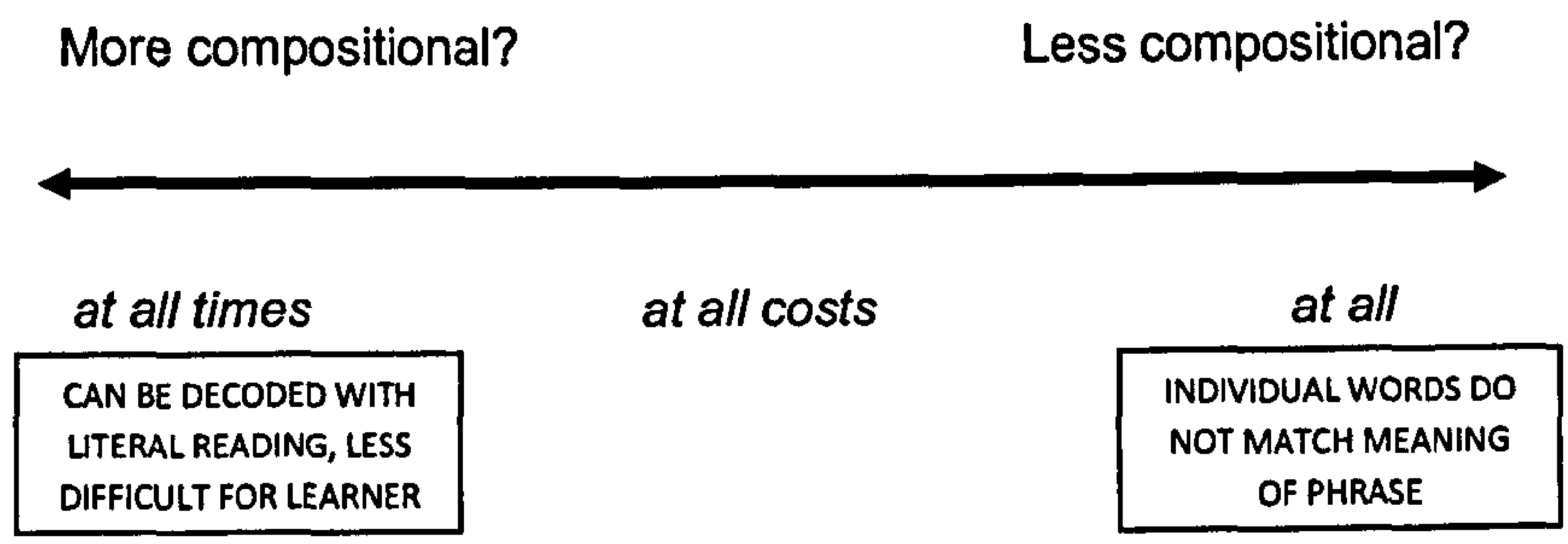


Figure 3.2. Degrees of compositionality for formulaic item selection

As seen in the previous chapter, Lewis (1993) observed that expressions vary in terms of the degree to which "the meaning of the whole is not immediately apparent from the meanings of the constituent parts" (p. 98), and called this varying compositionality a "spectrum of idiomaticity" (ibid.). In Figure 3.2, it could be argued that, while precise divisions are impossible to pinpoint, the expression *at all times* can be understood relatively easily from the meanings of its three component words, or as described in Grant and Bauer (2004), its meaning is still retained when each lexical word is replaced with its own definition (p. 52). The phrase *at all costs* has a more figurative, and likely less transparent, meaning – potentially making that item more difficult for a learner unfamiliar with that expression (Cooper, 1999; Spöttl & McCarthy, 2004). On the far right of the spectrum lies *at all* (e.g. *Do you exercise **at all**?*), whose individual components offer no more help to someone meeting that expression for the first time than do the individual letters that spell a word. In essence, *at all* behaves like a word in terms of form-meaning link. In terms of pedagogic value – especially with respect to so-called 'receptive skills' – it does not seem efficient to include compositional phrases which can be easily analyzed for meaning.

This avoidance of completely transparent formulaic sequences for the list can be said to also approximate the rationale for the Bauer and Nation (1993) concept of a word family discussed in Chapter 2 – a concept which forms the basis of the words which make up not only wordlists but also numerous vocabulary tests that draw on them. For example, if a vocabulary test assesses knowledge of the word *estimate*, Bauer and Nation would probably argue that it makes little sense to also test *underestimate* on the same instrument, since at least some of the knowledge of *estimate* is included in *underestimate*. (However, see critical discussion of the word family concept in Section 2.2.1.) Likewise, with a formulaic sequence like *at all times* – even if the item can be called (in Wray's terms) an MEU (e.g. 'constantly') – it would be nearly impossible to know on a vocabulary test if the candidate got that item right because (s)he recognizes the phrase as a whole, or because (s)he knows all the words in the phrase. Therefore, in the type of formulaic sequence list being targeted in the present research, items judged as lying to the far left of the spectrum of idiomaticity should not be included, preferring instead ones which learners may find difficulty in interpreting if they try to derive the meaning from the component words.

Considering all of the aforementioned issues, selection criteria can be arrived at which revolve around high frequency, meaningfulness, and relative non-compositionality. These criteria would select formulaic sequences that would in many ways be comparable to the individual words in a typical frequency-based wordlist, which in the case of the present research is as the ultimate list that results from the investigation should be meaningfully comparable to these wordlists. If so, then in conjunction with existing wordlists it should also result in a more inclusive overall description of the most frequent (and therefore useful) lexical items of English, both individual- and multi-word (see also Fig. 3.14, p. 91). It would also make it possible to insert these formulaic sequences into frequency-based vocabulary tests in order to gain a more valid measure of overall receptive vocabulary knowledge (see also Chapter 5).

3.2.3 Issues of taxonomy

The last issue of conceptualization concerns nomenclature. As reviewed in Chapter 2, terminology in the area of phraseology has always been somewhat disparate and inconsistent, with Wray (2002: 9) finding over fifty terms to describe the phenomenon of formulaic language. For example, in the Shin and Nation (2008) list described earlier, some examples of what constituted collocations in their work include *thank you, come on,* and *over there,* words which they identified as simply commonly co-occurring. This definition of a collocation would be in line with the thinking of other prominent researchers (e.g. John Sinclair) who subscribe to a 'frequency-based approach' (Nesselhauf, 2004), however differs somewhat from thinkers such as Cowie (e.g. 1981, 1984) who subscribe to a more 'phraseological approach' (Nesselhauf, ibid.) and would not call *thank you* and *come on* collocations, but instead 'formulae'. Moreover, these 'collocations' may variously be termed 'multiword units', 'prefabs' or 'lexical phrases' and still other names by different authors (cf. Biber et al., 1999; Lewis, 1993; Nattinger & DeCarrico, 1992). In the end, while one could quibble over the precise definition of such terms, perhaps the diversity of names given to the various types of phrases discussed in the literature is not dissimilar to the well-known parable of the blind men and the elephant, all touching on different parts of the same thing[29]. As Weinert (1995) has wisely put it, "[w]hile labels vary, it seems that researchers have very much the same phenomenon in mind" (p. 182). Binomial expressions (*cut and dried, up and running*), phrasal verbs (*take over, set up*), adverbial expressions (*in other words, for instance*), idiomatic expressions (*beat around the bush, over the moon*) – they are all multiword,

[29] According to Wikipedia, the 'Blind men and an elephant' parable originated in India and is "used to illustrate a range of truths and fallacies." The story tells of six blind men who are asked by the king to determine what an elephant looks like by touching it. One of the blind men touches a leg and says it feels like a pillar, another touches the tail and says it feels like a rope, and so on. The king then says, "All of you are right. The reason every one of you is telling it differently is because each one of you touched a different part of the elephant. So, actually the elephant has all the features you mentioned."

formulaic sequences of some kind – and though each can be assigned a different label, that does not mean that they are necessarily represented in different ways in the mental lexicon (Wray, 2008). Schmitt (2010) has proposed *formulaic language* as the umbrella term for the range of phrasal units which occur in language, and *formulaic sequence* as the term for each individual case of this phenomenon, a convention which as been followed thus far in the present thesis. However, the formulaic sequences which will be identified by the selection criteria here will clearly represent a limited subset of formulaic language, and therefore need a discrete descriptive name. After considering many possibilities, a transparent term was decided on, and therefore the particular category of formulaic language that will be identified for the proposed list will be called *phrasal expressions*. Thus, hereafter the list will consistently be referred to as the *PHRASal Expressions List*, or the *PHRASE List*.

3.3 Compiling the PHRASE List

Section 3.2 presented some of the central issues that were carefully considered in the planning stages of the PHRASE List. As mentioned in that section, what was realized as key to creating a valid and useful list was the formulation of effective and relevant criteria, which will be presented in the present section.

3.3.1 Honing the criteria

The first step in the compilation process was to operationalize the general criteria reached in the conceptualization stage. As discussed in the previous section, it was determined that the multiword items in the PHRASE List should not be easily 'decomposable' (i.e. understandable from individual parts), remaining consistent with the underlying constructs of wordlists like the GSL and the pedagogical instruments derived from it, such as the VLT. Automated identification can reliably identify only some of the potential PHRASE List candidates that would meet this condition, as discussed in Section 3.1 of this chapter. Moreover, since phrasal expressions are formulaic sequences that are identified by semantic rather than formal properties alone (i.e. can be any multiword expression that meets the semantic criteria, irrespective of grammatical features), a challenge that needed to be met in pinning down the criteria was how to not make them too restrictive. The criteria to some degree would need to be relevance-sensitive, applicable to certain phrases but not necessarily to others. Wray and Namba (2003), for example, outlined a set of eleven criteria designed to help the researchers justify intuitions regarding what may or may not be formulaic from a raw dataset of potential candidates. As the criteria were designed to guide them in assessing any potential formulaic candidate, the diagnostics are somewhat broader in scope and outcome than those needed for the PHRASE List, which has a more specific intended application. The Wray and Namba criteria, however, are similar to the needs of the PHRASE List because, unlike other criteria, such as those in the Shin and Nation (2008) study, they are not cumulative (i.e. not all criteria are meant to necessarily be met). Moreover, their criteria were designed to be used post-hoc (i.e. to help justify strings first identified by computer), and are therefore in support of qualitative judgments. Those features match the intended use of the six criteria used in the present study, outlined below, which have been developed in consideration of every element so far discussed in the present chapter, in addition to theoretical issues explored in Chapter 2. After some initial trialing and experimentation (see Section 3.3.2) it was realized that, although it was desired that the criteria be flexible and not mutually exclusive, some kind of systematic selective hierarchy would be conducive to consistency. Therefore, the criteria were divided into two sets: the 'core criteria', and the 'auxiliary criteria'. The core criteria are those which were used to determine the candidacy of a given n-gram to inclusion in the list, while the auxiliary criteria were occasionally consulted to add support to decisions.

Phrasal Expressions: Core Criteria

1. Is the expression a Morpheme Equivalent Unit (MEU)? Wray (2008) has suggested that one definition of a phraseological lexical item is that it is processed as if it were one morpheme "without recourse to any form-meaning matching of any sub-parts it may have" (Wray, 2008: 12), and especially among high-frequency expressions, there is psycholinguistic evidence for this assertion (Kapatsinski & Radicke, 2009; Sosa & MacFarlane, 2002). Such a criterion is consistent with the construct of 'word'. After all, a person reading the word *might* does not break it down into any sub-parts: it is clearly one morpheme, processed as such. An example of an MEU, then, would be *might as well*, as one who knows the expression is unlikely to resort to form-meaning matching of its sub-parts. As noted in Chapter 2, however, 'morpheme equivalence' is more of a "theoretical position" that certain wordstrings "contain semantically viable parts that are not taken into account" when they are processed (Wray, 2009: 31). Therefore, deeming a formulaic sequence to be an MEU is not a hard and fast science, but the criteria start with this "theoretical position" first of all, and use indicators (below) to justify selection judgments (Wray, 2008: 113).

2. Is the expression semantically transparent? To reiterate, the general idea regarding the items to be included in the PHRASE List is that they should be ones which are identified as potentially causing difficulty for learners of English, particularly on a receptive level. The expression *at this time*, for example, may qualify as an MEU because it means essentially the same thing as 'now', but even a learner who has never met this expression before and who encounters it in a text for the first time would stand a very good chance of unpacking its meaning simply by virtue of understanding *at + this + time* (i.e. the meaning remains even if each component word is replaced with its own definition). However, like all the criteria used, this one demanded careful subjective evaluation for each potential item. As has been suggested by Taylor (2006), "[f]ull compositionality is rarely the case" (p. 61) in multiword exressions, and "[t]he distinction between the idiomatic and the non-idiomatic may not be so clear-cut..." (p. 62) – hence "even the simplest of collocations may contain difficulty for learners" (Lewis, 2000: 136). This was borne in mind for every expression considered.

3. Is the expression potentially "deceptively transparent"? This question is also related to the issue of compositionality. Laufer (1989b) has shown that some lexical items in English can be "deceptively transparent" – words learners "think they know but they do not" (p. 11). Examples include *every so often* (which can be misread as 'often') and *for some time* (potentially misunderstood as 'a short amount of time'). When selecting phrasal expressions, an item was also often judged to potentially fit into this category when the most common and familiar meaning of at least one of the words in the expression was likely to pose confusion, especially if even a dictionary would not offer clear-cut help. For example, a survey of three advanced learner dictionaries[30] for the word *further* shows that the first and highlighted senses of the word are to do with distance and extent. However, in the multiword expression *a further* ('another') that meaning does not hold, and in the dictionaries surveyed the definition of 'additional' is not found until the third or fourth senses in the entry.

<u>Phrasal Expressions: Auxiliary Criteria</u>

1. Does the expression have a one-word equivalent? For example, *put up with* is synonymous with *tolerate.* This is evidence of that expression being an MEU. Indeed, even if there is no one-word equivalent in English, but there is one in another language, it may also be evidence that the expression represents a single morpheme (Zgusta, 1967). For example, there is no other English equivalent for *used to*, but there is evidence in Spanish (*solía*) and Portuguese (*costumava*) that it

[30] The *Cambridge Advanced Learner's Dictionary* (2008), the *Macmillan English Dictionary for Advanced Learners* (2008), and the *Collins COBUILD Advanced Dictionary* (2009).

represents an MEU – not a series of separate words. Although not all items in the list necessarily must meet this criterion, the ability to roughly equate a multiword lexeme to a single one also facilitates its ability to be included in vocabulary tests with item formats that require form-meaning matching, like the Nation and Beglar (2007) multiple-choice *Vocabulary Size Test* (see Chapter 5).

2. Could the learner's L1 negatively influence accurate interpretation? Take, for instance, the phrasal expression *out there*, which on the surface may seem marginally semantically transparent. Although there is a metaphorical mapping at work (THE WORLD IS OUTSIDE) which ostensibly could make it easy to understand by a learner, the source domain does not necessarily operate the same way in languages other than English. The sentence *She wants a job but there's simply nothing* ***out there*** *right now* would be translated thusly in Portuguese: *Ela quer um emprego mas por enquanto não tem nada* ***por aí*** ('...there's nothing around...'). This criterion is also related to cognate words, of course. The word 'addition' in the phrases *in addition to* and *in addition* might seem at first glance to be easily decodable by a speaker of a Romance language, for example, but when one considers the formulaic equivalents in such languages (in Spanish: *aparte de, ademas*; Portuguese: *alem de, mais*; French: *en outre, en plus de*; Italian: *in piu, oltre che*, etc.) it seems plausible that a focus on the cognate may actually render a spurious interpretation.

3. Does the meaning and/or opacity of a word change due to the grammar of the expression? The expression *no doubt* may violate the precepts outlined in Wray's MEU definition (since recourse to sub-parts may occur), but consider the discoursal difference between *I have no doubt she'll arrive* and *The president has no doubt taken his share of criticism*. While the first sentence is likely readily interpretable, the grammar of the expression has changed in the second: it is still preceded by a subject, but as an adverb rather than a noun phrase. As such, it also potentially qualifies as being 'deceptively transparent'. This criterion was often particularly relevant to passive constructions. For example, the fact that a beginner recognizes the meaning of the word 'know' does not mean that the same learner will understand a sentence like 'He's *been known to* do that before'. The verb 'expect', according to most learner dictionaries, is related to what one 'thinks will happen'. However, that meaning does not really remain in examples like 'bathers *are expected to* shower before entering the pool' (= 'bathers *must* shower before entering the pool) and 'as a host *I'm expected to* be courteous' (= 'as a host *I'm supposed to* be courteous').

As in the Wray and Namba (2003) research, the criteria used for selection of expressions in the PHRASE List were consulted to qualitatively "reveal the basis of intuitions already made" (Wray, 2008: 116) about items in the initial quantitatively-derived list, and not as a cumulative list of prerequisites. However, all expressions had to meet at least one of the core criteria. What all the criteria had in common was that they were designed to help justify why it was thought the items chosen might pose some difficulty for a learner on a receptive level. Regardless, it is clear from the above criteria that the intuition, subjectivity and general heuristics involved in the decision-making process necessitated a qualitative approach that no computer can yet achieve. The criteria (and heuristics), in turn, were also influenced and informed by the author's 20+ years of English language teaching and teacher training experience in a broad diversity of educational contexts (e.g. monolingual, multilingual, ESL, EFL, test preparation, EAP, etc.). Although such a methodology is obviously time and labor intensive, the end result is a PHRASE List that one could argue is enhanced pedagogically. (See Section 3.3.4 for a report on how the criteria were validated in order to establish how consistently they can be applied.)

3.3.2 *Materials*

The next step was to decide on the corpus source of the language data for the list. After careful consideration, the full 100 million word BNC was deemed the best choice from among the publicly available large corpora for a number of reasons, including its size, diversity and reputation. It is widely accepted that size does matter when it comes to corpus studies (e.g. Biber, Conrad & Reppen, 1998; Hunston, 2002; Yang, Lee & Cantos, 2002), especially when conducting qualitative research to explore patterns – as is the case in the present study. The BNC offers one of the largest publicly available corpora, which can be purchased on DVD-ROM. Not only is the BNC sufficiently large in size for conducting quantitative analyses, it is also derived from fairly diverse sources[31], helping to avoid data possibly skewed due to over-sampling from too few data sources (Table 3.3).

Table 3.3 Text types represented in the BNC

	texts	words	%	sentences	%
Spoken demographic	153	4233955	4.30	610557	10.13
Spoken context-governed	755	6175896	6.27	427523	7.09
Written books and periodicals	2685	79238146	80.55	4395581	72.94
Written-to-be-spoken	35	1278618	1.29	104665	1.73
Written miscellaneous	421	7437168	7.56	487958	8.09

The fact that the BNC consists predominantly of written texts (Table 3.3) was appropriate to the present study since some of the more popular tests of vocabulary knowledge currently in existence – one of the envisaged applications of the PHRASE List – are of written word recognition. Other currently available corpora, such as the 22 million-word *American National Corpus* (Ide & Suderman, 2004) also rely mostly

[31] A full listing of all the sources contained in the corpus can be viewed at http://www.natcorp.ox.ac.uk/XMLedition/URG/bibliog.html.

on written data, but are far less diverse[32]. However, perhaps one of the best rationales for using the BNC for the present research is the fact that it is a tried and tested corpus, cited and used in more academic studies than any other corpus[33]. Moreover, the BNC is the corpus that has most recently been used in the design of vocabulary research, lists and tests (e.g. Leech, Rayson, & Wilson, 2001; Nation, 2006; Nation & Beglar, 2007; and the BNC-20 *Vocabulary Profiler* available on the *Lextutor* website) – instruments into which the PHRASE List may be usefully incorporated. If one were aiming to develop a test of mostly spoken English, for example, use of the full BNC would be of questionable validity; however, as it is hoped that the PHRASE List will eventually be used much like the GSL (i.e. researching reading comprehension, vocabulary levels tests, etc.), the makeup of the BNC more than suffices for the present research.

Of course, a corpus of the size of the BNC cannot be easily analyzed without the use of some kind of specialized software to be able to observe patterns using all the data contained in it. The main software package used for the analysis of the BNC was *WordSmith Tools*, version 5.0. There are in fact a number of different software packages available for the analysis of corpora (e.g. *MonoConc Pro* [Athelstan], *WMatrix* [Rayson, 2008]), but few have gone through as much refinement as *WordSmith Tools* (Scott, 2005), now in its fifth version. Moreover, *WordSmith Tools* in recent years has been developed to be easily compatible with the latest BNC XML edition (the one used in the present study). The package is really three programs in one: 'Concord' for concordancing, 'WordList' for making frequency lists, and 'KeyWord' for exploring salient lexemes. For this study, the researcher was primarily interested in the

[32] The Corpus of Contemporary American English ('COCA') (Davies, 2008-) is an exception, consisting of 385 million words from more than 150,000 spoken and written texts. However, the COCA is only available through web interface, and therefore was not an option for the present study, which often required consulting the source text. Furthermore, the present study is especially concerned with n-gram analysis, which is not available via the COCA website. N-grams derived from the COCA are available for purchase, but at a price of $1 (US) per 1,000 *n*-grams. The BNC generated a list of over 4 million n-grams, which means that an equivalently extensive list from the COCA would have cost over 4,000 US dollars. (And at any rate, the author in that case would be dubiously relying on a secondary source for data.)

[33] According to Google Scholar, accessed April 30th, 2009.

first two, as will be discussed further in the following section. Although the BNC DVD-ROM which was purchased for the study does come with corpus analysis software, it does not have the ability to search for recurring strings of words (or 'n-grams'), and hence could not be used.

3.3.3 Procedure

First, the entire 100 million word BNC corpus was downloaded from the DVD onto the hard drive of the computer used in the study[34]. Since the text files are unreadable in the XML format they are encoded in for the DVD, *WordSmith Tools* was used to convert all the files into a plain unformatted text. The 'Wordlist' application embedded in *WordSmith Tools* was then used to upload all the texts (over 4000) and construct what is called an 'index' of all the words in the corpus. An index analysis collects vital information about each word in the corpus (e.g. dispersion, collocation, and so on), and is necessary if one wishes to run an analysis of recurrent word strings. The full index of the entire BNC took approximately 7 hours of computer time.

Once the index was complete (a file of over 1.5 gigabytes), the list was further analyzed and restructured in a process of lemmatization and grouping into word families. There were two reasons for this. First, most existing wordlists used in pedagogy are not listed by individual word forms (e.g. 'go', 'going', 'went'), but in their lemmatized, base forms. Further, as will be discussed in greater detail in Chapter 5, certain tests of vocabulary knowledge – such as the Vocabulary Levels Test (Nation, 1990) – assume that a learner who can recognize *interest*, for example, will also readily discern the meaning of derived forms such as *interested, uninterested, interesting* and *interestingly* (i.e. the word family of 'interest'). However, *WordSmith* (as all corpus analysis software) simply recognizes word forms when compiling

[34] HP Pavilion series notebook computer with 1 gigabyte RAM and 100 gigabytes of ROM; AMD Turion64 processor, running Windows XP Home Edition.

frequency lists, so this process must be done manually. Lemma lists do already exist, but these simply include inflected forms. In order to maintain the same construct as used in the VLT (and frequency levels), a full list of all the word families used by Nation and others for the VLT was downloaded from the Victoria University of Wellington website (http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx). However, these lists had to be manually reformatted to be readable by *WordSmith Tools*, which entailed regrouping all the words and merging them into one text file (14,000 basewords, or around 62,000 different word forms). (An example of the formatting difference is provided in Figure 3.3.) This of course was quite time consuming, but once finished the software was able to very quickly re-analyze the entire corpus and re-group all the appropriate words into their respective base-words.

```
•
        AN 1
ABLE 4
        ABILITY 1
        ABLER 0
        ABLEST 0
        ABLY 0
        ABILITIES 0
        UNABLE 1
        INABILITY 1
ABOUT 1
ABSOLUTE 2
        ABSOLUTELY 1
        ABSOLUTIST 0
        ABSOLUTISTS 0
ACCEPT 4
        ACCEPTABILITY 0
        ACCEPTABLE 1
        ACCEPTABLY 0
        UNACCEPTABLE 1
        ACCEPTANCE 1
```

```
A -> AN
ABLE -> ABILITY, ABLER, ABLEST, ABLY, ABILITIES, UNABLE, INABILITY
ABSOLUTE -> ABSOLUTELY, ABSOLUTIST, ABSOLUTISTS
ACCEPT -> ACCEPTABILITY, ACCEPTABLE, ACCEPTABLY, UNACCEPTABLE,
ACCEPTANCE, ACCEPTED, ACCEPTING, ACCEPTS, UNACCEPTABLY
ACCOUNT -> ACCOUNTED, ACCOUNTING, ACCOUNTS
ACHIEVE -> ACHIEVABLE, UNACHIEVABLE, ACHIEVED, ACHIEVEMENT, ACHIEVEMENTS,
ACHIEVER, ACHIEVERS, ACHIEVES, ACHIEVING
ACT -> ACTED, ACTING, ACTION, INACTION, ACTIONS, ACTIONABLE, ACTS, ACTOR,
ACTORS, ACTRESS, ACTRESSES
ACTIVE -> ACTIVELY, ACTIVITIES, ACTIVITY, INACTIVE, INACTIVITY, ACTIVIST,
ACTIVISTS, ACTIVISM
ACTUAL -> ACTUALLY, ACTUALITY
ADD -> ADDED, ADDING, ADDITION, ADDITIONAL, ADDITIONALLY, ADDITIVE,
ADDITIVES, ADDITIONS, ADDS
ADDRESS -> ADDRESSED, ADDRESSES, ADDRESSING, ADDRESSEE, ADDRESSEES
ADMIT -> ADMISSION, ADMISSIONS, ADMITTEDLY, ADMITS, ADMITTED, ADMITTING,
ADMISSIBLE, ADMISSIBLY, INADMISSIBLE, ADMITTANCE, READMIT, READMITTED,
READMITTING, READMITS, READMITTANCE, READMISSION, ADMISSIBILITY,
ADMISSIBILITIES, INADMISSIBILITY
ADVERTISE -> ADVERTISING, ADVERTISES, ADVERTISER, ADVERTISERS,
ADVERTISED, ADVERTISEMENT, ADVERTISEMENTS, ADVERTIZE, ADVERTIZING,
ADVERTIZES, ADVERTIZER, ADVERTIZERS, ADVERTIZED, ADVERTIZEMENT,
ADVERTIZEMENTS, AD, ADS, ADVERT, ADVERTS
```

Figure 3.3 Two versions of word family formatting

This sorting into word families allowed the researcher to arrive at the crucial frequency band cut-off points (Table 3.4), such as those used in existing wordlists and vocabulary tests. Indeed, since each word form has its own token frequency in the corpus, and since that frequency is added to the frequency to whatever word(s) it is grouped with, the frequency counts increased substantially in a number of cases, particularly among the most common words.

Table 3.4 1000-level frequency cut-offs (BNC)

Frequency band	Token frequency cut-off*	**Frequency band**	Token frequency cut-off
1,000	12,271 +	**8,000**	434 +
2,000	4,455 +	**9,000**	356 +
3,000	2,089 +	**10,000**	295 +
4,000	1,217 +	**11,000**	249 +
5,000	787 +	**12,000**	213 +
6,000	620 +	**13,000**	184 +
7,000	547 +	**14,000**	162 +

* Per 100 million, including all tokens within a word family

The frequency band cut-off data in Table 3.4 might also be usefully visualized graphically as in Figure 3.4, further evidence of Zipf's Law (Chapter 2).

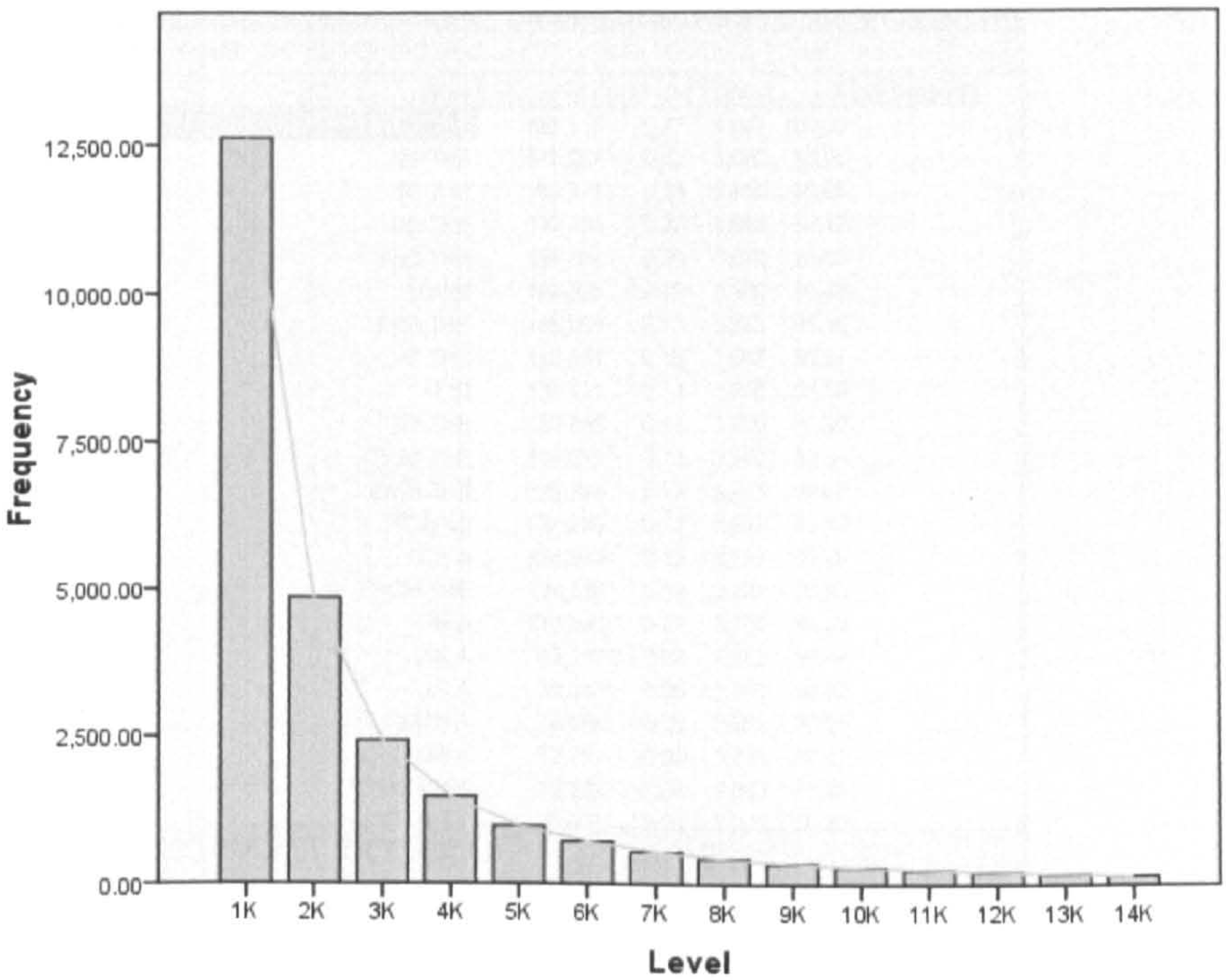
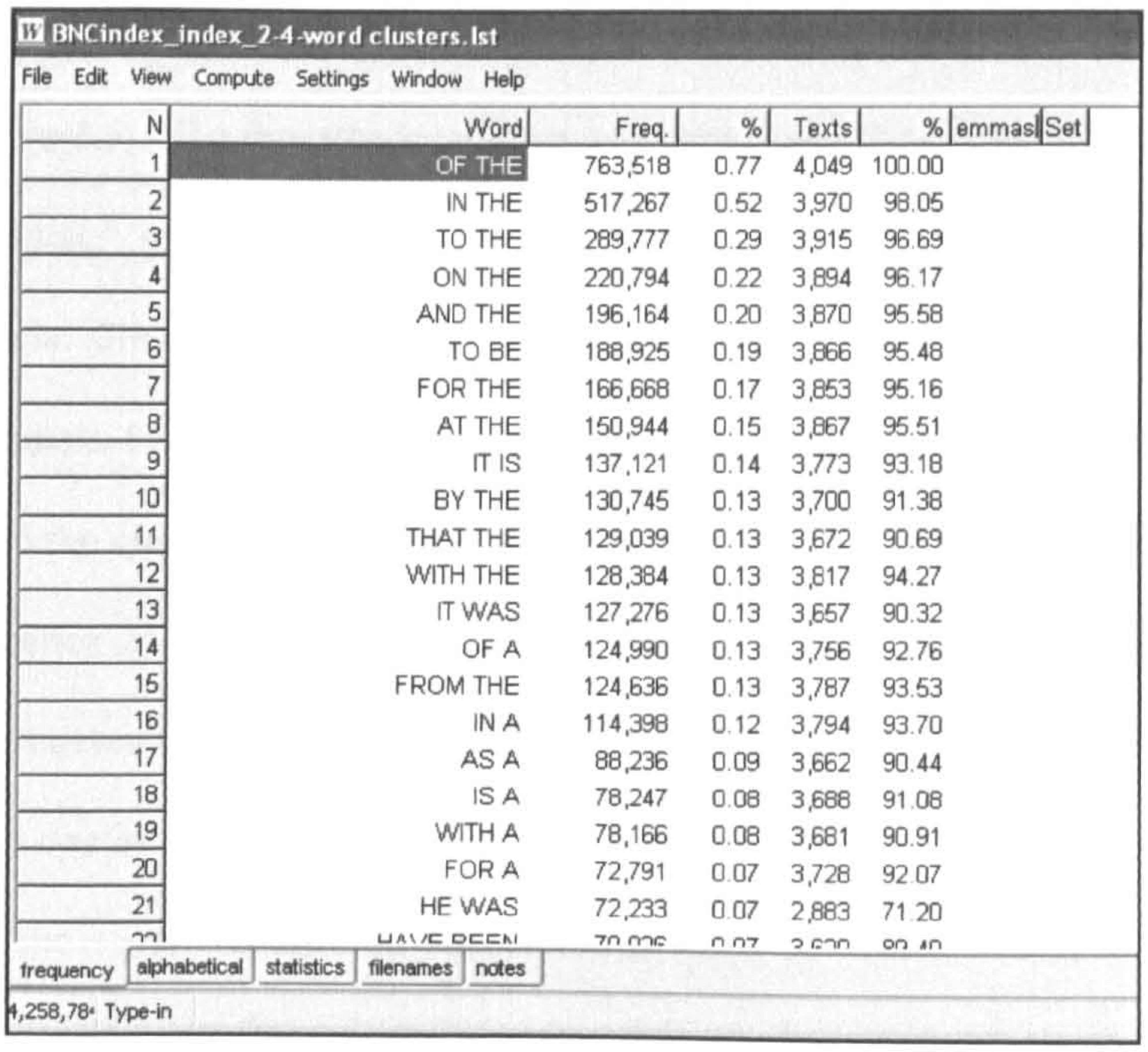


Figure 3.4. Diminishing returns in frequency bands

This new list was to be the one into which the eventual phrasal expressions identified could theoretically be integrated, so this was necessary as a preliminary step. The list was saved as a separate wordlist, as the original index from which it derived needed to be used to proceed to the next step in this initial research phase.

Thus began the actual extraction process by using *WordSmith Tools* to interrogate the indexed list for any and all n-grams between two and four words long repeated in the corpus at least five times. This search rendered a list of over 4.2 million n-grams. It is interesting to contrast this figure with the single-word index list, which reached a limit of just over 750,000 total word forms (a sample is provided in Figure 3.5). (Possible implications of this size discrepancy is elaborated on further in Chapter 6.)



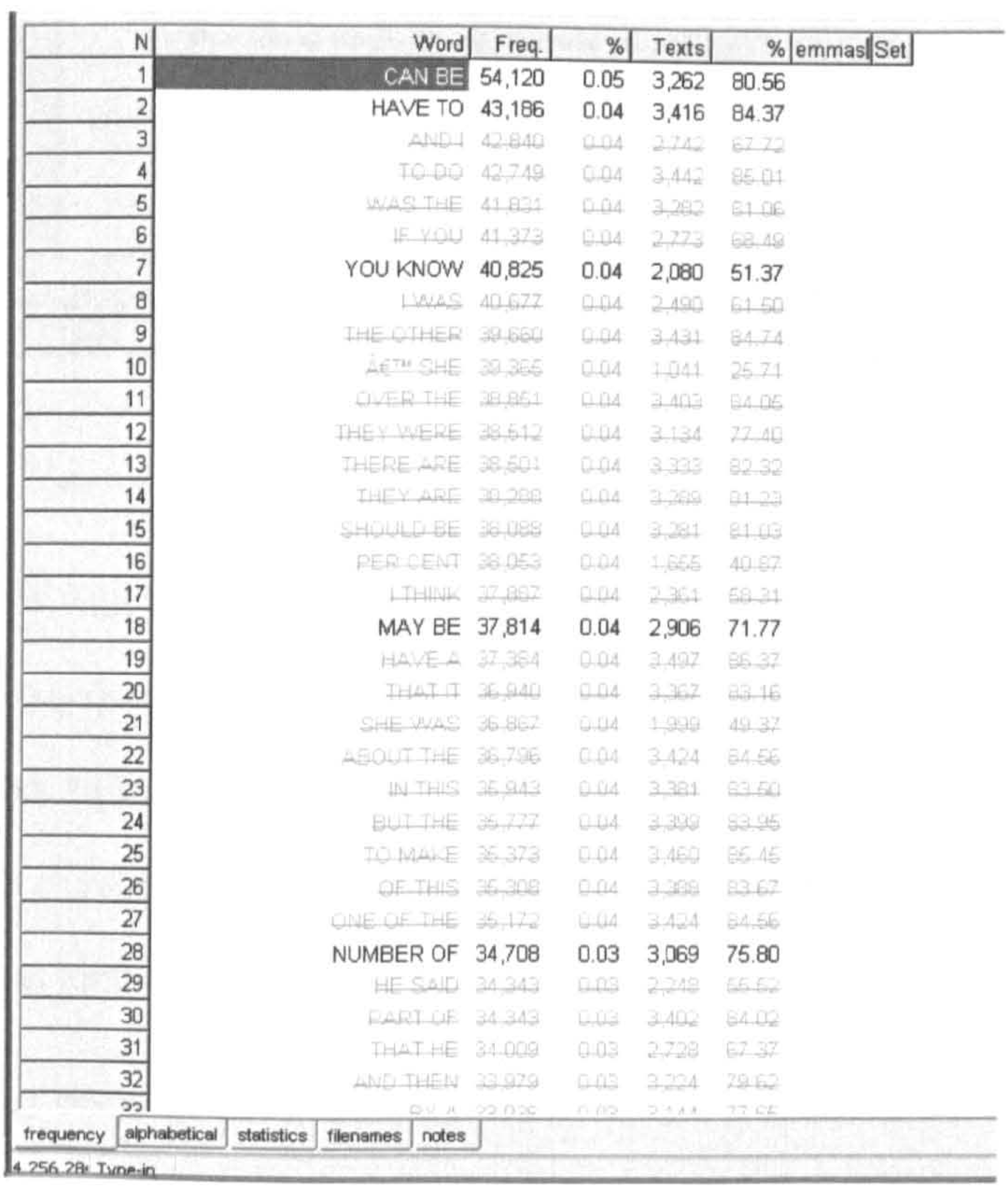
N	Word	Freq.	%	Texts	%	emmas	Set
1	OF THE	763,518	0.77	4,049	100.00		
2	IN THE	517,267	0.52	3,970	98.05		
3	TO THE	289,777	0.29	3,915	96.69		
4	ON THE	220,794	0.22	3,894	96.17		
5	AND THE	196,164	0.20	3,870	95.58		
6	TO BE	188,925	0.19	3,866	95.48		
7	FOR THE	166,668	0.17	3,853	95.16		
8	AT THE	150,944	0.15	3,867	95.51		
9	IT IS	137,121	0.14	3,773	93.18		
10	BY THE	130,745	0.13	3,700	91.38		
11	THAT THE	129,039	0.13	3,672	90.69		
12	WITH THE	128,384	0.13	3,817	94.27		
13	IT WAS	127,276	0.13	3,657	90.32		
14	OF A	124,990	0.13	3,756	92.76		
15	FROM THE	124,636	0.13	3,787	93.53		
16	IN A	114,398	0.12	3,794	93.70		
17	AS A	88,236	0.09	3,662	90.44		
18	IS A	78,247	0.08	3,688	91.08		
19	WITH A	78,166	0.08	3,681	90.91		
20	FOR A	72,791	0.07	3,728	92.07		
21	HE WAS	72,233	0.07	2,883	71.20		

Figure 3.5. A sample of unedited 2-4 grams list derived from BNC

As can be seen in Figure 3.5, however, this rather sizable amount of phraseological data was not very useful without a manual selection of lexical items deemed appropriate for inclusion in the PHRASE List.

Since the analysis of these n-gram clusters involved identifying any recurring string pattern for every single word form in the entire 100 million word corpus, the automated portion of the analysis took 93 straight hours of non-stop computer running time, or just under four full days. The BNC index of individual words when lemmatized and organized into word families indicated that any lexical item that occurred more than 787 times was frequent enough for the 5,000 word-family cut-off. Therefore, all n-grams occurring at least 787 times were considered for inclusion in the PHRASE List. This lowered the n-gram candidate list to approximately 14,500 items.

The time-consuming qualitative stage of analysis then began, involving a line-by-line data deletion phase (Figure 3.6). The researcher meticulously went down the n-gram list item-by-item looking for "plausibly formulaic" multiword items (Wray, 2009: 41), guided by the established selection (and exclusion) criteria. Great care was taken to not overlook potential expressions which at first may not appear formulaic. Even when an n-gram did not apparently seem valuable, there were a number of occasions when the qualitative inspection of the cluster in the actual corpus rendered very interesting results. The sequence *at that,* for instance, may on the surface appear incoherent, but when more carefully investigated reveals interesting idiomatic patterning as in the sentence *CEOs took a pay cut in 2009, and a big one at that.* Corpus-informed dictionaries were also regularly consulted as external confirmation that the n-gram constituted a lexical item.



N	Word	Freq.	%	Texts	%	emmas	Set
1	CAN BE	54,120	0.05	3,262	80.56		
2	HAVE TO	43,186	0.04	3,416	84.37		
3	~~AND I~~	~~42,840~~	~~0.04~~	~~2,742~~	~~67.72~~		
4	~~TO DO~~	~~42,749~~	~~0.04~~	~~3,442~~	~~85.01~~		
5	~~WAS THE~~	~~41,831~~	~~0.04~~	~~3,282~~	~~81.06~~		
6	~~IF YOU~~	~~41,373~~	~~0.04~~	~~2,773~~	~~68.49~~		
7	YOU KNOW	40,825	0.04	2,080	51.37		
8	~~I WAS~~	~~40,677~~	~~0.04~~	~~2,490~~	~~61.50~~		
9	~~THE OTHER~~	~~39,660~~	~~0.04~~	~~3,431~~	~~84.74~~		
10	~~â€™ SHE~~	~~39,365~~	~~0.04~~	~~1,041~~	~~25.71~~		
11	~~OVER THE~~	~~38,851~~	~~0.04~~	~~3,403~~	~~84.05~~		
12	~~THEY WERE~~	~~38,512~~	~~0.04~~	~~3,134~~	~~77.40~~		
13	~~THERE ARE~~	~~38,501~~	~~0.04~~	~~3,333~~	~~82.32~~		
14	~~THEY ARE~~	~~38,288~~	~~0.04~~	~~3,289~~	~~81.23~~		
15	~~SHOULD BE~~	~~38,088~~	~~0.04~~	~~3,281~~	~~81.03~~		
16	~~PER CENT~~	~~38,053~~	~~0.04~~	~~1,655~~	~~40.87~~		
17	~~I THINK~~	~~37,887~~	~~0.04~~	~~2,361~~	~~58.31~~		
18	MAY BE	37,814	0.04	2,906	71.77		
19	~~HAVE A~~	~~37,364~~	~~0.04~~	~~3,497~~	~~86.37~~		
20	~~THAT IT~~	~~36,940~~	~~0.04~~	~~3,367~~	~~83.16~~		
21	~~SHE WAS~~	~~36,867~~	~~0.04~~	~~1,999~~	~~49.37~~		
22	~~ABOUT THE~~	~~36,796~~	~~0.04~~	~~3,424~~	~~84.56~~		
23	~~IN THIS~~	~~35,943~~	~~0.04~~	~~3,381~~	~~83.50~~		
24	~~BUT THE~~	~~35,777~~	~~0.04~~	~~3,399~~	~~83.95~~		
25	~~TO MAKE~~	~~35,373~~	~~0.04~~	~~3,460~~	~~85.45~~		
26	~~OF THIS~~	~~35,308~~	~~0.04~~	~~3,388~~	~~83.67~~		
27	~~ONE OF THE~~	~~35,172~~	~~0.04~~	~~3,424~~	~~84.56~~		
28	NUMBER OF	34,708	0.03	3,069	75.80		
29	~~HE SAID~~	~~34,343~~	~~0.03~~	~~2,249~~	~~55.52~~		
30	~~PART OF~~	~~34,343~~	~~0.03~~	~~3,402~~	~~84.02~~		
31	~~THAT HE~~	~~34,009~~	~~0.03~~	~~2,728~~	~~67.37~~		
32	~~AND THEN~~	~~33,979~~	~~0.03~~	~~3,224~~	~~79.62~~		

Figure 3.6. Example of initial data deletion phase (faded n-grams are deleted ones)

Another example was when looking at the bigram *is to*. At first, it appeared as though the item was simply another meaningless string, among thousands of others in the n-gram list. Just in case, however, the item was checked against a concordance and it was found that there are many instances in which the sequence variously denotes 'will' and 'meant to' (Figure 3.7).

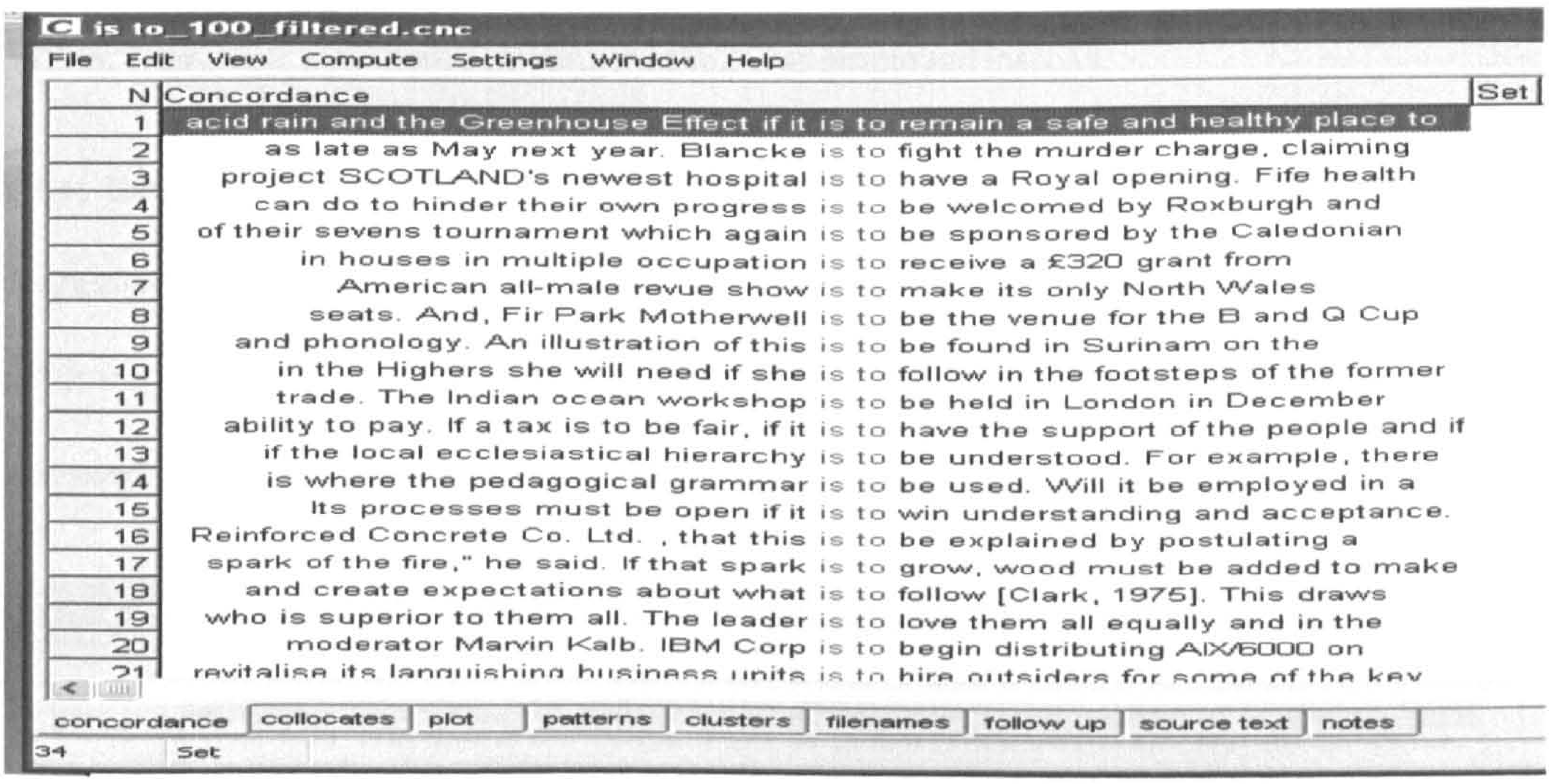


Figure 3.7. Concordance of n-gram *is to*

The item with that particular meaning only occurs 34% of the time (as verified by more than one random sample), but since *is to* as a general sequence appears 29,234 times in the corpus, the ultimate frequency figure for that item was a sizeable 9,939. Since it is also a verb phrase, it was also combined with its plural form (*are to*), independently randomly sampled, and then a total adjusted frequency arrived at for that phrasal expression (15,232).

A further challenge was phraseological polysemy. It was quickly discovered that the number of phrasal expressions with unique form-meaning mappings was relatively limited. While there were many expressions of the variety that would receive a 'ditto-tag' under CLAWS tagging (e.g. *in spite of*, *rather* than, as if) because of their discrete and fixed forms, the vast majority required further investigation in order to determine their true frequency in the corpus. An example is the phrasal expression *at first*. At a glance, it may seem clear that *at first* is an adverbial ('initially'), but with each potential phrasal expression identified an additional concordance of that item was run, and then it would become clear that *at first* also has non-phrasal expression manifestations, as in *love at first sight*. However, since an item like *at first* has a frequency of over 5,000 in the corpus, line-by-line searching was not a viable option. Therefore, a random sampling method was employed instead.

What the random sampling entailed was simply generating a concordance of the potential phrasal expression in question using the entire BNC corpus (a process that takes an average of about 5 minutes per item). Once generated, the concordance was saved and then a special command – 'delete to N' – was used to reduce the concordance lines to a random sample of just 100 (Figures 3.8 and 3.9).

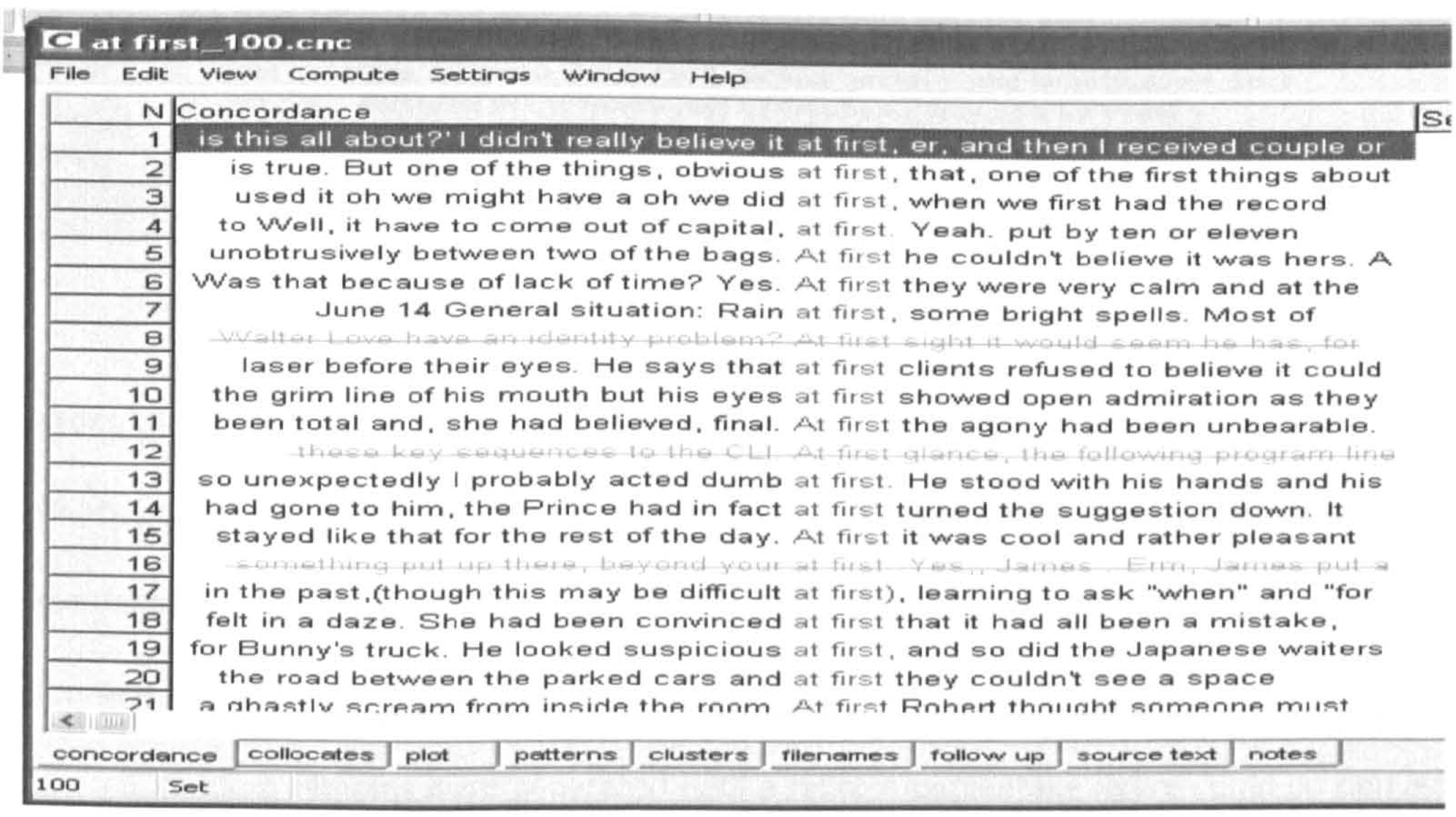


Figure 3.8. An example of a random sample for the purposes of data reduction

Once the concordance was reduced to 100, each line was scrutinized and deleted if necessary (Figure 3.8), until the percentage of lines reflecting the desired use of the multiword item was arrived at (Figure 3.9). As seen in the bottom left-hand corner of the window in Figure 3.9, out of 100 randomly-selected concordance lines, 84 exemplars of *at first* in its phrasal adverbial sense remain – or 84% of the original total.

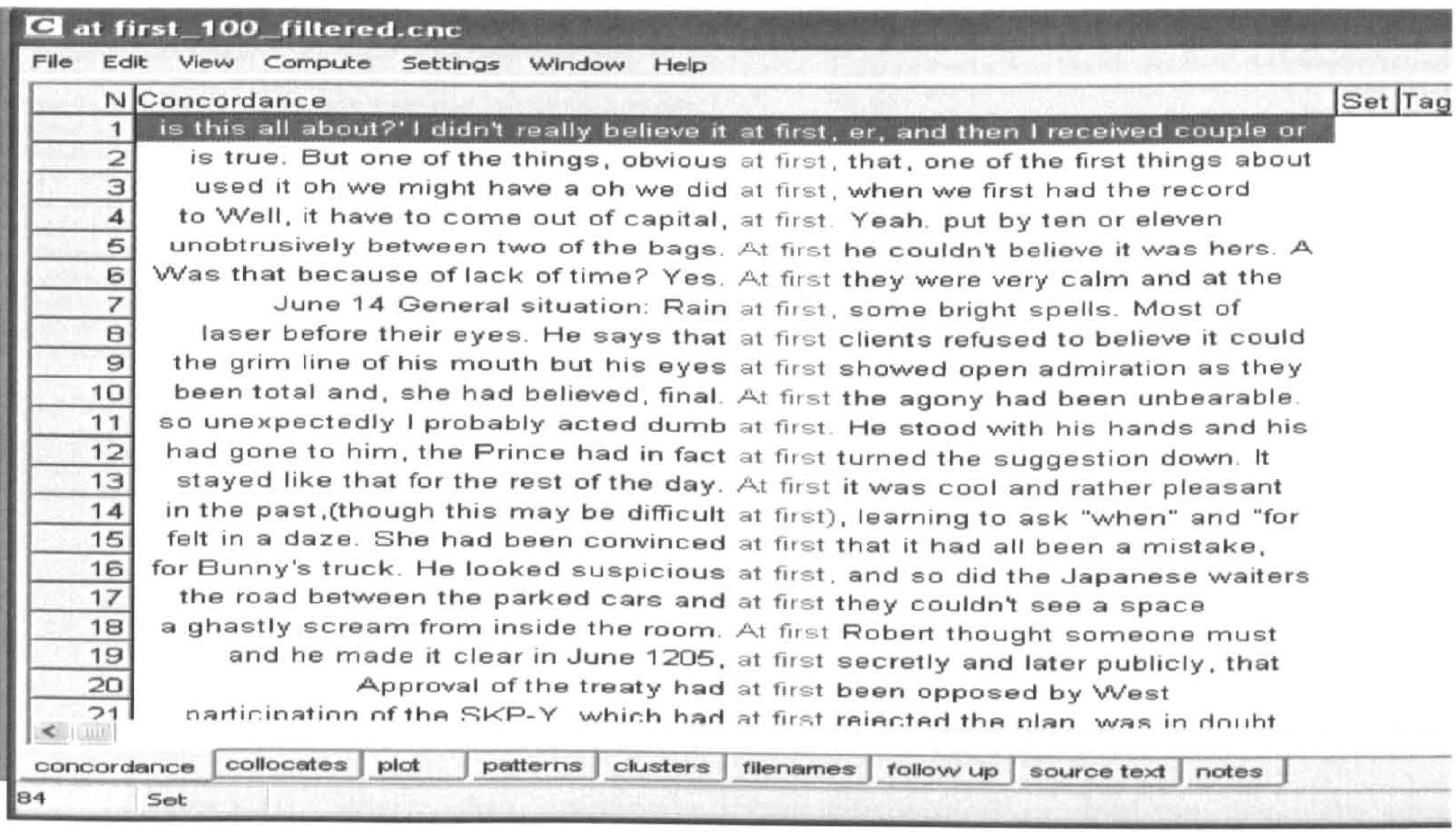


Figure 3.9. Final data deletion example

In order to validate this percentage, a second random sample was generated to check consistency. This method produced consistent results, and in cases of minor discrepancies the lower of the two percentages was used (e.g. the two random concordances for *at first* yielded 84% and 85%, so the 84% figure was used – see Figure 3.10). In the rare cases in which the figures did not match so closely, additional random samples were generated until a reliable percentage figure could be derived. Finally, the frequency figure for each multiword item was calculated by multiplying the total frequency figure by the percentage figure as explained above. For *at first,* this calculation was 5,090 (raw frequency) × .84

(% of desired use) = 4,275 (adjusted final frequency). The lines in the actual *WordSmith* word list were then edited to reflect the adjustment.

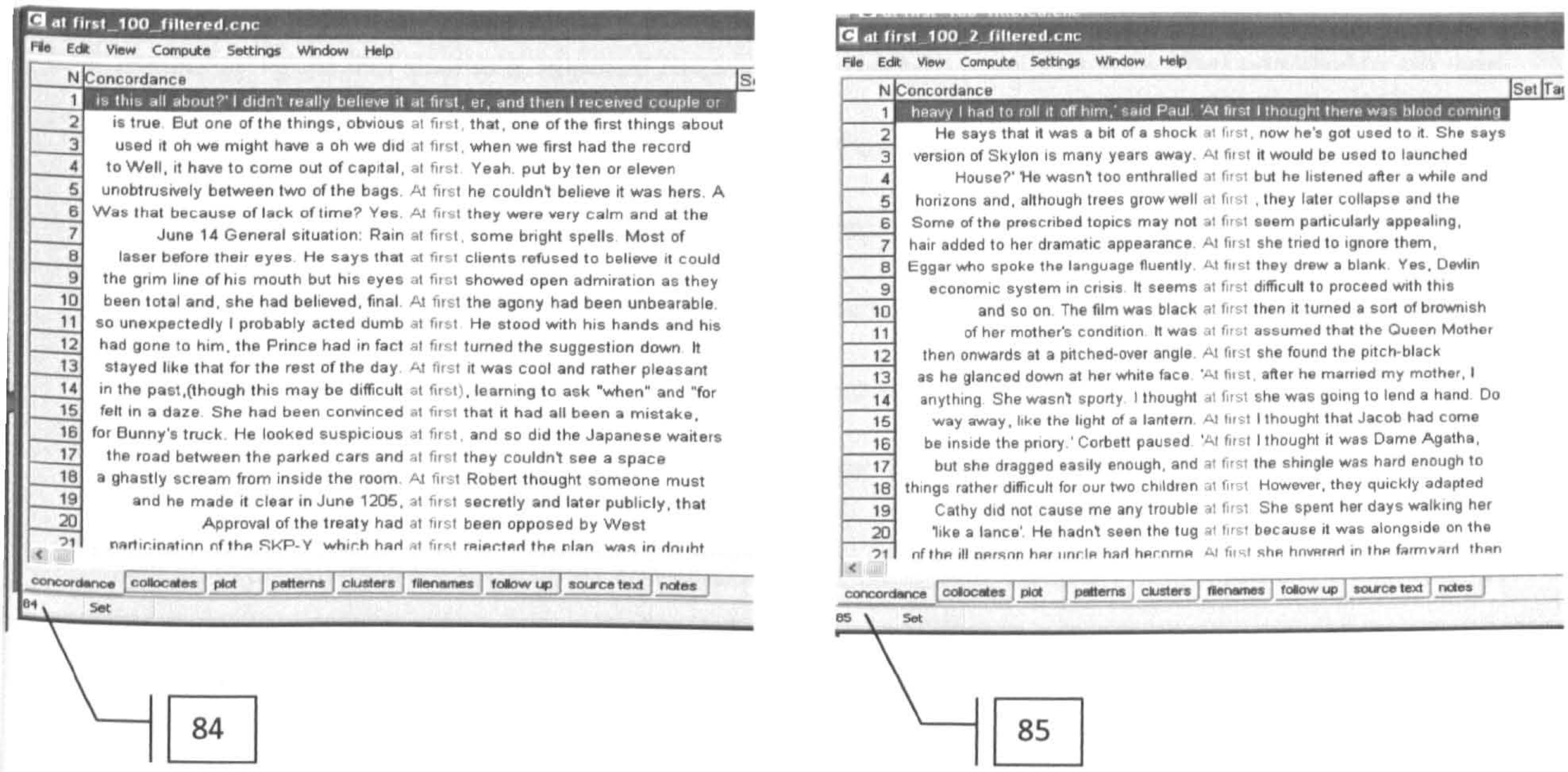


Figure 3.10. Side-by-side comparison of two independent random samples of the same phrasal expression

Also, frequency figures sometimes increased from their original levels. Since the current BNC-derived wordlists are lemmatized and organized into word families, the same needed to occur in the multiword item list. The expression *take place*, for example, in its uninflected form had a frequency count of just 3,248. However, the form can also be lemmatized:

take place → *takes place, taking place, taken place, took place*

In the case of *take place*, after conflating all of the inflected forms, the count increased from 3,248 to 10,556.

On other occasions, a subtractive method could be employed in order to arrive at a more accurate frequency figure. For example, *opposed to* essentially has two manifestations: *(be) opposed to sth*, and *as opposed to*. The n-gram list is not much help on its own since the program was asked to identify all recurring 2-to-4 word strings, and therefore *opposed to* is subsumed in *as opposed to*. In order to focus on just *opposed to*, it was possible to simply subtract the number of occurrences of the string *as opposed to* (1,615) from the number of times the bigram *opposed to* appears in the corpus (2,674), which rendered a difference of 1,059. In other words, the true frequency of just *opposed to* is 1,059.

Finally, expressions were sometimes encountered that contained variable components. For example, in the BNC, the first exemplar of *shake one's head* is actually 'shook his head' (1,698 occurrences). When a phrase with a variable component such as this one was identified (in this case, mainly the pronoun), a careful follow-up search was conducted in order to indentify all variable forms of that expression and arrive at a more accurate frequency count of it. Therefore, after considering *shook his head* (1,698), *shook her head* (1,241), *shook my head* (114), *shake my head* (30), *shaking my head* (17) and so on, the final frequency tally was 3,250.

Irrespective of the method ultimately employed, it was absolutely essential to take time to carefully examine each and every potential item to be included in the PHRASE List in order to both ascertain whether it met the selection criteria and to establish its true frequency.

At the end of the entire process of careful selection and deletion of n-grams, when it was felt that all items meeting the frequency cut-off had finally been identified, the last step would be to 'clean up' the data by excluding all the deleted n-grams form the analyzed raw data list. This can be done fairly easily in *Wordsmith Tools*, which has a command ('zap') that allows the user to eliminate all the lines that have been deleted (but which still appear in the n-gram list as sequences with lines running through them). Before performing this action, however, the file as it stood was saved with all deleted lines still visible and re-introducible if at a later date the original work that had been carried out required revisiting. Once saved, the data was cleaned ('zapped') and exported to an *Excel* file, and saved. This list, in turn, was imported into a *Word* document, which became the first version of the list (see sample in Appendix 2).

3.3.4 Validation of the criteria

Although a criticism can be made of phrase list development methods that apply statistical analyses to the exclusion of qualitative judgment, which was one of the points seen as a possible drawback of the AFL, Simpson-Vlach and Ellis (2010) point out an important rationale for approaching their methodology that way, as it avoids any hints of subjectivity. Clearly, while the criteria in the preceding section were based on theory and informed planning, if only one researcher applies the criteria, a question could raised regarding the reproducibility of the results. Therefore, to help validate the criteria and show that the resulting list is replicable, a rating exercise was carried out.

As explained in greater detail in the previous sections, the initial data extraction from the corpus used for the present study produced a phrasal expressions candidate pool of approximately 14,500 items to which to apply the criteria – naturally too large a sample for a rating exercise. Therefore, a selection of 550 of those items was instead extracted, and the first 50 used as a preliminary training sample (see Figure 3.11 below).

TICK ALL THE N-GRAMS THAT MATCH THE CRITERIA. IF YOU ARE UNSURE, YOU CAN MAKE NOTES ABOUT THE UNCERTAINTY IN THE RIGHT HAND COLUMN. (THE FIRST 50 ARE FOR TRAINING PURPOSES.)

1.	AND HE	25,723	
2.	TO TAKE	25,684	
3.	FROM A	25,672	
4.	DO YOU	25,441	
5.	AT A	25,372	
6.	THE WAY	25,203	
7.	BUT IT	25,143	
8.	AT LEAST	25,034	
9.	AND TO	24,198	
10.	NOT BE	24,050	
11.	IT WOULD	24,013	
12.	THE LAST	23,985	
13.	SUCH A	23,894	
14.	BETWEEN THE	23,887	
15.	MUST BE	23,874	

Figure 3.11. Sample of inter-rating training exercise

The rater was therefore asked to learn the criteria and try to apply them in the same way the principal researcher of this study did, with the only real difference being that the rater conducted the analysis on a paper hard copy, while the main research analysis was done on a computer. However, as the research for the PHRASE List was conducted while having access to the original corpus data, which often was necessary for nearly every item candidate (Section 3.3.3), it was decided that the rater should also have access to the corpus, using the same software used by the researcher. (These were supplied to the rater on a laptop.)

The rater chosen for the exercise was an applied linguistics doctoral student at the University of Nottingham. This person was chosen because of her familiarity with terminology and general familiarity with vocabulary issues, thus facilitating the training necessary. The rater was also selected because she is a non-native speaker of English (from Spain), and any resulting divergence from the results of the principle researcher might be revealing. Conversely, if a high rate of agreement could be achieved with a non-native, it would suggest that the applicability of the criteria is not limited to native speakers of English.

The rater completed the training exercise in less than fifteen minutes. Upon checking the items, it was seen that out of the eight n-grams that had been identified by this researcher as meeting the criteria among the 50 in the training portion of the sheet, only one had been missed by the rater (number 28, *a good*). She admitted that she had not consulted the corpus for that item, and it was pointed out that upon consulting the actual source data in the corpus, how a number of examples are retrieved that do match the criteria (e.g. *I waited for a good hour or so*). Finally, reporting that she had no further questions, the rater was asked to proceed to the main rating portion consisting of 500 n-grams (Appendix 3). The researcher left the room at this point to avoid any undue influence his presence might exert.

The rater completed the exercise in less than two hours, and was then debriefed about any questions or comments she had. In general, what the rater reported was that she found the criteria easy and straightforward to apply, and that they helped her to eliminate many n-grams right away. Then the rater was able to narrow the options down to a pool of around 50 candidates, and then narrowed these down to further. Of the 40 n-grams predetermined to be valid candidates that meet the criteria on the rating sheet in Appendix 3, the rater only missed two (numbers 165 and 426, *those who* and *came to*, respectively) , and chose only two that did not agree with the predetermined list (numbers 135 and 493, *there is* a and *even if*, respectively). These results were very encouraging, as a discrepancy of only four items total out of 500 (the 50 first n-grams were not counted as they were for training purposes only), meant that both raters were 99.2% in agreement.

It is interesting to note that the rater also reported that she found the exercise interesting because if she had been asked to simply find 'phrases' there could have been a number of other n-grams that could have been chosen, but the criteria helped her be more selective and even notice a few that she reported she might not have noticed without the criteria (e.g. number 344, *each other*).

The rater did report that she often was not sure whether to select an item because she recognized it as a stem or fragment of a larger phrase (e.g. *on the other*), but this was to be expected as the principal researcher had access to the entire n-gram list (and could therefore check if the whole phrase appeared elsewhere), while the rater for this exercise did not.

In all, however, the rating exercise was able to show with some confidence that the established six criteria could be applied with a high degree of consistency by two independent raters, and therefore the results of an analysis involving the use of the criteria should be replicable.

3.4 Results and discussion

Following the analysis presented in Section 3.3, the first version of the PHRASE List was produced. However, the list eventually underwent a number of changes as a result of further analyses and consideration, and the details of the alterations that were made and the rationale for them will be discussed in this section.

3.4.1. The PHRASE List: the first draft

A portion of the first version of the PHRASE List is shown in Appendix 2, but this list would undergo a number of changes. (A sample is shown in Figure 3.12.)

IN CONTRAST	2229
THIS STAGE	2223
ALL BUT	2214
ABOVE ALL	2212
RID OF	2212
IN ANY CASE	2159
THANKS TO	2159
GO AWAY	2150
ONCE MORE	2146
OH WELL	2129

Figure 3.12. A sample from the first draft of the PHRASE List

The total number of phrasal expressions in the original list was 526, but was eventually reduced to 505 – the number at which it stands at present. There were many different reasons for exclusion of some of the phrasal expressions. For example, in many cases phrases were basically repetitions of the same expression, but had mistakenly been listed separately. Some examples that were excluded can be seen in Table 3.5. If the phrase was a redundancy, the frequency of it was re-checked against the original data and adjusted if necessary. Still in other cases, there was an issue of transparency (i.e. the item would be too easily decoded from reading the individual words). In both the cases of data redundancies and semantic transparency, the items were removed when the list was proofread over the days and weeks following its initial instantiation, bringing its total then to 507 phrasal expressions. The phrases that were removed because they were too limited to a particular genre within the BNC (i.e. of limited 'dispersion') will be discussed in Section 3.4.3.

Table 3.5 Examples of phrasal expressions that were deleted from the original PHRASE List

Excluded phrase	Reason
it took	Redundancy. (The phrase 'it takes' already in list.)
a wide range of	Redundancy. (The phrase 'a range of' already in list.)
if you like	Transparency.
you know	Transparency.
party to	Dispersion.
third party	Dispersion.

3.4.2 Addition of contextual and rank information

When the first draft of the PHRASE List was seen by colleagues, one critical comment that was received was regarding the interpretability of the phrases. It seemed that while certain phrases were readily recognizable as lexical items (e.g. *might as well, in the first place, take for granted*) with clear, discrete form-meaning mappings, some other phrases eluded immediate interpretation (e.g. *or so, all but, yet to*). This was obviously a problem, particularly if the list was ultimately intended to be accessible and usable by a wide variety of end-users, including teachers and learners of English. In order to prevent the perceived usefulness of the list from being undermined by these more ambiguous phrases, the decision was taken to complement the existing list with an additional column that would put the phrases in an example sentence (Figure 3.13).

Integrated List Rank	Phrase	Frequency (per 100 million)	Example
3149	**IN CONTRAST (TO)**	2229	The inside was amazing **in contrast.**
3152	**THIS STAGE**	2223	We can't at **this stage.**
3157	**ALL BUT**	2214	She **all but** sent him chocolates and flowers.
3160	**ABOVE ALL**	2212	It is **above all** what people care most about.
3162	**RID OF**	2212	She was happy to be **rid of** it.
3197	**IN ANY CASE**	2159	It's not due till tomorrow **in any case.**
3199	**THANKS TO**	2159	And it's **thanks to** her research that we know that.
3205	**GO AWAY**	2150	The problem won't just **go away.**
3207	**ONCE MORE**	2146	I call on you **once more** my fellow citizens.
3220	**OH WELL**	2129	It was due yesterday? **Oh well.**

Figure 3.13. A sample from a later revision of the PHRASE List

As shown in Figure 3.13 above, the contextualization of the phrases seems to have enhanced the interpretability of the items in all cases, with the meaning of a phrasal expression like *all but* becoming much clearer. (The need for context for many of these phrases also presented itself as an issue when designing a test that included them, discussed extensively in Chapter 5.) What is more, the example sentences also help illustrate the intended sense of phrases. For example, the phrase *oh well* on its own in a list could look like some kind of incomplete stem of a longer sentence (e.g. *Oh well that's interesting*), however the contextualization helps to show that it is an item in its own right.

The frequency column remained the same, but in Figure 3.13 there is the addition of the 'Integrated List Rank'. What this last figure indicates is where the phrasal expression would rank on a frequency-ranked wordlist derived from the BNC. An extended sample of the integrated list is provided in Appendix 4, but an example can be seen in Figure 3.14. These data were added as it was felt they helped to conceptualize the relative importance of the phrasal expressions, even more than the raw frequency data.

RANK	BEFORE	FREQUENCY
4719.	CULT	1061
4720.	DESCENT	1061
4721.	STOCKING	1061
4722.	BELLY	1060
4723.	NUTRITION	1060
4724.	BRACKET	1059
4725.	SOFA	1059

RANK	AFTER	FREQUENCY
4719.	CULT	1061
4720.	DESCENT	1061
4721.	FOR GOOD ('FOREVER')	1061
4722.	STOCKING	1061
4723.	BELLY	1060
4724.	NUTRITION	1060
4725.	BRACKET	1059
4726.	(BE) OPPOSED TO	1059
4727.	SOFA	1059

Figure 3.14. Example of integrated list of phrasal expressions and single words

However, although the integration of the phrasal expressions as shown in Figure 3.14 does probably help to visualize their importance as lexical items, it should be noted that there is a validity issue in full integration of the PHRASE List into a list of words. It is an issue that lexicographers have faced before when trying to report valid frequency information in various corpus-informed dictionaries, as in the case of the Longman Dictionary of Contemporary English (LDOCE):

> The question "how do you count the instances of *look* if you are also counting all its phrasal verbs" was one we wished to avoid, since each possible answer to it had some undesirable implications. (Kilgarriff, 1997: 145)

What Kilgarriff is referring to is the fact that a simple wordlist will provide a raw frequency count of a word, but that word may actually be part of other expressions. Put another way, if a phrasal expression like *look for* is also accepted as a 'word' in its own right (which Kilgarriff and his colleagues did for LDOCE), then at least part of that frequency count of just the word *look* is wrong, since the frequency count for *look for* should be subtracted from it. However, this subtraction in itself can lead to further challenges:

> (T)his course would be awkward to implement, not only due to the built-in difficulties of counting phrasal verbs (because sometimes a combination like *look for* is not a phrasal verb in the corpus), but also because the (frequency) threshold will keep changing. (Kilgarriff, 1997: 153)

The same issue related to valid frequency counts would occur in a wordlist absorbing the PHRASE List. The frequency cut-off points in the list (e.g. 1K, 2K, 3K, etc.) were derived from the lemmatized BNC wordlist organized into word families, but if those word families are teased out and re-organized into phrases with their own respective frequencies, those frequency thresholds change again, and as Kilgarriff observes, "the whole process would need repeating" (p. 153) each time a phrase gets its own frequency[35]. This issue of the validity of the frequency counts of the commonest words in wordlists will be revisited in Chapter 5.

3.4.3 Addition of genre information

As explained in Section 3.3.1, the PHRASE List was reduced from an initial figure of 526 expressions to 505, and most of the phrases excluded were due to redundancies and issues of transparency. However, there were a couple of items that were removed when the list was used in a validation study of another phrase list, the *English Profile Wordlists* (EPW) phrases, discussed at length in Chapter 4. The EPW researchers were able to carefully evaluate the phrasal expressions in the PHRASE List individually, and though they saw the merits of nearly all of them (see Chapter 4), two phrases caught their attention as seeming to be of questionable general usefulness: 'party to' and 'third party' (Table 3.5). These phrases were investigated further in the corpus, and though the phrases are indeed attested in the corpus thousands of times, it was found that the great majority of the contexts in which they were found involved legal documents of some kind. Around the same time as the aforementioned investigation, a colleague who had the opportunity to see the list had recommended the addition of information regarding whether the each phrase was more common in spoken or written English. Taken together, it appeared obvious that it would be necessary to investigate the relative frequencies of each of the phrases according to genres in the BNC.

There is one problem with such an endeavor, however. The BNC is composed of hundreds of different sub-corpora, and there is no easy way to isolate, say, just general spoken conversation and investigate

[35] The issue is actually even more complicated than Kilgarriff suggests. Not only would there likely never be a consensus on which phrases deserve their own entries (and which therefore would subtract from the frequencies of other headwords) – an issue which actually occurred and is detailed in Chapter 4 – with phrasal expressions that contain two more content words (e.g. *take place, make up your mind*), a frequency deduction would somehow have to occur on at least two headwords.

the frequency of a given phrase in those files. Furthermore, even the individual files in the BNC that are tagged as representing 'spoken' English, for example, are actually not what one would immediately think of with respect to that modality of communication, with many BNC files actually containing data of memorized and/or written language that has been read aloud. This problem of genre mislabeling in the BNC has actually long been recognized by users of that corpus (Lee, 2001). Fortunately, an index entitled 'The BNC World Edition (Bibliographical) Index', which exists in the form of a publicly-available Excel file, has been developed to address this issue (Lee, 2002: 1):

> The BNCW Index spreadsheet was created as one solution to the problem of BNC 'domain' categories being overly broad and too inexplicit, to fix classification errors and steer people away from misleading file titles, and to provide a proper navigational map for people wanting to deal with specific 'genres' (as generally understood by most people).

BNC World SPOKEN	No. of words	%	Big Genre	# of Files
S_brdcast_discussn	757,317	7.3%		53
S_brdcast_documentary	41,540	0.4%	Broadcast	10
S_brdcast_news	261,278	2.5%	10.2%	12
S_classroom	429,970	4.2%		58
S_consult	138,011	1.3%		128
S_conv	4,206,058	40.7%		153
S_courtroom	127,474	1.2%		13
S_demonstratn	31,772	0.3%		6
S_interview	123,816	1.2%	Interviews	13
S_interview_oral_history	815,540	7.9%	9.1%	119
S_lect_commerce	15,105	0.1%		3
S_lect_humanities_arts	50,827	0.5%		4
S_lect_nat_science	22,681	0.2%	Lectures	4
S_lect_polit_law_edu	50,881	0.5%	2.9%	7
S_lect_soc_science	159,880	1.5%		13
S_meeting	1,377,520	13.3%		132
S_parliament	96,239	0.9%		6
S_pub_debate	283,507	2.7%		16
S_sermon	82,287	0.8%		16
S_speech_scripted	200,234	1.9%	Speeches	26
S_speech_unscripted	464,937	4.5%	6.4%	51
S_sportslive	33,320	0.3%		4
S_tutorial	143,199	1.4%		18
S_unclassified	421,554	4.1%		44
TOTAL	10,334,947	*100.00%*		*909*

Figure 3.15. Sample from the British National Corpus World Index (Lee, 2002)

An example of Lee's meticulous work can be seen in Figure 3.15. Figure 3.15, however, is simply a sample of the overview provided in the BNCW Index. The actual *Excel* spreadsheet file is fully searchable and contains a breakdown of the genre and specific contents of every file in the BNC. Nonetheless, Figure 3.15 does offer an idea of the sheer variety of specific genres that have been classified under 'spoken' in the BNC, with files like 'courtroom' and 'sermon' probably being less representative of what is generally thought of as everyday speech than files like 'speech unscripted' and even 'interview'. Initially, it was hoped that four genres could be isolated and investigated for the PHRASE List: spoken general English (i.e. general conversation), written general English, spoken academic English, and written academic English. However, upon closer examination of the BNCW Index, it was decided that there were not enough corpora of the spoken academic genre present in the corpus to allow for a representative and comparable sample to be generated. Therefore, two million words of the other genres were isolated from the BNC, and merged to form three sub-corpora for further analysis. The 505 phrasal expressions in the PHRASE List were then individually checked for their relative frequency. An example of the data that resulted, and the alterations that took place as a result in the PHRASE List, can be seen in Table 3.6.

Table 3.6 Sample of the PHRASE List with numerical genre-sensitive frequency information

Integrated List Rank	**Phrase**	**Frequency** (per 100 million)	Spoken general	Written general	Written academic	**Example**
107	**HAVE TO**	83092	**1479**	**502**	**89**	I exercise because I **have to.**
165	**THERE IS/ARE**	59833	**1133**	**997**	**668**	**There are** some problems.
415	**SUCH AS**	30857	**130**	**591**	**620**	We have questions, **such as** how it happened.
463	**GOING TO (FUTURE)**	28259	**587**	**194**	**12**	I'm **going to** think about it.
483	**OF COURSE**	26966	**511**	**327**	**41**	He said he'd come **of course.**

Table 3.6 shows the first attempt at listing genre frequency data alongside the phrasal expressions. As can be seen, all the other data columns (e.g. 'Integrated List Rank') all remain. These new frequency data pertaining to genre were ultimately found to be more of a hindrance than help, unfortunately. Every person who had the opportunity to look at the new version of the list found the new numbers confusing, and understandably so. First of all, the original data columns containing frequency information are still in the list, so the addition of three new sets of numbers is somewhat daunting (even for the academic researchers who offered feedback on this iteration of the list, so non-expert teachers and students would likely find them even more difficult). Second, the new sets of frequency information actually are un-interpretable in practical terms. As an example, the phrase 'such as' is shown to have 130 attested examples in the spoken general corpora analyzed. However, even when juxtaposed with the subsequent two columns with higher figures, what does that number of 130 mean? To a user picking up the list for the first time, one can imagine it would be very difficult to determine if that figure of 130 means that it is rare, and if it is, how rare it is relative the other genres.

WordSmith Tools does offer a 'Keyness' feature which will automatically produce statistical measures of distinctiveness between two corpora (e.g. log likelihood, chi squared), and while this was naturally considered as an alternative to frequency reporting, it was discarded for a number of reasons. First of all, the keyness feature is useful for exploring which lexical items stand out as identifying a particular genre (e.g. Culpeper, 2009; Durrant, 2009), and not necessarily for understanding relative frequency among phrases from three different genres. Moreover, the keyness measures do not offer much more help to the user of the PHRASE List as, once again, they are simply numbers that together with the frequency information already in the table may only serve to confuse, and in any case still require interpretation.

Therefore, since interpretation would be required irrespective of the measure employed, it was decided that a more practical system would simultaneously facilitate the interpretation while avoiding adding more numerical information that might overwhelm and confuse the user of the list. This led to the development of a new system, which is the one used in the final draft of the PHRASE List (sample provided in Table 3.7).

Table 3.7. Genre-sensitive frequency information represented by system of symbols

***	=	phrase most common in this genre (or as common)
**	=	phrase less common in this genre
*	=	phrase infrequent in this genre
X	=	phrase rare or non-existent in this genre

Integrated List Rank	**Phrase**	**Frequency** (per 100 million)	Spoken general	Written general	Written academic	**Example**
107	**HAVE TO**	83092	***	**	*	I exercise because I **have to**.
165	**THERE IS/ARE**	59833	***	***	**	**There are** some problems.
415	**SUCH AS**	30857	*	***	***	We have questions, **such as** how it happened.
463	**GOING TO (FUTURE)**	28259	***	**	x	I'm **going to** think about it.
483	**OF COURSE**	26966	***	**	*	He said he'd come **of course**.

As seen in Table 3.7, a system of four symbols was devised which correspond to the frequencies of each item, previously represented numerically. Each phrasal expression now has at least three stars ('***') in at least one genre, representing the genre in which that phrase occurs the most. If the frequency of that same phrase in another genre was found to be the same (within 30 percent), the same amount of stars was assigned. However, if the token frequency was between 30 to 70 percent less than the highest value at three stars, it was assigned two stars, and if representing only between 5 to 29 percent of the highest value, just one star. Anything less frequent than 5 percent of the highest value was assigned an 'X', designating the phrase as rare or non-existent in that genre.

A further review of this new system retrieved much more positive feedback among colleagues, who felt that the symbols, rather than frequency information, were a much better solution for a general audience of applied linguistics practitioners. At a glance, one can fairly easily tell if the phrase is common or not in each genre, and even achieve an understanding of the relative frequency among them. This new system, it was decided, better served the applied purposes of the PHRASE List. Hence, this final iteration of the format of the list – and the phrasal expressions contained in it – was taken as the final draft, and the full version of it can be found in Appendix 1.

3.4.4. The PHRASE List: the final draft

As one final measure to enhance the usefulness and user-friendliness of the list in its final form, a kind of 'PHRASE List User's Manual' was drafted to be used in conjunction with the list. Seen in full in Appendix 5, it contains some background on the development of the list, its rationale, and its intended applications. Perhaps most importantly, it also contains information about what inherent limitations there are to the list, including the variety of English (almost entirely British) and the predominance of written versus spoken English. It is hoped that together with the PHRASE List in its current form, the list will be found to be useful and practical by a wide audience of end-users.

As mentioned at the end of the preceding section, a full and final PHRASE List can be found in Appendix 1. This of course represents the last version of the PHRASE List as of this writing, but by no means does it suggest that no further changes should be made. (See 'Conclusion' below). However, the list does allow the summary of the results as they currently stand.

The PHRASE List consists of a total of 505 multiword items. This is actually quite a substantial number, and indeed, if integrated into and calculated as part of the 5,000 most frequent word families, the 505 multiword items would constitute over 10 percent of the total items. This figure also counters the assertion sometimes made that the number of commonly-occurring opaque multiword expressions in English is low (e.g. Grant & Nation, 2006; Moon, 1998; O'Keeffe, McCarthy & Carter, 2007), and thus "should not be a major learning goal of a language learning programme" (Grant & Nation, 2006: 11). According to the analysis conducted for the present study, there is a sharp increase in the number of phrasal expressions identified after around 12,000 occurrences, or the 1,000 (1K) word-family level, surging from 32 items in the first band to 85 items in the next (2K) level. This trend of increase appears to continue to the 4k level, and then levels off after 5K (Figure 3.16). This may be a reflection of a tendency for the most frequently recurring word combinations to sometimes become 'grammaticized' (described in Section 2.3), often losing compositionality (Bybee, 2003). Since the 5000 word frequency level seems to represent the upper limits of high frequency lexis, it may be that the number of such non-compositional phrases – precisely the kind chosen for the PHRASE List – diminishes to some degree after that frequency threshold. (Though this still needs to be researched – see comments on future directions in Chapter 6.)

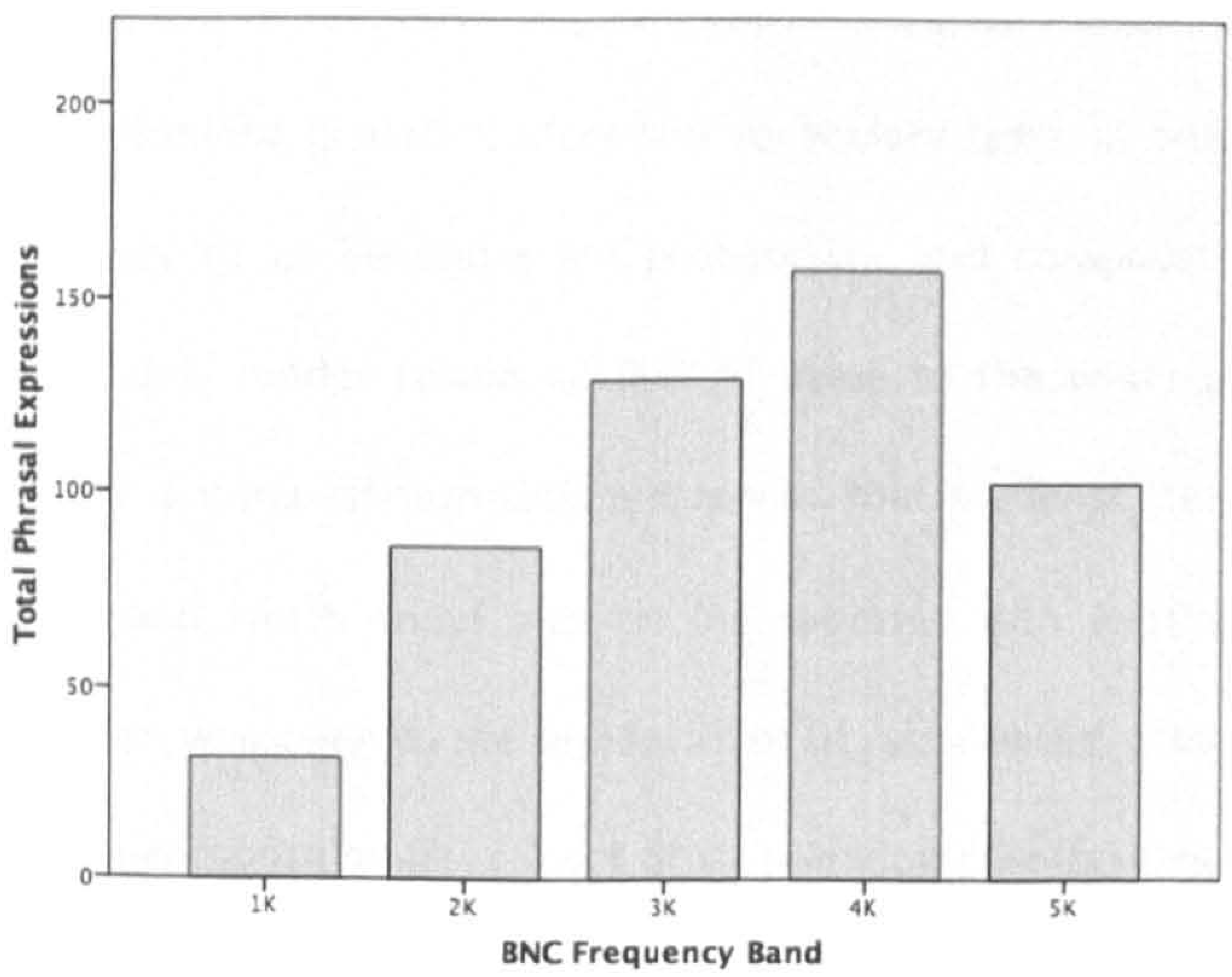


Figure 3.16. Phrasal expressions across frequency bands

Moreover, lending strength to the assertion long held by a number of authors working in the area of formulaic language that the most common words in existing wordlists are merely the tips of phraseological icebergs (e.g. Sinclair, 1987), as illustrated in Figure 1.1 of the first chapter, an analysis of the expressions in the PHRASE List shows that the 505 expressions are almost entirely comprised of the top 2000 words in English, with the vast majority in the top 1000. (95% in the first 1000 and 2.88% in the second.) Since the phrases were chosen not only for their frequency but also because of their relative lack of semantic transparency, it is not unreasonable to guess that L2 learners processing those expressions might therefore actually believe they understand them (if they identify them) simply because the individual words are so well known (Martinez & Murphy, 2011).

3.5. Conclusion

This chapter described an attempt to compile a list of formulaic sequences that could complement existing lists of the most frequent words in English that are currently used to inform areas of language instruction like graded readers and vocabulary tests. In order to do so, it was determined that it was necessary to go beyond mere probabilistic and computer-driven automated retrieval methods that would only render results of limited value to the wider applied linguistics community. In order to identify a range of formulaic sequences that students, teachers and researchers could recognize as useful and which could also be incorporated into existing wordlists and vocabulary tests, careful attention was given to the semantics of the sequences, prioritizing those with relative semantic opacity (or non-compositionality). A set of six selection criteria were also established, and these criteria passed an inter-rater test of consistency. In the final analysis, the product of the research – the PHRASE List – does appear to largely fulfill the intended outcome of the present study. The mixed-methods procedure employed in the identification and selection of the phrasal expression candidates was generally effective in facilitating the inclusion of the specific type of formulaic sequence targeted for the list, including phrases that might not have been picked up by machine or human scanning. Indeed, there will inevitably be disagreements regarding certain individual items that were included in (or even excluded from) the list when it reaches the public domain. Many expressions, such as *on the other hand* and *take for granted* are readily identifiable as formulaic, while others, such as *no one* (*'I can think of **no one** better'*) and *a good* (*'It takes **a good** three days'*) may not fit the stereotype of 'formulaic expression' in an obvious way. Users of the PHRASE List are advised to carefully consider such expressions in the light of the established criteria, and how what may at first seem easily understandable may in fact not be – even in context (Bensoussan & Laufer, 1984; Haynes, 1993; Kaivanpanah & Alavi, 2008) – especially for lower-proficiency learners. Nonetheless, critical evaluation of the list is welcome.

Finally, it should be mentioned that while the list can now be considered fairly comprehensive in terms of the type of phrase that was chosen for the PHRASE List and frequency levels identified as relevant, there were many more phrasal expressions that were identified and noted during the selection process, ones which were not frequent enough to be listed in the PHRASE List, but which were well within the frequency of the top 10,000 words in English, for example. Therefore, it would seem that just as the BNC frequency lists currently span all the way to the top 14,000 word families, it would be interesting to eventually revisit the phrases noted during the research described in this chapter and endeavor to create lists of lower frequency phrasal expressions as well. Such an investigation would take longer than the one that was carried out for the present study (due to frequency effects related to Zipf's Law), but would undoubtedly be worthwhile to reach still a better understanding of the nature and size of the phraseology that permeates the English language.

There are of course a number of future applications which are envisaged for the PHRASE List, and some of these will be discussed in Chapter 6. First, however, two case studies involving practical applications of the PHRASE List which have already been carried out will be reported on and explored in depth in the two chapters which follow this one.

Chapter 4 – Case Study 1: Validation of the *English Profile Wordlist*

One way of validating a wordlist is to compare it to others extracted by different means but which purports to comprise the same or similar construct (i.e. concurrent validity). After all, if the corpora from which the lists are derived are in any way comparable, then the output – if indeed of the same or similar construct – should also be at least somewhat comparable, even when compilation methodologies differ (e.g. Hoftland & Johansson, 1982). This principle should apply to all lexis, single or multiword. The current chapter describes a validation study that was conducted between December 2009 and May 2010 which involved the comparison of the PHRASE List with a list of phrases produced through research at the University of Cambridge ESOL Examinations and Cambridge University Press. Section 4.1 provides some background to the English Profile project in general, and the English Vocabulary Profile in particular. Section 4.2 gives an overview into how the English Profile wordlists were developed. Section 4.3 reports on the methodology and results of a comparison between the PHRASE List and the English Profile lists. Section 4.4 explores some issues that arose during the validation exercise in greater depth, and 4.5 concludes the present chapter.

4.1. The *English Profile Programme* and the *English Vocabulary Profile*

The English Profile Programme was set up in 2006 as a collaborative project to develop reference level descriptors linked to the four proficiency levels (Table 4.1) of the Common European Framework of Reference (CEFR), designed initially as a non-language-specific proficiency framework, predicated mostly on 'can-do' statements (e.g. 'Student can engage in simple informal conversation'), but which until very recently lacked guidelines containing linguistic specificity (Milanovic, 2009). Therefore, the primary remit of the English Profile Programme is to use empirical evidence – corpus data in particular – to add linguistic specificity (e.g. grammar, lexis, functional exponents), also known as 'criterial features' (Saville & Hawkey, 2010), to the CEFR levels.

Table 4.1 Levels of the Common European Framework of Reference (CEFR)

Proficient user	**C2 Mastery**
	C1 Effective Operational Proficiency
Independent user	B2 Vantage
	B1 Threshold
Basic user	A2 Waystage
	A1 Breakthrough

One key partner involved in the English Profile Programme, the University of Cambridge ESOL Examinations, commissioned research in 2007, led by Annette Capel (formerly of the COBUILD[36] project), to create a wordlist aligned with the CEFR levels. Initially called the English Profile Wordlist(s) (EPW)[37], the purpose of the list was to verify what lexical knowledge learners demonstrate at each of the CEFR bands. Conceivably, such information could then be used by such parties as examination writers and syllabus designers wishing to more accurately link tests and textbooks to claims of CEFR alignment. (See, however, word of caution on this matter at the conclusion of the present chapter.)

[36] The Collins-Birmigham University International Language Database, a corpus-based research project led by John Sinclair in the 1980s, originally intended to help inform lexicographic research, but which eventually led to important insights relevant to the nature of lexis, such as the notion of the 'Idiom Principle' (Sinclair, 1991).
[37] At the time of this writing, the project name has changed to the 'English Vocabulary Profile'.

4.2. Development of the English Profile Wordlists (EPW)

As one of the English Profile partners who commissioned the wordlist research was Cambridge University Press, lexicographic research that had already gone into the production of the *Cambridge Advanced Learner's Dictionary* (CALD) was used as a starting point. CALD had been compiled drawing on the Cambridge International Corpus (a corpus used almost exclusively by Cambridge University Press for in-house projects, consisting of one billion words of written and spoken English), and of particular use to the EPW project, contained sensitive frequency information for the most common 6000 words in English. What the lexicographers on CALD had done was to manually count concordance lines of each of those words to determine the relative frequency of each sense (Capel, 2010). For example, the word 'book' may occur 20,000 times in the Cambridge International Corpus, but even if one teases out grammatical information from the raw frequency (e.g. 'book' as a noun versus 'book' as a verb), in order to arrive at a truer frequency that also accounts for semantics (i.e. 'sense'), one needs to look at the words in the actual data – otherwise, automated retrieval methods may not reliably distinguish an example like 'the room was booked' from 'the player was booked'. Such data proved a valuable point of departure for the EPW project since it was vital for the objectives of the project to make sure that the wordlists were sense-sensitive.

The CALD lexical data for the top 6000 words had been tagged with three frequency levels, 'E' ('Essential'), 'I' (Improver'), and 'A' (Advanced), roughly representing bands of 2000 words each. Assuming that the 'E' and 'I' levels would prove the most likely be within the range of the CEFR levels, these frequency bands were scrutinized first. Therefore, these levels derived from CALD were used in order to cross-check against external evidence for rough calibration of CEFR level. The main source of such calibration were data from the Cambridge Learner Corpus, comprising 40 million words of mainly written language produced by learners worldwide in proficiency examinations. These proficiency examinations, in turn, have been aligned to the CEFR levels (Figure 4.1) by Cambridge ESOL (Milanovic, 2009; Salamoura, 2008)[38]. Hence, EPW researchers checked each word in the CALD lists to ascertain whether or not they were produced by candidates in the Cambridge Learner Corpus, and if so, at what point those items began to become represented in the various Cambridge ESOL proficiency exams (e.g. First Certificate, Certificate of Proficiency).

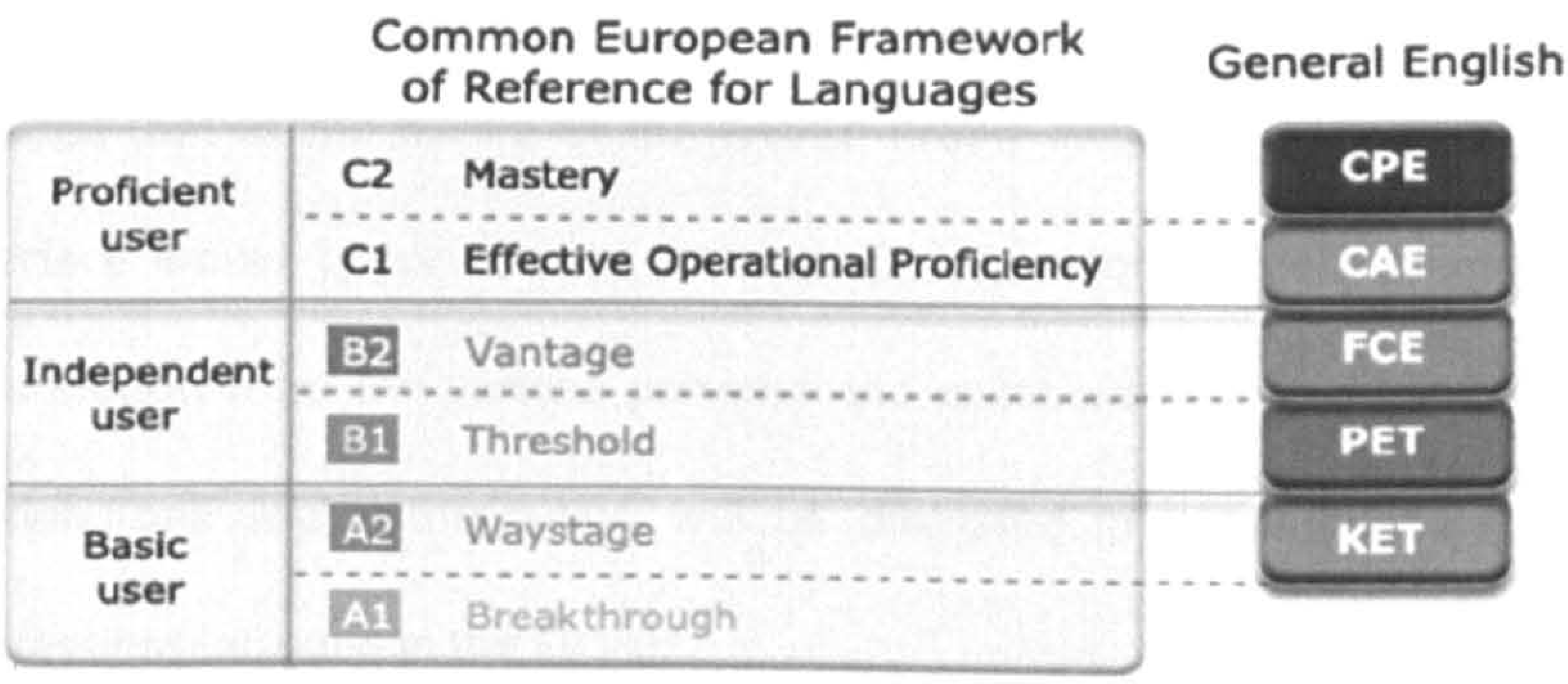


Figure 4.1. Alignment of CEFR with Cambridge ESOL examinations (as reported by Cambridge ESOL)

Annette Capel has claimed, however, that probably the most useful reference for her as the main researcher on the EPW project was Hindmarsh's (1980) *Cambridge English Lexicon* – a list of words prescribed by Hindmarsh as pedagogically relevant – which Capel consulted regularly throughout the list compilation process (Capel, personal communication). The reported reason for this perceived usefulness was the fact that Hindmarsh's list was intended to "produce a lexicon worth teaching and learning at the intermediate level of FCE" (Hindmarsh, 1980, cited in Capel, 2010: 3), and was divided into six levels that Capel asserts find counterparts in the CEFR bands. What is more, consistent with the aims of the EP lists, the items in Hindmarsh's list are all sense-sensitive.

[38] There is still little independent research into of the validity of this alignment, however.

The checking and cross-checking of data eventually produced a list of around 4,700 entries from A1 through B2 levels, though this number may yet change somewhat. First of all, the C levels – arguably the least defined of the CEFR reference-level descriptors in general (Capel, 2010) – are still, at the time of writing, in the development stage. Furthermore, though the list was originally intended to be sold in book and/or CD-ROM form, after some discussion at the early stages of the EPW development, it was decided that as the nature of the English Profile Programme was one of collaboration, a web-based interface would be preferable as it would allow for constant changing and updating as research contribution from the various partners and collaborators developed and increased. In the following section, one such contribution will be discussed in detail, that of the validation of the phrasal expressions included in the EP lists.

4.3 Comparing the PHRASE List to the phrases in the EPW

A number of researchers at the University of Cambridge who were also involved in the English Profile project came to learn of the PHRASE List development at a language testing conference in November of 2009. Those researchers expressed research interest in the list and passed that interest on to Annette Capel that same month, who immediately contacted this researcher to explain what role she saw for the PHRASE List in the validation of the EPW. In essence, Capel had carefully identified all the multiword expressions – which she called 'phrases' – in all the works of lexicography she had utilized during the EPW development, including those phrases that did not have their own dictionary entries but instead were embedded within other headwords.

Hence, although Capel felt that the work had been carried out with a high degree of thoroughness, the English Profile project wished to enhance the list by inviting input from outside researchers, and therefore the PHRASE List could, it was believed, help validate the work thus far realized by mostly one researcher (Capel). In particular, since the PHRASE List development had no *a priori* list from which it was derived – unlike the phrases in the EPW – there was a chance that there could be phrases in the PHRASE List missing in the EPW. The PHRASE List of course represents not all phrases but a set that met certain semantic criteria (the EPW phrases do not account for transparency), but there was bound to be overlap, and this overlap in itself would be of interest. To wit, since the PHRASE List contains frequency information, it was also requested that any apparent level mismatches between the CEFR levels attached to the EPW phrases and the frequency of the PHRASE List items by flagged and reported, and to the extent possible, indicate if those phrases might be good candidates for the C1/C2 levels in the EPW (at that time still underdevelopment). Therefore, the main research questions can be summarized as follows:

RQ1: Are all the phrases in the PHRASE List also in the English Profile Wordlist?

RQ2: Are any of the PHRASE List phrases that are missing in the English Profile Wordlist potential CEFR C-Level candidates?

RQ3: Do the CEFR levels assigned to the phrases in the English Profile Wordlist seem consistent with the phrases in the PHRASE List on the basis of their token frequency?

4.3.1. Checking the EPW for missing phrases

In order to carry out the comparison of the two lists, this researcher was given full access to the lists which were already online, but not yet open to the general public. The lists were contained in the form of a searchable online database, an example of which is shown in Figure 4.2.

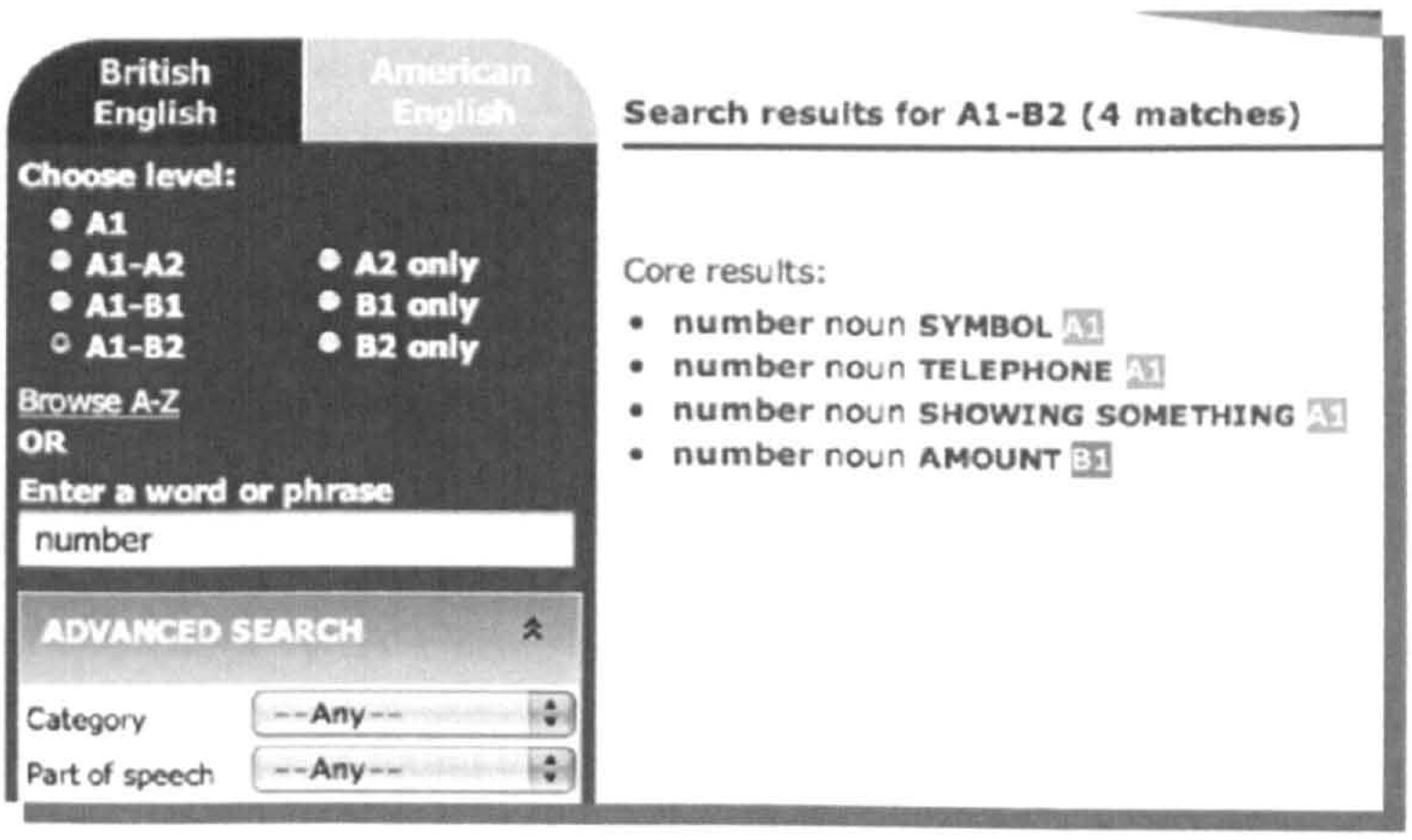


Figure 4.2. Example of EPW search query

On the word search page, one can choose to retrieve search results only from specific CERF levels, but there is also an option to choose everything from A1 to B2[39], which was always selected in the case of this research. Each phrase from the PHRASE List was then entered into the search query field (which may contain a word or phrase), to determine whether or not the phrase was already in the EPW. This exercise rendered a number of phrases which were common in both the EPW and PHRASE lists, listed in their entirety in Appendix 6, but shown partly in Table 4.2 for the purposes of exemplification.

[39] Eventually C1 and C2 levels will also be searchable.

Table 4.2 Partial list of phrases that overlap in both the EPW and PHRASE lists

Phrase	EP CEFR Level	BNC Frequency	Frequency Band
a lot	A1	22332	1K
as well (as)	A1	18041	1K
half past	A1	1325	4K
of course	A1	26966	1K
there is/are	A1	59833	1K
Would you like...?	A1	1133	5K
a bit	A2	19618	1K
a variety of	A2	4283	3K
all sorts of	A2	1535	4K
all the time	A2	3527	3K
and so on	A2	4584	2K
as usual	A2	1287	4K

However, the comparison also revealed 203 phrases that appeared to be missing in the EPW, or 40.2% of the entire PHRASE List. This initial account of all phrases that were found to be lacking in the EPW phrase list can be found in Appendix 7.

This initial discrepancy naturally raised the question of why the difference was so apparently large. (This question will be discussed further in the Discussion section.) However, when this disparity was first discovered, one of the first steps taken was to not assume that the phrases were in fact all missing but to instead carefully re-examine the research methodology employed in the search for them.

Therefore, as a next step, instead of entering the phrases as multiword items in the search query field of the interface, this time the component words of the phrases identified as 'missing' were entered. For example, one of the phrases in the PHRASE List that initially was found not to be in the EPW was 'a number of'. When entered as a phrase (i.e. *a number of*) in the search query field, no results were retrieved (Figure 4.3). However, when just the word *number* was entered, the 'core results' shown earlier in Figure 4.2 appeared. As can be seen in Figure 4.2, one of the senses listed for *number* is 'amount'. This sense immediately led the researcher to suspect that the phrase may indeed be in the EPW, but embedded in another headword (in this case, *number*), rather than as its own entry. Indeed, this was the case, as shown in Figure 4.4.

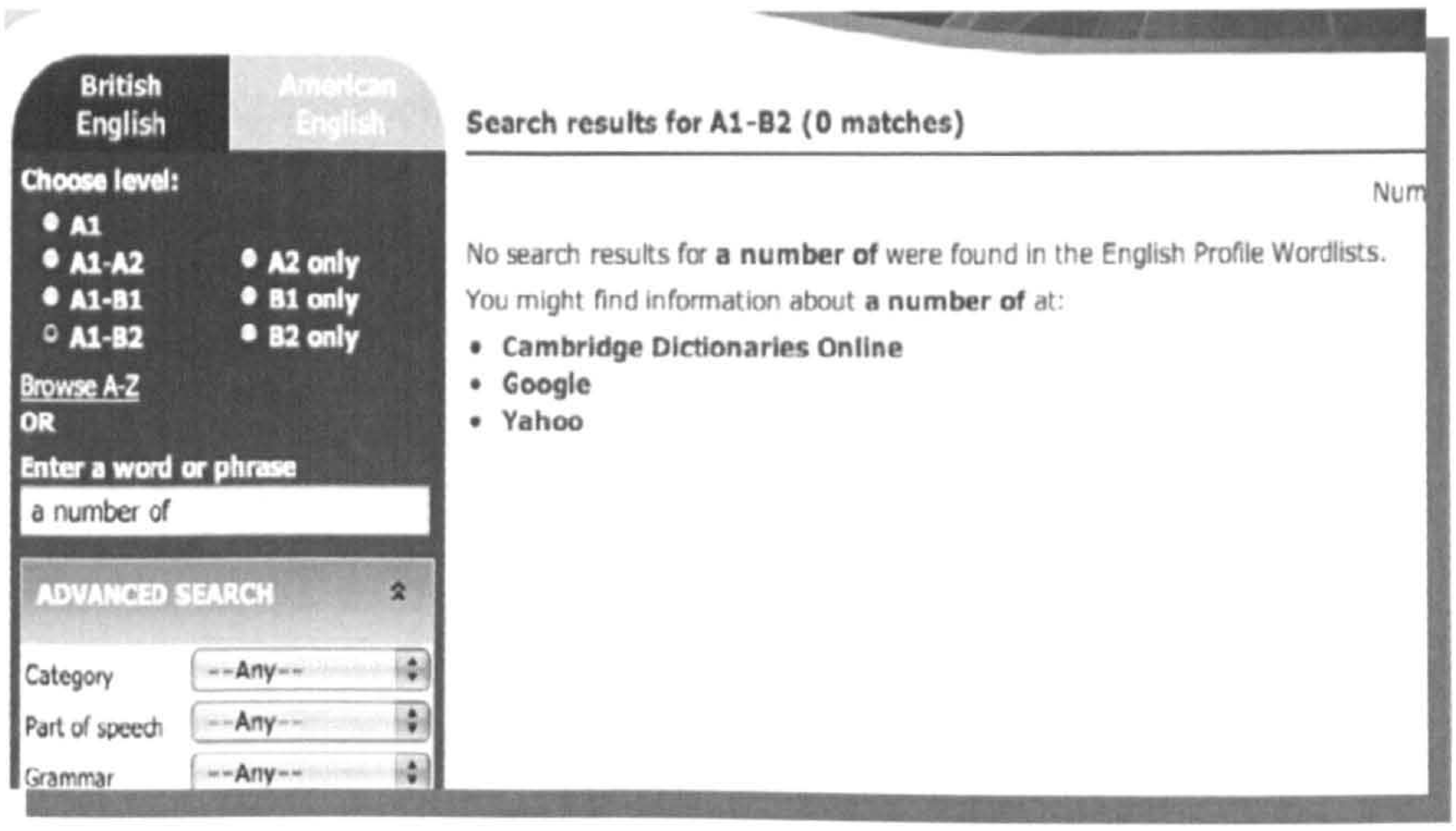
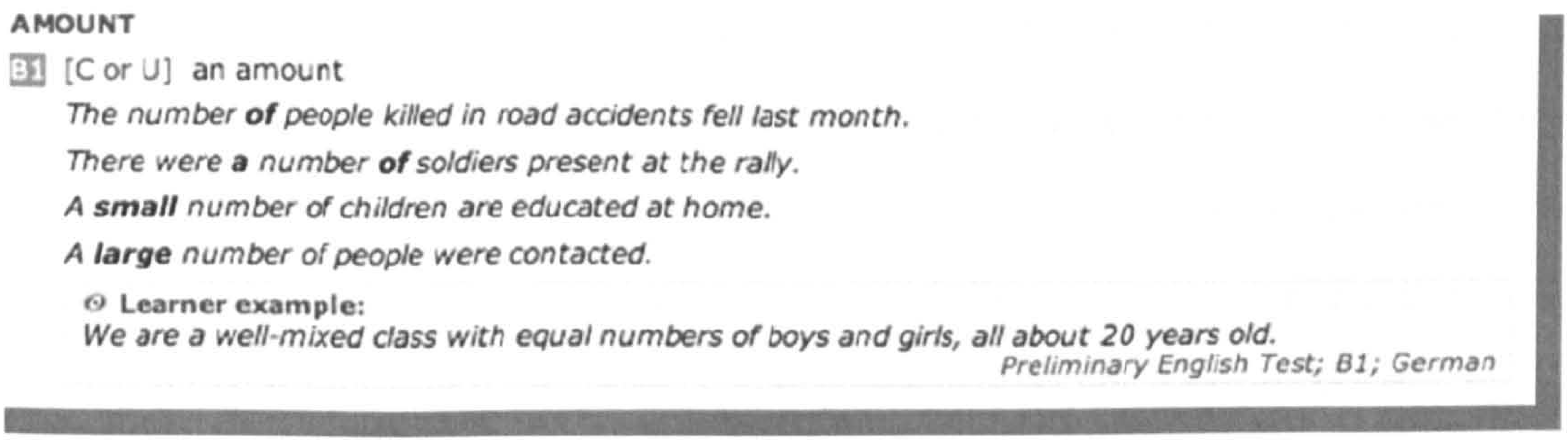


Figure 4.3. Example of phrase search query rendering no EPW results



AMOUNT

B1 [C or U] an amount

*The number **of** people killed in road accidents fell last month.*

*There were **a** number **of** soldiers present at the rally.*

*A **small** number of children are educated at home.*

*A **large** number of people were contacted.*

Learner example:

We are a well-mixed class with equal numbers of boys and girls, all about 20 years old.

Preliminary English Test; B1; German

Figure 4.4. Sense entry for *number* under 'amount', within the headword *number*

Figure 4.4 reveals a number of limitations of this first iteration of the EPW as it pertains to the phrases it contained, and these limitations will be discussed in greater depth in the Discussion section of this chapter. However, Figure 4.3 is a representative example of the procedure that was undertaken and the results that were obtained in this phase of the research. In case after case, the phrases that were initially identified as 'missing' in the EPW were in fact found to be embedded within the senses of other headwords. A full listing of these can be seen in Appendix 7. Also listed in Appendix 7 are phrases that were truly missing completely in the EPW. Many of those, however, were determined to be potential C-Level candidates, and this is explained in greater depth in the following section. It was only after the identification of the frequency level of the potential C-Level phrases that a truer number of missing phrases was finally arrived at.

4.3.2 Investigating issues of frequency and level in the EPW

Although the second and third research questions asked to explore issues related to level between the CEFR level assigned to the EPW phrases and the frequency band level of the phrases in the PHRASE List, it was clear from the beginning that the investigation could not render any kind of definitive, exact results. For one, the corpora from which the EPW lists are derived are different from the one which produced the PHRASE List (the BNC). However, more importantly, the very notion of comparing CEFR level to BNC frequency level carries problems. While it is generally true that the most frequent words are those which are taught/learned first in an L2, it is also true that there are words that feature in the first levels of English courses that are actually not very common in the BNC, relatively speaking. Words like *pencil, dictionary, teacher* and *classroom* – vocabulary often explicitly taught at the earliest level for their functional necessity in the context of formal instruction – may not necessarily fall into the highest frequency bands in wordlists derived from the BNC, but they are assigned CEFR A1 and A2 levels in the EPW. Moreover, there are very little data related to alignment of vocabulary size and the CEFR, though some research does exist:

> With vocabulary size linked to the level of particular examinations, and as these examinations are tied into the CEFR framework, it is possible to link vocabulary size to the CEFR. It might be expected ... that the vocabulary knowledge associate with each level would equate to some degree with the size of these lists. (Milton, 2009: 186)

In particular, Meara and Milton (2003) were able to correlate vocabulary size scores from Greek and Hungarian EFL learners taking the X-Lex test (a yes/no checklist test; see also Chapter 5) with various examinations, including the CEFR-aligned Cambridge ESOL examinations (Table 4.3).

Table 4.3 X-Lex scores linked with Cambridge ESOL examinations (Meara & Milton, 2003)

X-Lex score	Cambridge ESOL examination	CEFR alignment
2000 - 2740	KET	A2
2750 – 3240	PET	B1
3250 – 3740	FCE	B2
3750 – 4240	CAE	C1
4250 – 4490	CPE	C2

It should be noted that the X-Lex scores reported in Table 4.3 represent those from students who demonstrated that they could "take and pass" (Milton, 2009: 180) the respective Cambridge ESOL examinations, meaning they could achieve at least the minimum score on each of those tests. X-Lex may be considered particularly appropriate since one of the wordlists that helped create that test was the Hindmarsh (1980) *Cambridge English Lexicon* – the same ones used by Annette Capel in the development of the EPW. On the other hand, the X-Lex lists do not correspond directly with the BNC-derived wordlists, so one would need to qualify any claims of equivalence. Nonetheless, assuming the Cambridge-CEFR alignment is correct (see Discussion below), it should be possible to estimate a very general cut-off vocabulary threshold for the BNC lists. Milton (2009) suggests that students "will probably need 3750 words or more to move from CEFR B2 to C1" (p. 187). It was therefore initially assumed that any PHRASE List item less frequent than the 3000 word band (or < 2100 occurrences in the BNC) and which was also not in the EPW in any form (i.e. not as its own entry nor embedded in senses of other headwords) could be considered a C-level candidate. On the surface, although very rough, the Meara and Milton (2003) data seemed to provide some general parameters by which to judge possible level mismatches.

Additional support for the Meara and Milton alignment data could also be checked against the missing phrase analysis conducted. In Figure 4.5, the number of phrases identified as truly missing in the EPW is shown in relation to their BNC frequencies.

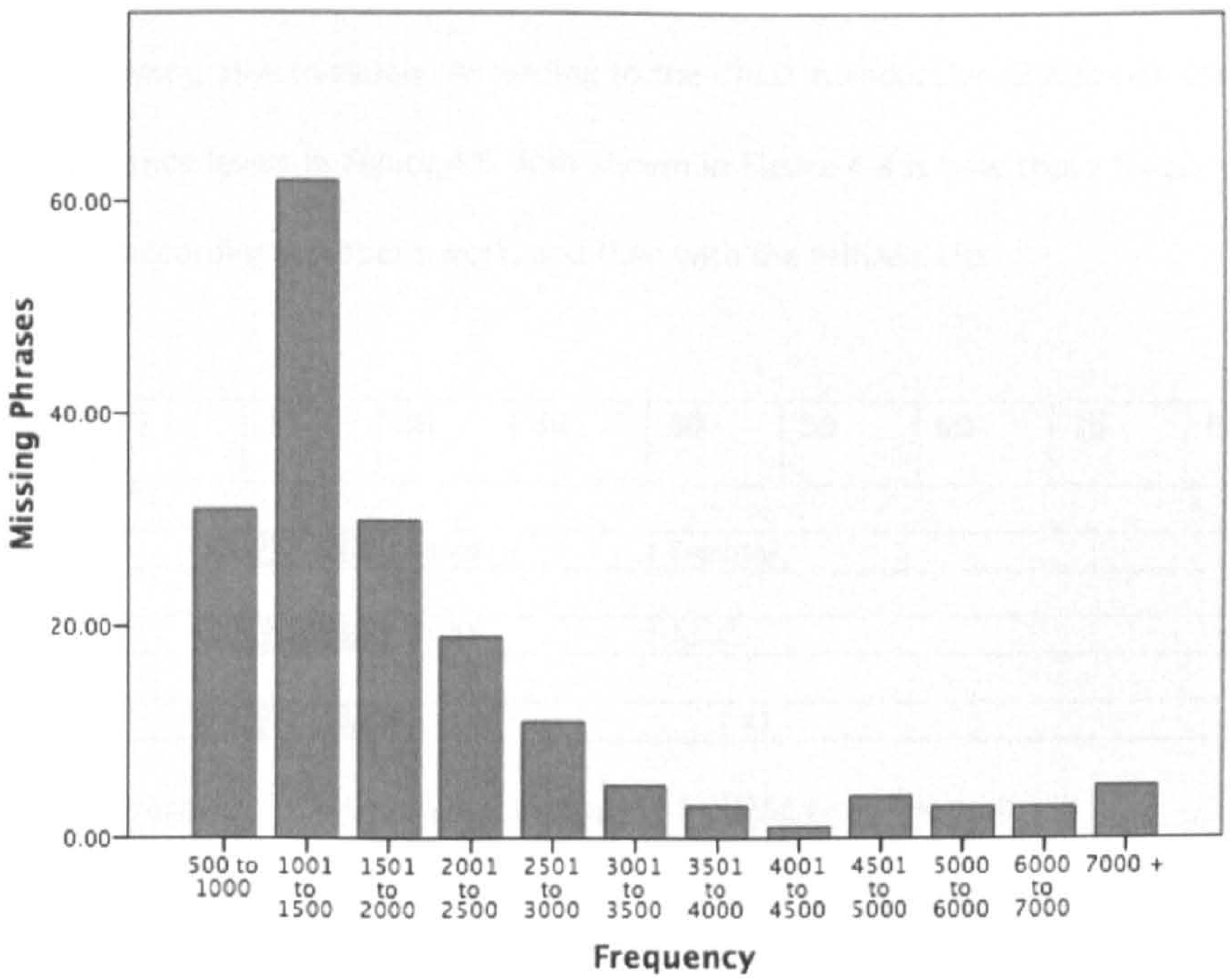


Figure 4.5. Frequencies of missing EPW phrases

What Figure 4.5 would seem to show is that from a BNC token frequency of about 3000 words, most PHRASE List phrases are accounted for in the EPW. Since at the time of the validation exercise only CEFR levels A1 to B2 were covered in the EPW, it can be assumed that, broadly, phrases in the PHRASE List with a BNC frequency lower than 2500 – which is about the middle of the 3K level – may be considered potential C-level CEFR candidates, rather than oversights on the part of the English Profile researchers. These rough cut-points would seem to be generally in line with the data from the Meara and Milton study.

The frequency data also make some sense when Capel's compiling rationale and methodology is revisited. According to Capel (2010: 3), she roughly equated the lemmas tagged as 'Essential' and 'Improver' as aligning with A1-B2[40]. This equivalence was merely a starting point for Capel in her compilation work, but the fact that those CALD tags are associated with corpus frequency data make them relevant, albeit vaguely. According to the CALD introduction (2008: viii), these tags correspond to the frequency levels in Figure 4.6. Also shown in Figure 4.6 is how those frequency levels overlap with the CEFR according to Capel's work, and then with the PHRASE List.

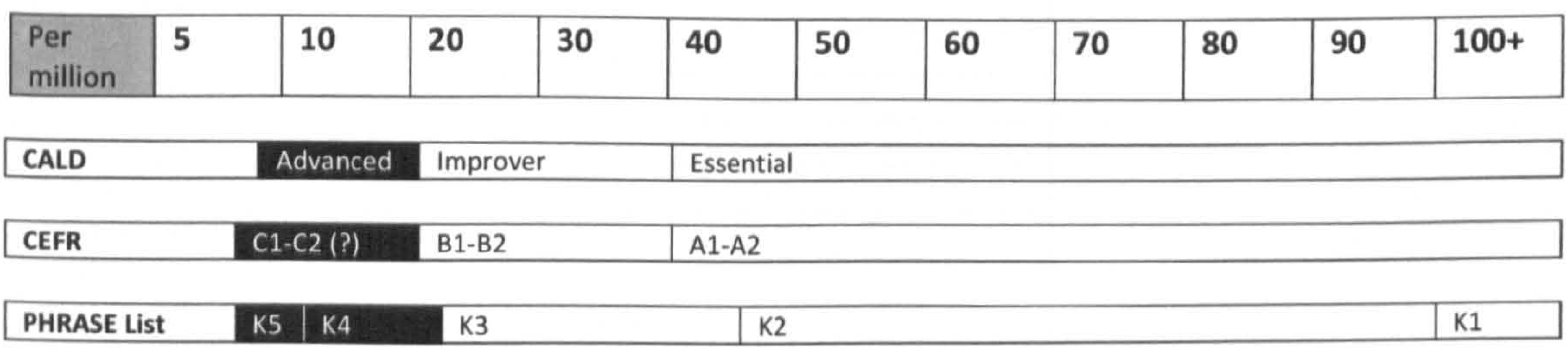


Figure 4.6. CALD frequency data and PHRASE List alignment

By extension, it can therefore be assumed that any phrases identified as truly missing in the EPW that also have a BNC frequency of over 2500 should be considered as A1 to B2 EPW candidates. These phrases (36 in total), and their respective frequencies, are listed in Table 4.4.

[40] It should also be noted, however, that Capel ended up also including over 50% of the lemmas tagged as 'A' in CALD upon cross-checking with other sources (Capel, 2010: 3), thus showing once again how any alignment can only be considered a very rough guideline, and how the C-Levels still were unclear at the time of this validation.

Table 4.4 Phrases indentified as truly missing completely in the EPW, and which are strong A1 to B2 candidates

Missing phrase	BNC Frequency
LED BY	2511
OUT THERE	2513
IN THE COURSE OF	2585
IN PART	2652
THAT'S IT (i.e. 'that's all)	2674
IN A WAY	2684
NO DOUBT ('SURELY')	2791
IN PLACE	2805
IN RESPECT OF	2909
THEY SAY	2962
IN THE EVENT (OF)	2998
KNOWN TO ('notorious for')	3091
SOMETHING LIKE ('AROUND')	3092
PRIOR TO	3110
NOT EVEN	3128
NO MORE THAN ('ONLY')	3226
IN TURN	3558
SAID TO BE	3586
IT TAKES	3670
OTHER THAN	4380
CONCERNED WITH	4619
IN (THE SENSE) THAT	4805
ALONG WITH	4948
COME TO ('EVOLVE TO')	4970
SUBJECT TO	5218
THAT IS (REPHRASING)	5737
SEEK TO	5937
A FURTHER ('ANOTHER')	6121
OR TWO	6192
CALL FOR STH	6243
YOU SEE	7102
WORK ON	7600
LAST NIGHT	7992
AS TO	11535
THOSE WHO	13951

Unfortunately, while there seems to be a very rough cut-off frequency in the PHRASE List that helps indentify which phrases might be good candidates for the CEFR C-levels, still to be defined and listed in the EPW, a look at the overlapping phrases in Appendix 6 shows that there does not seem to be a reliable frequency pattern that aligns the EPW and PHRASE List phrases. An analysis of the BNC frequencies and the CEFR levels of the overlapping phrases confirms this. Table 4.5 shows the corresponding frequency ranges for each of the EPW levels.

Table 4.5 BNC frequency ranges and the listed CEFR levels for the overlapping EPW phrases

EPW CEFR designation	BNC frequency range	Mean frequency
A1	1,133 – 59,833	21,605
A2	915 – 30,857	6,159
B1	788 – 21,085	4,635
B2	820 – 20,296	3,674

Indeed, a T-test reveals that only the difference between the A1 and A2 levels has any significant frequency difference ($t = 3.15$, $p = 0.008$). The difference between the B1 and B2 levels is not statistically significant in terms of frequency. Further, although there seems to be no clear definition of band cut-offs, a closer analysis within each band reveals a number of statistical outliers (Table 4.6).

Table 4.6 Phrases identified as outliers due to disparate frequency

Phrase	EP Level	BNC Frequency	Frequency Band
half past	A1	1325	4K
Would you like...?	A1	1133	5K
a bit	A2	19618	1K
I mean	A2	23616	1K
such a(n)	A2	23894	1K
such as	A2	30857	1K
at all	B1	14650	1K
carry out	B1	10753	2K
deal with	B1	13634	1K
in fact	B1	15983	1K
rather than	B1	21085	1K
used to (PAST)	B1	14411	1K
a little	B2	20296	1K
as if	B2	14470	1K
fail to	B2	10263	2K
lead to	B2	13555	1K
not only	B2	14110	1K
sort of	B2	13361	1K
tend to	B2	10504	2K

In finally answering the research questions, the phrases identified as truly missing in the EPW (Table 4.4), plus the phrases identified as clearly level-discrepant (Table 4.6), together with the suggested C-Level frequency cut-off (below 2500), were sent to Annette Capel with comments on those data and further commentary on insights which surfaced during the validation exercise. A summary of those comments is provided in the following section.

4.4 Discussion

In summary, all three research questions were addressed to some degree. There were phrases that were identified as not present at all in the EPW: a total of 178. However, 142 of those were determined to likely be good C-Level candidates, and therefore may only be missing in the EPW because at the time of the validation exercise the highest level represented in the list was B2. There were also phrases that were identified as statistical mismatches in terms of level, but only 20 items in total. On balance, it would seem that the PHRASE List and the EPW are largely in agreement, at least in as far as the phrases that overlap are concerned.

However, although the research questions were addressed, new questions arose during the exercise:

- There were many phrases that were initially identified as missing in the EPW but that were later discovered to be 'hidden' within other headwords. Why were they not listed as their own entries?
- There were 36 PHRASE List expressions that were identified as truly missing in the EPW. How can their absence be explained?
- There were 20 phrases identified as clearly being discrepant in terms of level. What might explain their disparity?

These questions will be explored further below.

4.4.1 Why did some EPW phrases have their own entries, and others not?

As illustrated earlier in Figure 4.3, a search for *a number of* as a phrase originally did not render any results from the EWP web interface, yet the phrase was found to be embedded within the sense of 'amount' under the headword 'number'. Why, therefore, had this phrase – along with several others (Appendix 7) – not been deemed as a phrase worth pulling out as its own entry by the original lexicographers? (All of the entries are derived from at least one of the dictionaries used by the EPW researchers.) A careful semantic analysis of the examples listed under 'amount' in Figure 4.4 might provide some insight:

AMOUNT

B1 [C or U] an amount

*The number **of** people killed in road accidents fell last month.*

*There were **a** number **of** soldiers present at the rally.*

*A **small** number of children are educated at home.*

*A **large** number of people were contacted.*

First of all, it is worth noting that it is a common learner dictionary convention to set common collocates of words in bold. Therefore, the word 'small' is set in bold before *number* to signal that it is a collocate worth noticing. Interestingly, this would also indicate that in the conception of the lexicographers responsible for this example (originally derived from CALD), the morphemes 'a' and 'of' surrounding the word *number* are collocates of the word – which in turn suggests a de-composition of a phrase that in the PHRASE List is considered more holistically, as if it were essentially a single morpheme.

Indeed, if the word *number* consistently means 'amount' in all the examples provided in the dictionary, then the expression *a number of* would be more of a three-word combination than an MEU. However, while most of the examples above do show the word *number* with the sense 'amount', it is not clear that *a number of* carries that same meaning. If that were the case, then the word *number* should be able to be substituted by the word 'amount' in all the examples. However, consider the following:

AMOUNT

B1 [C or U] an amount

*The amount **of** people killed in road accidents fell last month.*

There were **an amount **of** soldiers present at the rally.*

*A **small** amount of children are educated at home.*

*A **large** amount of people were contacted.*

Although perhaps not grammatically, semantically all of the examples work above – except for 'an amount of soldiers'. Written this way, it becomes clear that 'a number of' actually carries its own holistic meaning as a phrasal expression, a kind of adverbial meaning 'several' (*There were **a number of** soldiers... = There were **several** soldiers...*). It is also worth pointing out that, again in Figure 4.4, the 'Learner example' (taken from the Cambridge Learner Corpus of exam scripts) strictly reflects the word 'number' meaning 'amount':

Learner example:
We are a well-mixed class with equal numbers of boys and girls, all about 20 years old.
Preliminary English Test; B1; German

Therefore, there seems to be compelling evidence that the phrase 'a number of' should re-examined both in terms of its merits as its own headword in the dictionary, and also in terms of its designated CEFR level. Indeed, this was the recommendation made to the EPW researchers and the entry was changed, as shown in Figures 4.7 and 4.8.

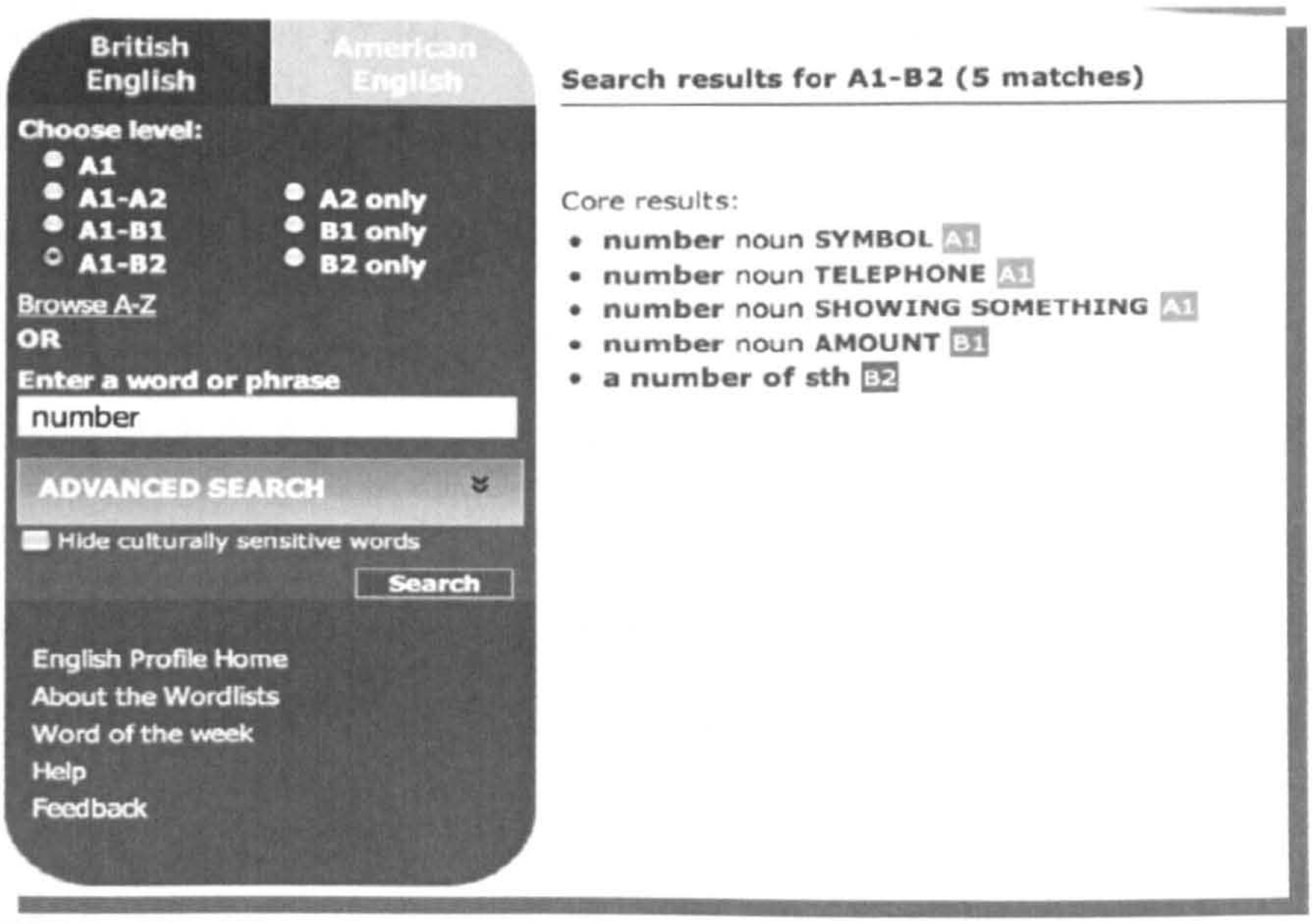
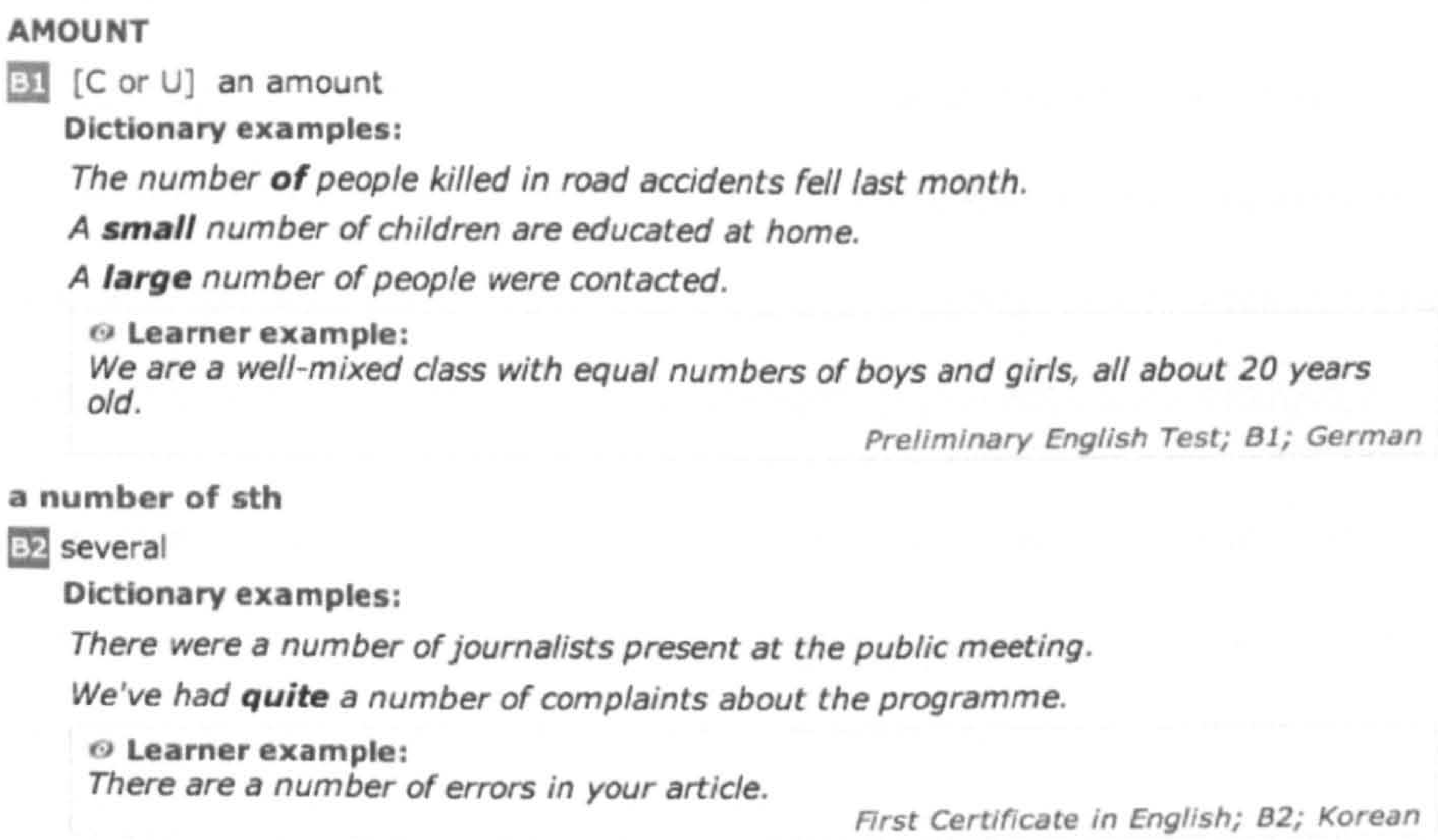


Figure 4.7. Updated search results for the word *number* following recommended changes



AMOUNT

B1 [C or U] an amount

Dictionary examples:

*The number **of** people killed in road accidents fell last month.*

*A **small** number of children are educated at home.*

*A **large** number of people were contacted.*

Learner example:

We are a well-mixed class with equal numbers of boys and girls, all about 20 years old.

Preliminary English Test; B1; German

a number of sth

B2 several

Dictionary examples:

There were a number of journalists present at the public meeting.

*We've had **quite** a number of complaints about the programme.*

Learner example:

There are a number of errors in your article.

First Certificate in English; B2; Korean

Figure 4.8. Updated entry for *number*, now showing *a number of* as separate phrase

As shown in Figure 4.7, the EPW researchers agreed that the semantics of *a number of* were indeed not the same as 'amount', and that it should be pulled out as a separate expression. Moreover, Figure 4.7 also shows that the phrase *a number of* has now been labeled as a B2 lexical item, different from the sense of 'amount', which has remained B1. This change in CEFR level for that phrase resulted from a consultation of the Cambridge Learner Corpus by Annette Capel (personal communication), which revealed that there was little to no evidence of that expression in any of the exam script data below those learners who passed the FCE. Moreover, as this phrase has now been reclassified as B2 in the EPW, it would also qualify as a phrase with a level disparity in relation to the PHRASE List The phrase 'a number of' occurs 15,090 times in the BNC, qualifying it for the first 1000-word frequency band, which would suggest that it would be more expected to appear in the A1 or A2 Cambridge Learner Corpus data – but for whatever reason does not. (This issue will be discussed further below in Section 4.5.3).

Figure 4.8 shows the way in which *a number of* is now displayed in the EPW, as a separate phrase. It also shows that the morphemes 'a' and 'of', which were set in bold in the earlier version of the entry, are now no longer set apart, implying that they are not separate components from 'number' in that phrase. Instead, Annette Capel consulted the corpus and found that a common collocate of *a number of* is the word *quite* (***quite** a number of*), and so the phrase *a number of* has received treatment here akin to a word, not broken down into individual words, with its own collocational behavior.

There were of course other phrases from the PHRASE List which were identified as 'hidden' in various senses in the EPW. A full listing of these is provided in Appendix 7, but Figure 4.9 below provides a fairly good example of this phenomenon.

ROUTE

A2 [C] the route you take to get from one place to another

Dictionary examples:

*Do you know the way **to** the train station?*

*I've only been living in Madrid for a couple of weeks so I don't really know my way **around** it yet.*

*We'll have to stop for fuel **on the** way to the airport.*

*Can you **find** your own way **out of** the building?*

*It's getting late - we should **make** our way home soon.*

*The coach stopped for us to eat lunch but within half an hour we were **on our** way/**under** way again.*

*There's no way **through** the centre of town in a vehicle - it's for pedestrians only.*

*Many people have **lost** their way in the forest.*

*Only a local person could **find** their way **through** the narrow streets of the old town.*

Learner example:

The shortest way to get there is going along Sarmiento Avenue.

Key English Test; A2; Spanish

Figure 4.9. EPW entry for 'way'

Under the sense 'route', there is a display of complex phraseology in the example sentences shown in Figure 4.9. Arguably, a great many of those examples should be pulled out as separate phrases. Similar to 'amount' in with the word *number*, it is not clear that the meaning of 'route' holds for all the examples (e.g. *The coach stopped for us to eat lunch but within half an hour we were *under route again*[41]). Furthermore, just as in the example of *in order to*, if one or more of those phrases can be considered separate lexical items, they too should also be re-analyzed for their own respective CEFR level designation. This is indeed what occurred with a few of the phrases in Figure 4.9, including *make your way*, now a separate entry (Figure 4.10).

[41] On the recommendation of this researcher, this phrase is no longer embedded within the sense of 'route' and now has its own entry as a phrase.

make your way

B2 to get to a place

Dictionary examples:

We slowly made our way down the river.

It's getting late - we should make our way home soon.

Learner example:

She made her way slowly to the waiting room and patiently waited for the train to arrive.

First Certificate in English; B2; Azerbaijani

Figure 4.10. New entry for *make your way*

As seen in Figure 4.10, the phrase now gets its own definition ('to get to a place') and the CEFR level has been elevated from A2 to B2, which is much more in line with the PHRASE List frequency data (BNC frequency = 1446, or the 4K level).

At the time of writing, there are still a number of phrases under consideration for being removed from their current place embedded within other senses to be listed as phrases in their own right. It is anticipated that the phraseology of the type displayed in the examples in Figure 4.9 will be more carefully analyzed as the EPW grows and develops. To some extent, the very fact that the researchers at EPW re-evaluated many of the PHRASE List phrases that were embedded in other senses and decided to make them their own entries also provides some validation for the methodology employed in the selection of phrases for the PHRASE List. In carefully considering semantic properties of phrases, such as transparency and non-compositionality, the PHRASE List was able to identify multiword items that lexicographers and other researchers had misidentified as extensions of the senses of their component words.

The following section will consider phrases that were not 'hidden' within other headwords, but in fact missing completely in the EPW.

4.4.2 *Why were some phrases missing in the EPW?*

As explained in detail in Section 5.4, although initially there were over 200 PHRASE List expressions found to be missing in the EPW, after finding that many of them were in fact embedded within senses of other headwords, and establishing a frequency cut-off for C-level phrase candidates, there were 36 phrasal expressions that remained unaccounted for. There are likely a number of explanations for their absence in the EPW, and the plausible explanations will vary from expression to expression. Table 4.7 lists a few of the missing phrases, with frequency information for each, and also what the feedback was on each of the phrases from the English Profile researchers working on the lists at that time (the names included are pseudonyms to protect anonymity). Through these phrases some of the aforementioned plausible explanations will be explored.

Table 4.7 Missing phrases + feedback

Phrase	**BNC / 100 m**	**EP comment**	**(Proposed) EPW level**
LED BY	2511	*C levels if at all, but I can't find this in the Macmillan English Dictionary (MED), and it feels rather rare to me. BNC written-biased corpus? Sally, any view?*	?
OUT THERE	2513	*I hear/see this a lot, but MED only has 'out there' as in outside and far away, whereas the 'new' use is 'widespread'?*	?
IN PART	2652	*A formal phrase, which is C1 in my view. There are a few cites at FCE but only poor attempts.*	C1
THAT'S IT (i.e. 'that's all')	2674	*I use this all the time! MED says it's spoken, but certainly common in email use. It is an omission from EPW, lots of cites at PET as an ending, eg 'That's it for now', and even some at KET.*	B1
IN PLACE	2805	*Evidence at CAE eg put in place, measures in place, etc.*	C1
THEY SAY	2962	*Help! Not sure how to include this one...*	
IN TURN	3558	*Decided this was above B2. It will be added for Cs.*	C1
IT TAKES	3670	*Sally, is this an omission, or save for Cs?*	?
ALONG WITH	4948	*C level phrase.*	C
OR TWO	6192	*Mainly spoken, I suspect. We should include it but I'm unsure of level!*	
YOU SEE	7102	*Spoken again. MED has phrase You see. I agree it is different. Sally, can we include this?*	?

AS TO	11535	*Formal use makes it C1 I think. Lots of cites at CAE.*	C1
THOSE WHO	13951	*Not sure how this could be presented... what would be a typical NS use?*	?

A look at the comments reveals a number of points. First of all, the researchers consulted dictionaries they had not used extensively in the original development of the EPW, in particular the Macmillan English Dictionary (MED). This consultation took place by those researchers in part to confirm that the phrase in question was also considered a phrase by another authority, but also, since the MED also indicates frequency information for many of the entries, it could also help with guidance as to level. Second, there are also some phrases (e.g. *out there*) that seem to be missing in all the dictionaries at the disposal of the EP researchers, but which the EP researchers believe should be included as phrases. Finally, there are a number of phrases that the EP researchers seem to feel are of a more formal register and so, despite their relative high frequency, should be allocated to the C-levels, and last, there are even a few phrases that the EP researchers do not feel very confident about.

In terms of the phrases that were found in the MED, the expression *that's it* in Table 4.7 stands out as interesting. Through the researchers' comments, it can be observed that they do not agree with the MED's reporting of the phrase as 'spoken'. This concern with items belonging to a spoken mode rather than written is a concern for the researchers since nearly all of the Cambridge Learner Corpus data are written. Therefore, there are a number of phrases which may be absent or at a different level (see also Section 4.5.3 below) because they are not written or written very often (e.g. *you see*). However, in the communication (to 'Sally'), it can be seen that the phrase is seen as perhaps important enough that it may merit consideration as a phrase.

Returning to the example of *that's it*, another insight that the EP researchers' commentary reveals is the process of determining a CEFR level. The phrase is determined to be useful for the EPW, but a level still needs to be decided. The researchers apparently went back to the learner corpora and found enough examples to confirm that it is definitely not a C-level phrase. The EP researchers also notice that while it is even present in the earliest Cambridge ESOL examinations (i.e. KET and PET), the use of the phrase seems to be mostly confined to formulaic memorized use, namely in the token *That's it for now*. The researchers therefore arrive at a suggested CEFR level assignment of B1.

It also appears that, at least to the minds of the EP researchers, the more formal the register, the more likely the phrase should be limited to the C-levels. In the example of *in part* in Table 4.7, for instance, the researchers seems to conclude that the phrase is *formal*, but cites no other rationale beyond his or her own idiolect or native speaker instinct. Moreover, despite some evidence that the phrase may be of a B2 or even B1 level, the 'feeling' the EP researcher has about the phrase seems to almost sway the researcher into relegating it to the C-levels.

Finally, there are a few phrases for which the researchers cannot find guidance in existing works of lexicography, and which they seem to not be sure about in general (i.e. in terms of level or presentation). The phrasal expression *led by*, for example, is at the cusp of the C-level cut-off (which itself is of course a very rough guideline), but the researcher may be correct in suspecting that it may be more appropriate for the most advanced of the CEFR bands. On the other hand, the researcher seems to question its merits of conclusion at all ("C levels if at all…"). The phrase *led by* met a number of the PHRASE List selection criteria, including (lack of) transparency, and the auxiliary criterion of grammar. While the verb *lead* in its active voice, transitive sense is mostly quite clear (e.g. 'Obama still leads the polls'), the phrase *led by* departs from that meaning (e.g. *Led by their greed, the managers bankrupted the company*) – a meaning closer to 'compel'. It is possible that lexicographers, who did not have an n-gram list at their disposal when including phrases in their dictionaries, simply never considered the phrase when writing the entry for *lead*. Lexicographers work from lemmatized lists of words (Kilgarriff, 1997; Sinclair, 1987), and any phraseology they include is either derived from existing works, or extracted from exploration of the words in a corpus, usually by looking at collocations in a concordancer. However, even if lexicographers did notice the phrase 'led by' when conducting searches for collocations, it is unlikely that they would have considered a separate phrase since they would be mostly concerned with common patterning as opposed to any kind of semantic opacity – which was key in the PHRASE List.

The EP researcher is most likely correct that the written bias of the BNC makes the phrase particularly common in that corpus, but this in itself should not form the basis for exclusion of the phrase altogether. The fact that the phrase appears over 2500 times in the BNC should be enough evidence that it at least deserves a closer look.

In general, as with the embedded phrases discussed in the previous section, there seems to have been some advantages to letting the n-gram data provide the raw material from which the PHRASE List was derived. At the very least, it showed that such a method can help identify phrases that perhaps would not otherwise have been noticed or considered – even by teams of lexicographers consulting very large corpora.

4.4.3 How can the level discrepancies be explained?

Table 4.6 showed that there were a number of phrases with clear-cut level discrepancies in the EPW, in terms of their divergence from PHRASE List frequency data. In addition, some of the phrases which were embedded in other senses, but which are now becoming their entries, are also currently receiving new

CEFR assignments (see the example of *a number of* in Section 4.5.1), and so the number of aforementioned clear-cut discrepancies is likely to grow larger still. Regardless of their number, the EPW phrases that seem inconsistent with the PHRASE List data can mostly be explained by two causes: corpora and task.

First of all, it is important to remember that only the phrases that were identified by the various works of lexicography used by the EP researcher <u>and</u> which were also present in the Cambridge Learners Corpus data are represented in the EPW. Further, it is not until a phrase is judged to be frequent enough in the learner examination data that a level is generally assigned. Hence, even if there are few examples of 'in turn', for instance, in the examination data at PET (B1) level, it will be assigned a B2 level if it is determined that widespread use of the phrase only occurs in the FCE (B2) data. In short, the level designations in the EPW are not a function of 'real-world' frequency, but instead of learner output, which of course is likely to be different.

Indeed, the output itself is also an artifact of the examination itself, and therefore the language that is used is also influenced by the tasks contained in the tests. With that in mind, it becomes clearer why a phrase like 'half past' is assigned such a low CEFR level in the EPW when it is so relatively low in frequency in the PHRASE List (Table 4.7). It is well known that one of the first functions that is commonly taught in EFL classes is that of telling time, often in the context of talking about daily routine (itself usually a pretext for practice of the present simple tense). It is also true that tasks that involve the telling of time and talking about routine often feature in the lowest proficiency examinations, such as KET and PET. At least according to the BNC data, the time of day is not communicated so often in naturally-occurring discourse.

Furthermore, since representation in the learner examination data was the greatest determinant of level assignment in the EPW, the levels are largely reflective of productive rather than receptive knowledge of vocabulary. (Although the learner dictionaries and the Hindmarsh (1980) volume also were used to inform decisions regarding level.) Therefore, it is conceivable that 1) the phrases used by the students in the written examination data are mostly those with which they feel most confidence in using in that high-stakes situation, which may not include phrases like *in fact* and *rather than* (Table 4.7), but which they actually do recognize and understand; and 2) that the tasks required of the students by the exams may not necessitate and/or lend themselves the use of the phrases in Table 4.7.

Finally, there is one more variable, not yet mentioned in the present discussion, which may be exerting profound influence on the level assignments of the phrases, and it is to do with the validity of the alignment of the CEFR levels with the Cambridge ESOL examinations. Though an extensive discussion of the validity issues that may subvert the relevance of that alignment lie beyond the scope of the present chapter, concerns have indeed been raised in the language testing community. The CEFR was never intended to be linked directly to language tests, and suggestions have been made that the alignment may be more driven by commercial interest than true demonstrated equivalences (de Jong, 2010). Obviously, if the alignment of the Cambridge ESOL examinations is not accurate or lacks validity, then the Cambridge Learner Corpus which has largely informed the decisions regarding level must also be revisited, and levels perhaps reconsidered.

4.5 Conclusion

In the final analysis, although the PHRASE List was of course instrumental in answering, or at least attempting to answer, the research questions posed at the beginning of this chapter, the validation exercise also provided some new insights into the advantages of using the methodology employed in the development of the PHRASE List. With its unique attention to the semantic properties of phrases during its compilation, the PHRASE List helped to identify phrases that were later determined by English Profile researchers to have been mistakenly embedded within the senses of other headwords, when in fact they merited entries of their own. Moreover, allowing an exhaustive n-gram list to inform the PHRASE List, rather than using existing lists, was shown to be effective in identifying phrases that had apparently been missed by lexicographers and authors of previous works, even those informed by large corpora.

However, perhaps what proved most valuable in this validation exercise was that, in validating the EPW, the EPW also helped to validate the PHRASE List. Although there were important differences between the lists, as discussed at length in previous sections, the fact is that the vast majority of PHRASE List expressions had also been identified as phrases by the EPW compilers (only 36 unaccounted for in the A1-B2 range). What is more, since the PHRASE List phrases identified as truly missing in the EPW then required further evaluation by the EPW researchers, and since in nearly all cases they were eventually deemed worthy of inclusion in their list at some level, even the items which did not overlap found independent validation. As mentioned in Chapter 3 (Section 3.5), it was in part thanks to the close scrutiny of people like Annette Capel that more careful attention to genre was eventually allocated in the PHRASE List. All in all, it can be said that both lists were enhanced in some way by the exercise.

Chapter 5 – Case Study 2: Development of a PHRASE test

The present chapter reports on the preliminary development of a test designed to measure knowledge of the phrasal expressions on the PHRASE List. As lexical items of the type contained in the PHRASE List had heretofore not been tested in the same way individual words are commonly assessed, the focus in the present chapter is to some extent on the process of test development, rather than the product, reflecting the exploratory nature of the research. (Nonetheless, the chapter does present a usable version of the test in Appendix 8.) It should be emphasized that the validation of the test as reported in the present chapter is therefore incremental and in no way should be interpreted as research that requires no further development. Indeed, in many ways, the validation discussed in this chapter should be viewed as only a beginning. Therefore, general issues surrounding vocabulary testing are discussed in Section 5.1, including a review of some of the more popular existing vocabulary tests. Section 5.2 reports on the method employed in the test development, including how phrasal expressions were selected for item writing, and how a test format was ultimately arrived at. Section 5.3 presents the results of a field test of three versions of the test that was developed, here entitled the Phrasal Vocabulary Size Test (PVST), including a validation exercise conducted on the last version of the instrument developed for the present research. Section 5.4 concludes the present chapter, discussing paths for further development of the test.

5.1 Vocabulary testing, tests, and issues

The main question addressed in this chapter is *How can a test of phrasal vocabulary be developed?* However, to begin to answer that question, it is first important to review how 'traditional' vocabulary is currently tested.

To a certain degree, all language tests assess knowledge of vocabulary. After all, without understanding at least some of the words that one is reading or hearing, there can be very little comprehension. Therefore, in integrative language tests (i.e. tests of language skills rather than discrete linguistic systems), especially when involving receptive skills (i.e. reading and listening), lexical deficiencies tend to at least be partially responsible for lower scores (Schmitt, Jiang & Grabe, 2011; Staehr, 2008). Although the focus of this chapter will be on the specific testing of vocabulary, or what Read (2000) has called testing vocabulary as a discrete construct (Figure 5.1), it may not be useful to conceive of vocabulary testing as an isolated measure.

Discrete	⟷	**Embedded**
A measure of vocabulary knowledge or use as an independent construct		A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective	⟷	**Comprehensive**
A measure in which specific vocabulary items are the focus of the assessment		A measure which takes account of the whole vocabulary content of the input material (reading/listening tasks) or the test-taker's response (writing/speaking tasks)
Context-independent	⟷	**Context-dependent**
A vocabulary measure in which the test-taker can produce the expected response without referring to any context		A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Figure 5.1. Three dimensions of vocabulary assessment (Read, 2000: 9)

The DIALANG (DIAgnostic LANGuage) assessment system (Alderson, 2005), for example, is a web-based low-stakes test[42] designed to assess proficiency in 14 different European languages. Although the main part of the test includes more 'embedded' testing of vocabulary through the skills of reading, writing and listening, the online interface also includes an adaptive mechanism that can be calibrated at the outset of the test through a 'discrete' yes/no test[43] called the Vocabulary Size Placement Test (VSPT),

[42] According to Fulcher (2010: 322), a *low-stakes* test is "[a]ny test ... in which the outcome has few or no consequences for the takers, institutions or society."
[43] A yes/no vocabulary test (also known as a 'checklist' test) requires the test-taker to indicate whether s/he knows the meaning of a word, and usually includes non-words to help control for guessing.

based on the Eurocentres Vocabulary Size Test (Meara & Jones, 1988, 1990 –described more later in the present chapter). What is interesting, however, is that Alderson (2005) found high correlations between the discrete vocabulary measures in DIALANG, and the embedded ones (p. 79):

Reading	.64
Writing	.70
Listening	.61

Alderson therefore concludes that discrete measurement of vocabulary knowledge is a worthwhile pursuit even on its own:

> What this would appear to show is that the size of one's vocabulary is relevant to one's performance on any test, in other words, that language ability is to quite a large extent a function of vocabulary size. (...) From the point of view of the diagnosis of language strengths and weaknesses, a measure of vocabulary size would therefore appear to be of considerable value in its own right, let al.one as a quick and reliable placement procedure for more detailed diagnoses of different aspects of language ability. (p. 88)

Although very few people would argue that discrete tests of vocabulary alone can replace integrative tests of general proficiency such as the IELTS or TOEFL examinations altogether (Hughes, 2003: 79), the findings of Alderson (2005) at the very least provide evidence of the degree to which vocabulary assessment is integral to just about any language test.

Four existing vocabulary tests, the Vocabulary Levels Test (Nation, 1983; Nation, 1990; Schmitt, Schmitt & Clapham, 2001), the Eurocentres Vocabulary Size Test (Meara & Jones, 1988), the Productive Vocabulary Levels Test (Laufer & Nation, 1999), and the Vocabulary Size Test (Nation & Beglar, 2007), will be described in Section 5.1.1. These tests have been chosen for the present discussion because they are some of the most widely used instruments for the discrete testing of vocabulary, and as all four currently only test individual orthographic words, the extent to which the PHRASE List can be utilized to integrate phrasal expressions into them will also be explored. Important concepts in testing, such as issues to do with test *validity* (a concept revisited later on in this chapter), will also be explored through the discussion of these tests.

5.1.1 *Current vocabulary tests of single orthographic words*

As Paul, Stallman and O'Rourke (1990: 1) have noted regarding the testing of word knowledge, "the choice of test format depends on the type of information desired." The four tests discussed in this section all test vocabulary discretely, but each also contains a different test format. The Eurocentres Vocabulary Size Test, for example, has a yes/no 'checklist' type of format that mostly asks if the task-taker recognizes the word and its meaning; such a test can be said to be assessing vocabulary on a 'receptive' level (Daller, Milton & Treffers-Daller, 2007). On the other hand, Laufer and Nation's (1999) 'vocabulary size test of controlled productive ability' – also called the Productive Vocabulary Levels Test – requires the test-taker to be able to retrieve word forms from long-term memory, and therefore can be thought to be assessing the construct of productive vocabulary knowledge.

However, while a number of authors have in the past drawn a distinction between receptive and productive vocabulary knowledge (e.g. Crow, 1986; Goulden, Nation & Read, 1990; Webb, 2005), there is also evidence that calls into question the validity of that distinction. Laufer (1998) used the Vocabulary Levels Test, the Productive Vocabulary Levels Test and the Laufer and Nation's (1995) Lexical Frequency Profile (a measurement of lexical richness in written L2 production) to test the vocabulary of two groups of Israeli learners of English during one week of instruction – one group of 10th graders, and another with one more year of tuition. Laufer found that that with the benefit of one more year of instruction in English, the 11th graders showed an increase in both active and passive vocabulary, but that the gap between the two types of knowledge widened particularly at the higher levels of proficiency (i.e. lower frequency words). Further, using the Lexical Frequency Profile, Laufer found that gains found between the two groups will nearly nullified at the free expression level.

Another, more recent, study by Webb (2008), however, finds that the evidence for the distinction is not as compelling as in Laufer (1998). Webb tested 83 adult Japanese EFL learners using translation tests: to assess receptive knowledge, participants saw L2 forms and then had to supply the Japanese character(s) for that meaning; for demonstration of productive knowledge, the same test-takers were given the L1 (Japanese) meanings alone and then were required to write the equivalents correctly in English. As in the Laufer (1998) study, the receptive scores were significantly higher than the productive scores. However, unlike the Laufer study, Webb also gave credit for partial knowledge of the written form in the tests (i.e. 'sensitive scoring'), and using that methodology found no statistically significant interaction for type of word knowledge. Hence, although further research is of course still warranted, what Webb's study seems to suggest is that although receptive vocabulary knowledge does seem to consistently exceed receptive knowledge, the distinction between the two is sensitive to frequency level, and is not a black and white construct, but rather a matter of degree.

Schmitt (2010) suggests a problem lies in the extant terminology employed in the literature on receptive verses productive vocabulary measurement. Without clear terms to delineate the constructs being measured on the vocabulary tests being reported on in various research studies, interpretation of research results can become confounded. To address this issue, Schmitt draws four distinctions for the purposes of describing items types in vocabulary tests (p. 86):

- **form recall**: meaning given, L2 form must be produced;
- **form recognition**: meaning given, L2 form must be recognized;
- **meaning recall**: form given, meaning must be produced;
- **meaning recognition**: form given, meaning must be recognized.

According to Schmitt, the function of the four distinctions is to clarify that the information that discrete item vocabulary tests in fact provides about the test-taker is not so much related to a passive/active or even receptive/productive distinction (which they cannot accurately nor conclusively measure), but simply that different item types will provide (i.e. 'tap into') certain types of word knowledge and elicit others. Furthermore, Schmitt suggests that it is easier to align real-world skills, such as reading and writing, to such terms. For example, when reading, a learner needs to tap into meaning recognition aspects of word knowledge, especially. On the other hand, a test that requires form recall finds more restricted real-world functions, such as looking up words in a thesaurus (Schmitt, 2010: 88).

Therefore, while the four tests discussed in the sections that follow below do have different test designs and therefore will tend to tap into different types of word knowledge to a degree, it can also be asserted that simply because a test item is of a meaning recognition design, for example, that it does not mean that a taker of the test would not perform well on a form recognition test item of the same word(s). Nevertheless, as stated in Schmitt (2010), any claims of vocabulary 'learning' always need to be qualified in the light of the four distinctions drawn and, ideally, both receptive and productive vocabulary measures should be employed when possible (p. 89).

Furthermore, it can be said that the four tests presented and discussed below assess breadth rather than depth of vocabulary knowledge (discussed in Chapter 2). Nonetheless, much as in the Webb (2008) research discussed earlier, the distinctness of the two types is perhaps less defined than once believed. Vermeer (2001), for example, tested Dutch L1 and L2 children on both breadth and depth of word knowledge; vocabulary depth was tested using a word-association task, and for the receptive task words were spoken and the child was required to point to the correct picture. Group correlations (Dutch L1 and L2) between the breadth and depth measures revealed that "measuring breadth matches up very much to measuring depth: if one knows more words, one can describe a stimulus word in greater depth"

(p. 9). In addition, those correlations – though smaller – remained significant even when Vermeer examined the L2 data alone. Vermeer concludes that these correlations are sensitive to frequency (i.e. the more common the word, the stronger the breadth-depth correlation), and this was true irrespective of Dutch being the first or second language.

Naturally, it would not be sensible to generalize on the basis of this one study, particularly when the participants were young children and not adults. However, as in the Webb (2008) study, what the Vermeer study suggests is that a measure of vocabulary breadth is also likely to have some correlation with depth of knowledge as well.

5.1.1.1 The Vocabulary Levels Test

One of the earliest and most cited attempts at discerning vocabulary thresholds through a test was the Vocabulary Levels Test (Nation, 1983). Nation, largely influenced by the work of researchers before him who had produced wordlists that included data related to frequency of occurrence, believed that the testing of vocabulary could be approached the same way, broken down into 'levels' of frequency. The basic principle behind the notion of the usefulness of focusing on the commonest words first in pedagogy is related to Zipf's Law (described in Chapter 2), which states that the frequency of any given word is usually inversely proportional to its rank in a frequency list (see, for example, Table 5.1 below). Realizing this, Nation sampled, combined and cross-checked from three main sources to construct his first version of the Vocabulary Levels Test (henceforth, VLT): the Thorndike and Lorge (1944) list of 30,000 words, the Kučera and Francis (1967)/Francis and Kučera (1982) computerized lists, and the General Service List (West, 1953). Nation then divided the data into five levels of frequency: 2000 words, 3000 words, 5000 words and 10,000 words.

Table 5.1 Frequency band and percentage of text coverage based on the Brown corpus (Francis & Kučera, 1982)

Frequency level	Cumulative coverage (%)	Coverage (%)
1st 1000	72.0	72.0
2nd 1000	79.7	7.7
3rd 1000	84.0	4.3
4th 1000	86.8	2.8
5th 1000	88.7	1.9
6th 1000	89.9	1.2

For Nation, only words at the 2000-3000 level were worth spending class time on since those words are what West (1954) called 'essential' words (p. 122), the ones without which one cannot function effectively in, accounting for somewhat over 80% of running words of most texts (Nation, 1983: 14). Nation determined that 5000 words represented a kind of benchmark; to cross this 5000-word threshold meant being able to go beyond general words one needs to merely function in English and enter into more specialized lexical fields (Nation, 1990). The 10,000 level merely represented a threshold of 'low frequency words' (Nation, 1983: 14).

The fifth level tested words from the University World List (Xue & Nation, 1984), which in turn was informed by the Campion and Elley (1971) Academic Vocabulary List[44] and the American University Word List[45] (Praninskas, 1972). Since the Campion and Elley list excluded the top 5,000 words of the 30,000-word Thorndike and Lorge list, Nation placed the section of the test that measured candidates' knowledge of the academic words after the 5000 word level (Nation, 1983).

---

[44] A list designed for overseas students in New Zealand, constructed using a corpus of 301,800 words based on published material (i.e. textbooks, academic journal papers) and a selection of university examination papers, covering the most important academic disciplines at New Zealand universities in the 1970s (Campion & Elley, 1971).

[45] Compiled using an academic-textbook-derived corpus of 272,466 words, aimed for use by non-native speakers at American universities. The list excluded words from the General Service List (West, 1953).

Nation adopted a kind of multiple-choice/matching format for the test requiring no production of the word form by the test-taker, also conforming to the 'form recognition' item type according to the Schmitt (2010) taxonomy:

1. business	
2. clock	_____ part of a house
3. horse	_____ animal with four legs
4. pencil	_____ something used for writing
5. shoe	
6. wall	

(Nation, 1983: 19)

Each level (i.e. 2000, 3000, etc) contained six such items, for a total of 18 words targeted for assessment. (Nation claims, however, that since the distractors are also words, that really 36 words in all are being tested at each level[46] (Nation, 1983: 14)). The multiple matching format was chosen for its marking facility, ability to test many words in a relatively short period of time, and its sensitivity to even limited amounts of knowledge about each word. In addition to using frequency data, Nation also chose to prioritize nouns in the test, in a '3:2:1' ratio (Beglar & Hunt, 1999: 139) of nouns:verbs:adjectives (though this weighting was not uniformly consistent throughout the test), reflecting their relative importance in the source lists.

The Nation (1983) version of the test gained widespread use and acceptance over the years (particularly after its republication in a popular book dedicated to the subject of teaching and learning vocabulary (Nation, 1990)) , eventually becoming what Meara (1996) called the "nearest thing we have to a standard test in vocabulary" (p. 38). It took a number of years from the publication of the original test for researchers to begin to look at it more critically.

[46] This assertion was based on observation of candidates as they took the test in its piloting phase. However, as noted by Read (1988) and later confirmed by Schmitt, Schmitt and Clapham (2001), the test-takers looked carefully at all the distractors only when they were not sure of the word(s), otherwise the distractors were largely ignored (p. 18).

For example, Read (1988) questioned whether the test really discriminated between stronger and weaker candidates. His analysis showed that although the scores the VLT produced were scalable (i.e. a perfect score on a higher level meant that a candidate had mastered its preceding levels), there was a problem with the validity at the 5000 word level and its relationship to the university word level. Specifically, Read found that learners who had been exposed to a more academic register (due, for example, to study at university) did well on the university word level, and therefore its placement after the 5000 word-level – and its implicit assumption that mastery of it also meant mastery of the 5000 word-level – was questionable. Moreover, since a number of academic words are Latin cognates, L1 speakers of Romance languages provided a confounding variable to the interpretability of the results of that section of the test. Those shortcomings notwithstanding, Read acknowledges that the test 'has proved to be a very useful tool for diagnostic purposes' (p. 18) at Read's institution (Victoria University of Wellington).

The test's use and dissemination continued in its original form without much further scrutiny for another ten years until Beglar and Hunt (1999) explored the validity of the 2,000 word level (chosen for that level's relative importance in reading comprehension (Nation, 2006; O'Keefe et al., 2007)). In their study, Beglar and Hunt used four versions of the VLT: the original one ('Version A'), and three others (versions 'B', 'C' and 'D') which had been written by Schmitt in 1993 (Schmitt, Schmitt & Clapham, 2001: 57). All four versions of the test were administered to Japanese high school students (n=496) and the researchers ran repeated measures ANOVAs to explore any variation among means from the different tests. Beglar and Hunt found that there was indeed a significant disparity among versions ($p < 0.004$), with the difference between the means of Versions C and D, for example, being more than 2 whole points (out of 18 possible), with Version C having a much higher standard deviation (4.52 versus 3.15) (Beglar & Hunt, 1999: 137).

In order to help determine how best to revise the tests, Beglar and Hunt also explored the reliability of the individual test items (total 72 – the sum of all fours versions) by running item-total correlations (correlation between individual item dichotomous [0 or 1] scores and the total score on the test), in addition to the between each item and all other items (Beglar & Hunt, 1999: 138). The authors found that there was a total of 12 items that needed to revised or eliminated due to their relatively low item-total correlations (under 0.30). Moreover, Beglar and Hunt also concluded that there were not enough items classifiable as 'good' (with correlations of 0.40 or higher) (p. 139). Therefore, the authors decided to not only revise or otherwise change the 12 items with the lowest correlations, but also to extend the test from the original 18 items, to 27 (adding three more sets of three tested words) by shifting items around that had more desirable correlation coefficients. This alteration rendered two new versions of the VLT (at the 2,000 word level) with very high levels of internal consistency ($\alpha = 0.90$), with the two versions combined rendering a coefficient of 0.95. Following a re-trialling, the authors conclude that the two new versions are superior to the original versions for being more reliable and homogeneous.

Schmitt (Schmitt, Schmitt & Clapham, 2001), the author of the versions of the original VLT that Beglar and Hunt adapted, decided to expand on Beglar and Hunt's validation study, extending it to the other levels (i.e. not just the 2,000 word level), including a revision of the University word level test using the more up-to-date and balanced Academic Word List (Coxhead, 1998, 2000). Like Beglar and Hunt, Schmitt et al. analyzed the reliability of individual items (or 'clusters' of three items) and discarded items that had skewed or otherwise unbalanced responses, and like Beglar and Hunt, Schmitt et al. also found that it was necessary to increase the number of items. The researchers therefore pooled the best performing clusters from the four versions and merged them into two new versions ('E' and 'F'), adding one more cluster to each level, increasing the total number of assessed words from the Beglar and Hunt (1999) 27 items to an even 30. Schmitt et al. then administered counterbalanced versions of the two tests to a substantial group (n=801) of learners of English from a variety different countries and both Indo-European and non-Indo-European L1s[47].

As in the Read (1988) study, Schmitt et al. found that their new versions discriminated well between stronger and weaker levels of proficiency, with scalability coefficients of Versions 1 and 2 at 0.971 and 0.978, respectively. The new versions were also found to be highly reliable, with Chronbach's alphas all above 0.90 for all levels of both versions. However, perhaps the most novel and valuable contribution to the validation of the VLT in the Schmitt et al. study is the added qualitative element, in the form of post-test interviews.

22 test-takers were interviewed immediately following the test to explore whether the answers given on the paper test would correspond with their verbalized knowledge of a stratified sampling of 50 words from the test. The researchers found a reasonably strong correlation between the answers supplied on the written test and those given in the oral interview (0.749, $p < 0.001$). However, since a perfect correlation (i.e. 1.00) was not found, the researchers explored the data further, and found that around 6% of the participants interviewed did not know a word that they had answered correctly on the test. It was further discovered that a portion of that 6% did not guess wildly, but rather on the basis of partial word knowledge. The authors conclude that these data provide further evidence that guessing is not a major issue in the results of the test, and that the instrument does indeed have a good degree of validity and reliability.

Not discussed, but indirectly alluded to, in the Schmitt et al. interview validation exercise is the fact that around 90% of the 22 interviewees were classified as 'knowing' a word through a verbalized definition. Depending on the extent of this verbalization (not discussed in the Schmitt et al. paper), there is a likelihood that this explanation of the words would also reveal a depth of word knowledge. If this is the

[47] It should be noted, however, that about 40% (n=322) of the participants were from Spain.

case, it would appear that the breadth of vocabulary knowledge demonstrated in the VLT also relates, at least to some degree, to depth of knowledge of those words (cf. Vermeer, 2001).

5.1.1.2 The Eurocentres Vocabulary Size 'Yes/No' Test

The Eurocentres Vocabulary Size Test (EVST) was commissioned by the Eurocentres Group and developed by Meara and Jones (1988) in order to make that institution's process of determining students' levels more efficient. Until the late 1980s, the Eurocentres network of private language schools used to place newcomer students on their short courses by means of an integrative skills test, and with sometimes weekly intakes of new students, a faster, less labour-intensive system was required. What Meara and Jones eventually developed between 1986 and 1987 was a test that was able to be taken in ten minutes, and that required no human intervention whatsoever.

The researchers achieved this by constructing a test which, like the VLT assesses knowledge of frequency bands of 1000 words[48], however does so by simply asking the test-taker to indicate if s/he recognizes the word and its meaning. Such a format easily lends itself to a computer interface – words appear on the screen, and the candidate simply presses a 'Yes' or 'No' button. For this reason, this type of test has come to be known as a 'Yes/No' test – particularly in its electronic format – and, in paper-based form, a 'checklist' test. Therefore, unlike the other tests described in this chapter, there is no direct check on task-taker understanding of the words tested, but rather relies on the 'honesty' of the person being tested. However, in order to control for guessing, Meara and Jones included non-words, or what they call "imaginary words" (p. 81), which would calibrate the test-takers' final scores depending on whether or not they indicated that they knew them or not.

[48] Using the Thorndike and Lorge (1944) list.

The mathematical algorithm employed by Meara and Jones actually has its origins in Signal Detection Theory, used by the Navy in the 1950s in order to help detect enemy submarines (Anderson & Freebody, 1983):

$$\text{True } h = \frac{\text{h - f}}{\text{1 - f}}$$

where h = 'hit rate' (actual words selected as known), and f = 'false alarm rate' (imaginary words selected as known).

Although the EVST was eventually implemented and used throughout the Eurocentres network with success, there has since been more critical analysis of its validity and reliability, centred especially around the non-words used in the test, and also metacognition (i.e. test-taker attitude and/or strategies when taking the test).

With respect to the non-words, Meara and Jones (1988) report that they "carefully constructed" (p. 85) to fit phonotactically into the set they share. However, the authors also report test-takers in the field testing – even native speakers – taking a "long time" with some of the imaginary words (p. 86). As discussed in Eyckmans et al. (2007: 61), learners who take tests in a Yes/No format may actually reflect learner attitude (e.g. self-confidence) rather than proficiency, which in turn could undermine the validity of the test:

> The Yes/No task makes a strong appeal to learners' self-rating of language skills that it tempts some participants to overestimate their knowledge, and this is penalized in the Yes/No testing system. (p. 62)

Therefore, one of the mechanisms designed to help increase the reliability and validity of the test could actually in some cases have the opposite effect, as "the validity of a test can be defined as the degree to which test scores accurately reflect the test-takers' various abilities" (Eyckmans et al., 2007: 61), and if what is being assessed is metacognition in addition to language ability (see further discussion of validity in Section 5.1.2), then that validity is debilitated (Beeckmans et al.,, 2001; Eyckmans, 2000; Eyckmans, 2004). However, there is evidence that there is perhaps no need for the non-word control at all. Shillaw (1996) administered a series of pen-and-paper tests to Japanese university students learning English and analyzed the scores using the Rasch Model[49], and produced some interesting results. Shillaw showed that a carefully-designed checklist (i.e. Yes/No) test with good items (e.g. items that discriminate well between stronger and weaker candidates) produces a high enough measure of reliability that non-words could in fact be dispensed with. In addition, the Rasch analysis also provides a way of identifying those learners mentioned in the Eyckmans et al. research (those who tend to overestimate their vocabulary knowledge) as such pattern anomalies show up in the Rasch model, particularly in three-parameter item response theory (Fulcher & Davidson, 2007).

5.1.1.3 A Vocabulary Size Test of Controlled Productive Ability

Thus far in this chapter, the vocabulary tests presented have the primary intention of assessing passive recognition of word forms/meanings. Laufer and Nation (1999) sought to use one of those – the Vocabulary Levels Test – as a basis for a test of 'controlled productive ability', that is, a test that would also measure one's ability recall the form of a word:

[49] Rasch is a statistical model applied in testing IRT (item response theory). Rasch assumes that responses to test items are a function of both the difficulty of the test item and the ability of the test taker, and therefore makes a calculation of difficulty on the basis of those two variables (e.g. if the test taker's ability and the item difficulty are the same, there will be a 50% probability of achieving a correct response).

> We use the term 'controlled productive ability' for the ability to use a word when compelled to do so by a teacher or researcher, whether in an unconstrained context such as a sentence-writing task, or in a constrained context such as a fill-in task where a sentence context is provided and the missing target word has been supplied. (Laufer & Nation, 1999: 37)

Therefore, as opposed to the VLT which supplies both the target word form and its meaning, requiring the test-taker to match the two, the Productive VLT (Productive Vocabulary Levels Test) provides a context and part of the form as a prompt for active recall of the whole word, as in the following example:

The book covers a series of isolated epis____ from history.

In theory, the test-taker who knows the word *episode* should be able to then produce it on the basis of the context, and not produce other words that might work instead (*incidents*, *chapters*, *events*, etc.) because of the four letters (*epis*...) provided.

Like the VLT, the Productive VLT assesses knowledge of the top 2000, 3000, 5000, and 10,000 word level, including the University Word List.

Although the test was underwent a validation exercise by Laufer and Nation, there may be some issues with the validity of the Productive VLT, in other words "the degree to which the inferences drawn from test scores to test-taker abilities are sound" (Fulcher, 2010: 324). Consider again the example provided above – an actual item on the Productive VLT:

The book covers a series of isolated epis____ from history.

To what extent is the word *isolated* as a collocate of *episode* facilitating the word 'episode' in that item? If it is helping in recall, that in itself is not necessarily a problem, but if that facilitation does not occur throughout the test, then the difficulty of the test and even the construct being tested is not uniform. A look at other items from the Productive VLT provides further examples of this potential problem. Consider the difference between the items listed below (from the 2000 word level, Laufer & Nation, 1999: 46):

productive paradigmatic knowledge (?)	syntagmatic relationship knowledge (?)
8. The rich man died and left all his we____ to his son. *9. Pup____ must hand in their papers by the end of the week.*	*3. Every working person must pay income t____.* *7. He takes cr____ and sugar in his coffee.*

In examples 8 and 9 above, there is no particular word with which the target words necessarily collocate, and therefore it could be argued that recall from long term memory involves *paradigmatic* word relations (i.e. words that could substitute for one another in that slot). In number 8, for example, the word *wealth* could easily be substituted by *money, inheritance* or *riches*. Likewise, in number 9, a read of the rest of the sentence could render *students*, or even *employees* – if the prompts were not given. By contrast, no further co-text need be heeded beyond *income t____* (in reality a compound noun) and *cr____ and sugar* (a common binomial) in order to complete those items (3 and 7, respectively)[50]. Those items have strong collocational associations with each other, and therefore can be argued to involve a more *syntagmatic* mechanism of recall.

This is not to say that that the Productive VLT should be completely discarded as a useful test of vocabulary. As argued in Schmitt (2010), in order explore the validity (and value) of the test, a concurrent validity exercise involving free written production of the words targeted for assessment could go a long way in determining whether or not the Productive VLT actually tests one's ability to

[50] Moreover, it should also be noted that the number of letters in the prompts also vary, which in itself is likely to affect test-taker response behavior.

produce the lexis it purports to. Ways in which the PHRASE List might be incorporated into this format will be discussed further on in the present chapter.

5.1.1.4 The Vocabulary Size Test (Nation & Beglar, 2007)

According to Nation and Beglar (2007: 9), "the Vocabulary Size Test was developed to provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1000 to the 14th 1000 word families of English." Therefore, like all four tests discussed so far in this chapter, the Vocabulary Size Test (VST) uses words sampled at 1000-word bands – in this case from the 10-million-word spoken portion of the BNC – in order provide an overall of one's vocabulary size. In the case of the VST, however, the frequency is extended 4000 words beyond any of the other tests.

Nation and Beglar contend that the VST is closer to a test of proficiency in vocabulary than the VLT or Yes/No tests, which they argue are diagnostic measures. Nation and Beglar base this assertion on a number of features of the VST, including its coverage of each frequency band (K1-K14), as opposed to only a sample of four (K2, K3, K5 and K10), and also the VST test format. Unlike the VLT with its form-recognition matching format, and Yes/No (or checklist) tests which are more of a self-assessment, the Nation and Beglar VST puts the target word in a non-defining context (i.e. one cannot derive its meaning via context) and requires the test-taker to choose from among distractors from the same frequency band in a *meaning*-recognition format (Schmitt, 2010), as shown below (from the 2000 word level):

3.	upset: I am **upset.**	
	a.	tired
	b.	famous
	c.	rich
	d.	unhappy

As can be seen in the above example, the context 'I am' does not offer any help to one would employ a guessing strategy, yet does provide part of speech information. The authors cite research on TOEFL vocabulary item writing by Henning (1991) in support of this format, which showed that such non-defining contextualization added validity to the measure (e.g. by not causing unnecessary task difficulty by presenting the word in isolation, arguably an abnormal condition under which to decode a word). Moreover, Nation and Beglar argue that since there are distractors, and that those distractors usually share elements of meaning, the VST is more challenging than either the VLT or Eurocentres tests, and therefore the test taker must also have a more developed knowledge of the words tested in order to accurately discriminate from among the options.

Following its release, the VST also underwent a validation exercise by Beglar using Rasch analysis (Beglar, 2009). The findings showed that most items on the test displayed good fit and a high degree of unidimensionality (i.e. testing of a uniform construct), with the Rasch model accounting for 85.6% of the variance. Beglar therefore concludes that test is a reliable and valid instrument to measure vocabulary size.

However, in addition to the lack of multiword items in the VST, there is a potential issue regarding the choice of corpora used to inform the content of the test. Nation and Beglar (2007) explain that the spoken section of the BNC alone was used (and not the full 100 million words) because of issues related to register, especially. For example, items which one might consider 'common', like *cat, hello* and *sun* only appear in the 4000 word level in the full BNC, but feature much more prominently when including spoken data alone. Likewise, some words which might be considered more formal, such as *civil* and *commission*, occur in the top 1000 words if considering both written and spoken. However, it is likely that when dispersion data are considered (i.e. how 'spread out' across the data files the items occur), some of the more problematic or counter-intuitive frequency rankings could have been resolved.

Instead, with the inclusion of spoken data alone, it is possible that a number of important items that recur frequently in the written mode especially get relegated to positions in the ranking that do not reflect their natural Zipfian distribution – if they feature at all. In fact, the decision to exclude 90 million words of written data proved of questionable value in the end, since the authors report that the sacrifice did result in some changes in ordering of certain items, but "the changes were not large" in the end (p. 10). Finally, since the VST "is a measure of written receptive vocabulary size" (Nation & Beglar, 2007: 11), the exclusion or written data may actually threaten its validity to some extent. More research is necessary in order to explore these potential threats to validity further.

Issues of validity regarding all of the tests presented in this chapter up to this point will be discussed in the next section.

5.1.2 Issues of validity in vocabulary tests

In early (i.e. pre-1990s) test theory, validity had evolved to be thought of as a concept comprising a number of different types of validity, such as 'criterion-oriented validity' (itself composed of predictive and concurrent validity), 'content validity', and 'construct validity' (Chronbach & Meehl, 1955)[51]. However, in more recent years greater attention has been paid to what inferences can really be drawn from the results of a test (e.g. scores). Thus, validity today is generally not thought of as a property of a particular test, but the degree to which the inferences that are made from test scores/results are actually meaningful and justifiable (Messick, 1995: 742; Weir, 2005: 12). So, although validity is still

[51] Roughly, 'criterion-oriented validity' is essentially concerned with the extent to which what is being tested makes a valid statement about an ability during or after the exam; 'content validity' relates to whether or not what is being tested is actually representative of the area of language it purports to represent; and 'construct validity' is about a broad conceptualization of the assessment as a whole, and whether 'vocabulary' (for example) is in fact a measurable concept (Fulcher & Davidson, 2007).

thought of as multi-faceted (Fulcher, 2010: 20), the prevailing thinking is toward validity as a more unitary concept, embodied by the term construct validity (Figure 5.2):

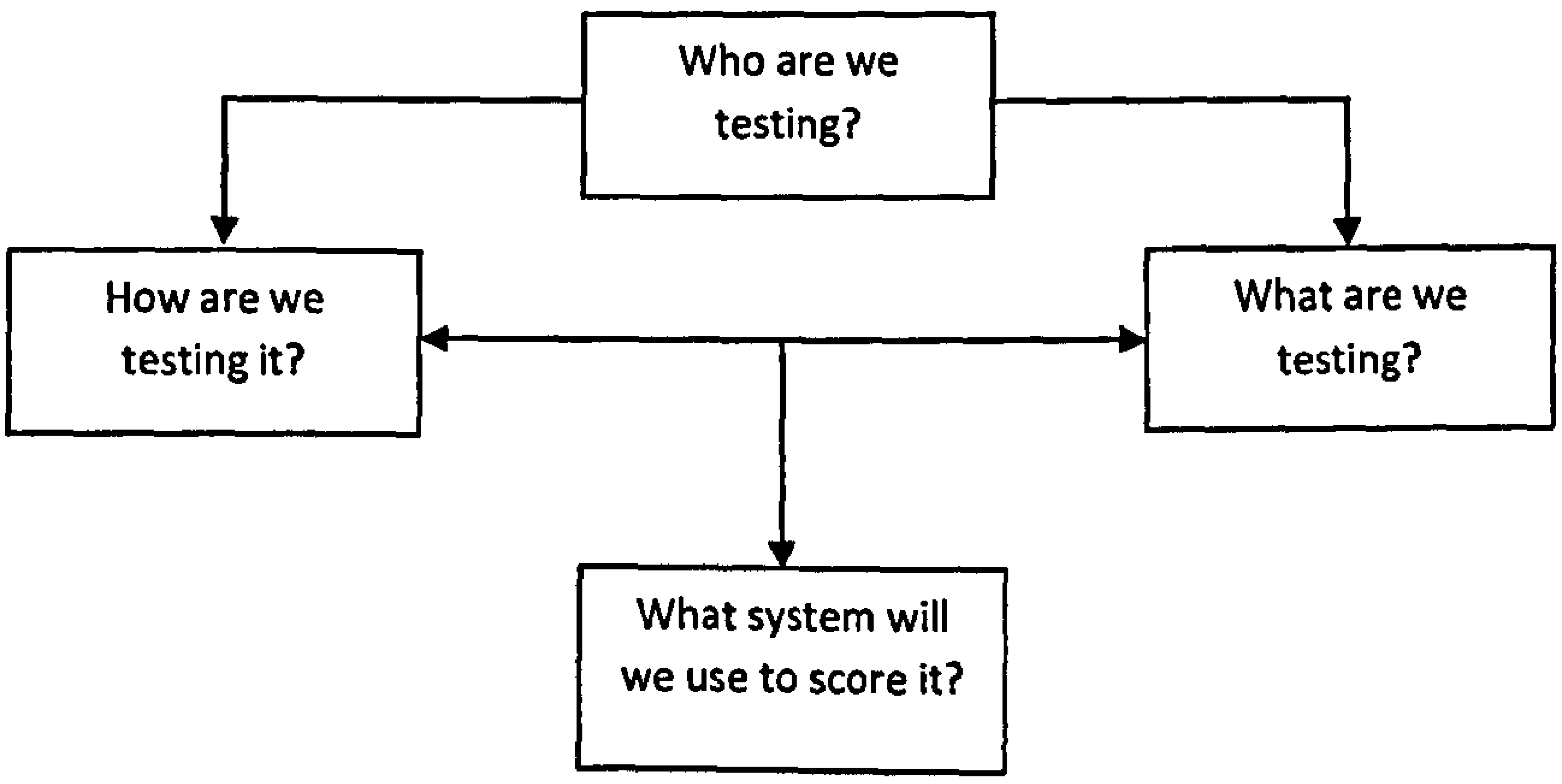


Figure 5.2. The 'core' of construct validity (O'Sullivan & Weir, 2011: 23)

As cautioned in O'Sullivan and Weir (2011), unless the questions in Figure 5.2 can be answered satisfactorily, a test is "unlikely to allow us to make valid inferences about the candidature" (p. 22).

In the specific case of the types of vocabulary tests thus far discussed in the present chapter, there can be a number of threats to that construct validity. For example, although vocabulary knowledge has often been associated with certain integrated skills abilities (e.g. reading, listening), the consequential validity of vocabulary test scores would possibly be threatened if those scores were used to make strong claims about how well (or poorly) the test taker will actually perform in those skills (cf. Qian, 2002; Shiotsu & Weir, 2007). Vocabulary is usually read or listened to in some kind of context, but tests like the Yes/No Test and the VLT present vocabulary in isolation, which may threaten the context validity of those tests (Weir, 2005). And validity may be affected by non-linguistic factors, such as strategy and test-taking behavior (i.e. metacognition). If, for example, words that a vocabulary test indicated were not known by a candidate were then actually discovered to be known in a retrospective protocol

(Anglin, 1993), that would be an indication of an issue with the construct validity of that test. However, perhaps one of the principal threats to the validity of the tests so far presented here may have to do with their content validity, or the words actually included on the tests.

Tests like the VLT are derived from frequency lists of words, themselves derived from corpora that are bound by elements such as the time period in which they were compiled and the type of discourse from which the data were sourced. Further, as discussed in some detail in Chapter 3, lists of words may themselves have a validity threat associated with them, namely misleading frequency counts. As explained in Chapter 3 (the example of *look* versus *look for*), many of the words that feature in the top 2000 word families may need to have their frequencies adjusted to account for the proportion of the counts that are actually due to multiword expressions containing that word. The word *time* is tested at the first 1000 level on the VST, for example, but *time* is also one of the words that most features within the expressions in the PHRASE List, and it is questionable to argue that getting the word *time* right on that test also means knowing phrasal expressions like *over time*, *at times* and *time and again*. Therefore, the PHRASE List is indicative of a validity issue with existing vocabulary tests: if data show that around ten percent of the top 5000 words should be inclusive of phrasal expressions (see Chapter 3), then any claims regarding the meaning of scores on tests that do not account for phraseology should at least be qualified.

Regarding validity, Robert Lado once wrote, "Does a test measure what it is supposed to measure? If it does, it is valid" (Lado, 1961: 321). In the case of discrete vocabulary tests, perhaps the most honest answer to Lado's question would be, 'well, mostly'. Ways in which that answer might change to something closer to a 'yes' will be discussed in the following section.

5.2 Development of a test of phrasal expressions

Since the PHRASE List was compiled using frequency information and divided into 1000-word frequency bands, the list in theory should be able to be integrated into any one of the tests described in the previous section. Take, for example, the phrasal expression *take place*, and how it might be incorporated into the various item format types:

Vocabulary Levels Test

1. take place	
2. have got to	____ do
3. seek to	____ try
4. fail to	____ happen
5. make sure	
6. carry out	

Eurocentres (Yes/No)

1 ☐ take place 2 ☐ seek to 3 ☐ see in for 4 ☐ fail to 5 ☐ make sure

Productive VLT

The wedding will t____ pl____ at a lovely church near the beach.

VST

1. take place: It did not **take place**.

   a. stop
   b. steal
   c. leave
   d. happen

As seen in the examples above, the phrase take place can fairly easily be incorporated into the various items types; however, each may have its own drawbacks and advantages. In the VLT cluster format, the efficiency of the form-recognition may not be the same as in its one-word counterpart: it is unknown whether the decontextualization of phrases would present any more of a challenge than just single word forms. The Eurocentres test presents a similar question, in addition to the challenge of coming up with a 'pseudo-phrase' for each item (as in 'see in for' in the example), and the validity issues that may raise. The Productive VLT format perhaps present the greatest validity threat, as it would require the production of at least two words, two (or more) words that are part of a phrase that may or may not have representation in the mental lexicon as two or more separate morphemes – the cognitive processing differential and what effect it may cause on test performance may be a major issue. Finally, the VST format seems to lend itself well to the phrasal expression lexical item type, as at least there is surrounding linguistic context, but there is the challenge of having to come up with four multiple choice options for each phrase. Taking into account all of the aforementioned features, it was decided that the VLT and VST formats would make the best candidates for the initial test development prototyping.

However, it can also be said that there are actually two approaches that can be taken to the inclusion of phrasal items in a vocabulary test: *integrative* and *additive*. An integrative approach would simply involve inserting phrasal expressions into existing vocabulary tests. For example, the current VLT (Schmitt, Schmitt & Clapham, 2001) contains 10 clusters per frequency band, and each cluster tests three words. One cluster that tests knowledge of three phrases could be added justifiably into each band (K2, K3, and K5), representing 9% of the total words tested at each band (3÷33). Likewise, the VST tests ten words at each frequency band, and therefore one phrasal expression per frequency band could be included in that test. Although there are good reasons for taking an integrative approach to phrasal inclusion (e.g. positive washback, the implicit message that vocabulary is more than just individual words, etc.), the items from the PHRASE List might best be included in existing vocabulary tests in an additive fashion.

The additive approach is, of course, not without precedent. The University Word List and the Academic Word Lists have consistently been included in tests like the VLT as a kind of add-on level – though still generally considered of great importance. The PHRASE List could be incorporated into existing tests in a similar fashion, and such an approach can offer a number of advantages. First of all, it is questionable the impact and power a single item added onto a test would have. (If a candidate shows knowledge of a single formulaic sequence, what inferences about that person's phraseological abilities can really be drawn?) However, taking an additive approach, it is possible to include a special measure of phraseological competence that includes multiple items.

5.2.1 *Exploring the performance for two test formats*

As described in the previous section, a test that measures knowledge of phrasal expressions would ideally complement existing vocabulary tests, and both the VLT and VST were identified as good candidates in terms of test format. In order to determine which format was superior (if either), tests which adopted both formats were developed for comparison.

Both the VLT and VST sample from different word classes from frequency lists, generally representative in proportion to their relative presence in the lexicon (nouns, verbs and then adjectives). However, the PHRASE List is clearly a different type of frequency list, requiring its own grammatical analysis, summarized in Table 5.2 below.

Table 5.2 Grammatical analysis of PHRASE List items

Band	NP	VP	Adv.	Adj.	DET./PRO	Other
1K (k = 32)	0	7	23	0	1	1
2K (k = 85)	1	38	26	1	14	5
3K (k = 128)	2	45	63	3	12	3
4K (k =158)	4	38	97	3	12	4
5K (k = 102)	4	32	56	2	5	3
Total = 505	**11**	**160**	**265**	**9**	**44**	**16**
Cum. %	2.17%	31.68%	52.47%	1.78%	8.71%	3.16%

As can be seen in Table 5.2, it was found that the items in the PHRASE List could be broken down grammatically into noun phrases or *NP* (*point of view, well being*), verb phrases or *VP* (*catch up, let alone*), adverbial phrases (*along with, on the way*), adjective phrases (*the odd, key to*), determiner/pronoun phrases (*the following, each other*), and then miscellaneous category that included interjections and other less frequent items (*oh dear, that is*). Clearly, adverbial phrases dominate the grammatical categorization, representing over 52%, followed by verb phrases at over 31%. Taken together, the verbal and adverbial items in the PHRASE List represent over 84% of all phrases to the 5K level, and therefore it was from these two functional categories that phrases were selected for possible incorporation into vocabulary test items[52]. The items, once written, then underwent a moderation phase (Hughes, 2003: 63) in which two colleagues scrutinized the test for any items that needed rewriting or even rejection.

[52] It is worth noting that in calculating the grammatical categories of the items, it was their function rather than superficial formal features which were taken into account. For example, the phrase *to blame* is a verb phrase on the surface, but functionally speaking is actually an adjectival form synonymous with *responsible* (*the weather is to blame = the weather is responsible*). Likewise, the phrase *this stage* appears to be a noun phrase, but since it basically means 'now', can be classified as a time adverbial (*I thought I'd be done by this stage = I thought I'd be done by now*).

As suggested in Hughes (2003: 63-64), it is important to trial any new instrument on a cohort of native speakers. Clearly, if a significant number of native speakers have difficulty with the test and/or test items, there is then compelling evidence that the test/items should be carefully revised before administering the same instrument to non-native speakers. As further suggested in Hughes (2003: 64), it is not imperative that native-speaker test-takers of a trial version of a pilot test be in any way linguistic experts; in fact, it is preferable that they not be in order to not bias the results. Therefore, native speakers (n = 10) were recruited to try both test versions who were familiar to the researcher but known not to have any formal training in any linguistically-related academic discipline. These participants (5 men and 5 women) all had at least completed a secondary education in the United States, were all speakers of American English, and ranged in age from 16 to 39 (M = 28; SD = 14.56).

Three versions (A, B and C) of the Phrasal VLT (henceforth, PVLT) were developed for the initial trial (Appendix 9), and one long version of the Phrasal VST (henceforth, PVST) (Appendix 10)[53]. Participants were administered the test individually in their spare time, and committed to the trialling completely voluntarily. The purpose of the test was explained to participants, as well as the role of the native speaker participants in the test validation process. Participants were asked to first look at the example on the first page of the test (Appendix 9 and 10), and were asked if they understood and if they had any questions. It was always stressed to participants that in no way was this test assessing their vocabulary; on the contrary, it is assumed, the were told, that as educated adult native speakers they already know all the items on the test. All participants were therefore always assured that what was being tested was the test itself, and not to worry if they struggled at any point since this would be an indication of a problem with the instrument rather than a reflection of their knowledge or lack thereof. When ready, participants were requested to begin the test and to indicate (for example, by circling or underlining)

[53] Three versions of the PVLT were developed in order to have the phrases on both the PVLT and PVST roughly overlap

any item(s) that they found 'tricky' in any way or otherwise wished to query. They were also encouraged to make notes of their thoughts on the items or test as a whole on the script itself. The results of the native speaker trial are reported in Tables 5.3 and 5.4.

Table 5.3 Phrasal VLT native speaker pilot: central tendency (Max. poss. = 30)

Version	N	Minimum	Maximum	Mean	SD
A Score	10	24	29	27.60	1.42
B Score	10	27	30	29.30	1.06
C Score	10	27	29	28.60	0.84

Table 5.4 Phrasal VST native speaker pilot: central tendency (Max. poss. = 50)

N	Minimum	Maximum	Mean	SD
10	46	50	49.61	1.01

As should be expected from a test designed for non-native speakers that is taken by native speakers, the scores overall were fairly high, with no participant scoring less than 24 on any of the versions of the PVLT, or less than 46 on the PVST. However, it is clear that there were some problems with all versions. First of all, when the items that were answered incorrectly were reviewed with each participant, it was evident in each case that the participant knew the item and had simply answered wrong due to problems in the items. For example, five participants reported having difficulty associating the word 'if' with *whether or not* in the PVLT. However, a potentially more serious problem presented itself when participants were asked about their thought processes while taking the test. All participants reported that while in some cases the answers were fairly obvious in both test formats, with most items in the

PVLT they also found themselves putting the phrase in an example sentence in their heads; in other words, since the phrases are presented without any contextual support in the PVLT format, they felt they often needed to provide it in order to help them resolve the items.

As discussed earlier in the section that dealt with validity in vocabulary tests (Section 5.2), ideally tests should measure what they claim to measure, and scores should be meaningful. However, if the participants in the PVLT format pilot unanimously report that they were performing an additional operation – not provided for in the test – in order to resolve the items, then there is an element of metacognition, rather than the language itself, that is also being tested. If that is true, then other extraneous variables may affect the outcome of the test, such as the working memory of the participant, and his or her ability to come up with that strategy on his or her own when taking the test. Finally, the fact that participants reported needing to add context calls into question the 'cognitive validity' (Weir, 2005) of the test format (also related to 'context validity'), which is a property of "the relevance of the individual's test responses to the behaviour under consideration, rather than on the apparent relevant of the item content" (Anastasi, 1988: 131). Therefore, although it was believed at the design and prototyping phases of the PVLT that the phrasal expressions would naturally be testable in the same way as in the 'original' VLT, the change in lexical item – a phrasal expression with no context – did apparently cause an unexpected and unwelcome change in the response behaviour. O'Sullivan and Weir (2011) suggest that

> ...even small changes to parameters of context validity are likely to impact significantly on cognitive validity and subsequently on the score or grade a candidate receives on a test. (p. 28)

In order to confirm that this cognitive differential occurred among non-natives as well as native speakers, ten students in a private language school in Oxford (UK) agreed to participate in a small-scale study, and as compensation received cinema vouchers (Appendix 11). All had taken a placement test upon starting their courses, and were reported by the school to be of at least 'intermediate' proficiency. Each was at least 18 years of age (M = 24, SD = 5.60), three of whom were male. The L1s of the participants broke down as follows: Japanese (n = 4), Spanish (n = 3), German (n = 2), and Russian (n = 1). The same instruments used by the native speakers were used for the non-native trial, with a few of the items rewritten (those which had posed problems for the native speakers according to their post-test interviews). All participants were administered both tests in counterbalanced order (alternating Phrasal VLT first) to allow for the checking of any order effects. Participants were reminded of the purpose of the test, and told that they would receive their cinema vouchers following a post-test interview that should only take no more than ten minutes. They were then asked to look at the example items on both tests (Phrasal VLT and VST), and allowed to ask any questions. Candidates completed the tests at their own pace, and once finished, were invited into a separate room for their interview. Students were first asked which version of the test they had taken (Phrasal VLT first or Phrasal VST first), and then asked how they found both tests. If the examinees did not volunteer the answer, they were also asked specifically which test (Phrasal VLT or Phrasal VST) they preferred, and why. This phase of the post-test interview generally lasted no more than five minutes.

The second part interview consisted of a questionnaire comprised of 48 items (a sampling of 34 percent of a total of 140 items across all tests) that overlapped on both the VLT and VST formats. The aim of the questionnaire, following Schmitt, Schmitt and Clapham (2001), was to determine the extent to which the knowledge expressed on both tests reflected participants' actual knowledge. Participants were allowed to look at each phrase and define it aloud. If participants could not produce a phrase by simply looking at the lexical item without context, they were allowed to look at the phrase in a non-defining sentence (the same in the test). If the participant was able to provide an acceptable description, the phrase was marked with a value of '1' on the rater's score sheet (all rating was conducted by the researcher personally), and if not, then the item received a '0'. It was usually very clear to the rater when the item was known, usually because a close synonym was provided or a good contextualizing sentence was given. On occasion, a participant would offer a sentence whose context did not satisfactorily provide evidence of knowledge, in which case the researcher elicited a second defining sentence.

All candidates, irrespective of the order in which they took the tests, indicated that they found both tests challenging, but that they preferred the Phrasal VST. When asked why, the most common comment reflected what the native speakers had reported in the first Phrasal VLT trial, that when taking the test in the VLT (de-contextualized) format they would often not recognize the lexical item, or would have to try to put it in a sentence in their heads first. By contrast, no such extra cognitive processing was reported in the VST format.

The results of the item questionnaire of overlapping phrases tested on both test formats were matched against candidates' tests, looking specifically for any discrepancies between declared knowledge and knowledge demonstrated on the instruments (Appendix 12). The full tabulation of the comparison between both test formats is in Appendix 13, but a summary of the knowledge discrepancies (e.g. item shown to be known, but incorrect in Phrasal VLT) encountered can be found in Table 5.5. The analysis revealed that, indeed, the number of times the knowledge as expressed in the post-test interviews did not agree with the knowledge demonstrated in the test was significantly higher in the Phrasal VLT ($t = 5.439$, $p \leq 0.001$), with the difference also significant at each frequency band except for the first 1000 (Table 5.6). Out of a total of 480 total answers (48 overlapping items X 10 participants), participants answered incorrectly – even though they demonstrated knowledge of the item – a total of 77 times on the VLT test format (or 16.43% of the time). By comparison, test takers only showed such knowledge discrepancy on the VST format a total of 11 times, or a relatively modest 2.29%. Irrespective of whether or not both percentages can be considered acceptable, it is clear that the VST format consistently renders more favourable response behaviour.

Table 5.5 Knowledge discrepancies between PVLT and PVST forms (per individual participant)

Test	N	Participant Minimum	Participant Maximum	Mean	SD
PVLT Discrepancies	10	3	15	8.80	4.18
PVST Discrepancies	10	0	3	1.50	0.97

Table 5.6 Breakdown of knowledge discrepancies per frequency band

Frequency Band	PVLT	PVST
1K	3	4
2K	21	0
3K	19	2
4K	21	2
5K	13	3
Total (Max.= 480)	77	11

In Tables 5.5 and 5.6, an item was judged to be 'discrepant' when the post-interview declared knowledge (or lack of it) did not match with either the Phrasal VLT, Phrasal VST, or both. Out of the ten participants who took the test, there was not one case in which the verbal declared knowledge did not match at least one of the choices written on the test paper (either the PVLT or the PVST), indicating that students generally did not guess when they did not actually know the expression. Moreover, no significant interaction was found between *frequency* x *discrepancy* on either test, suggesting that the format itself – and not frequency or other factors of difficulty – may be the main explanatory variable for the differential performance.

Naturally, following a careful item analysis, it would likely be possible to decrease the incidence of knowledge discrepancies in the Phrasal VLT; however, this also holds true for the Phrasal VST, and so it seems clear that both the qualitative and quantitative data point to the same conclusion: the Phrasal VST format is more likely to produce a more valid instrument, providing a more reliable portrayal of receptive knowledge of phrasal expressions.

5.3 Field testing and validation

As discussed in Section 5.2 above, it was determined, through careful prototyping of both VLT and VST formats, that the VST format shows itself to consistently be a more accurate measure of receptive knowledge of the phrases. It was therefore decided that any further larger-scale piloting should occur with the VST format.

There were only ten non-native participants that took the Phrasal VST in the parallel trial and therefore any robust quantitative item analysis was not realistic; however, the instrument was subjected to a qualitative analysis on the basis of the parallel trial performance, and on that basis a number of items altered and/or rewritten For example, it was found that the wording of 'created an increase in' as a distractor for the phrase 'gave rise to' (item number 8, fourth 1000, Appendix 10) might have caused two of the candidates to get it wrong, so it was instead changed to 'increased the number of'. However, again, the main objective of the prototyping of the first form of the Phrasal VST was to examine whether there was differential cognitive processing evident in it vis-à-vis the Phrasal VLT, and whether therefore it was beneficial to proceed with the VST format instead for larger scale piloting. It would be the aim of a larger pilot to conduct more thorough analysis of the items.

Carol Spoettl of the University of Innsbruck was planning to conduct a country-wide field-testing of a proficiency test developed by that university, and offered to send out a pilot version of the Phrasal VST along with the regularly scheduled papers to be taken by the same candidates taking the university tests. Further detail regarding that piloting is described in the sections that follow.

5.3.1 Materials

The University of Innsbruck tests were going to be sent in a test booklet, and although the opportunity to use the large candidate pool that the university had recruited presented itself as a unique opportunity, it also presented a few challenges. In particular, due to graphic design and printing constraints, only three sheets of A4 could be included in the test booklets. The problem with limiting the test to only three sheets was that, since one main aim of test piloting is to determine if items are behaving as they should (e.g. discriminating between stronger and weaker candidates), a common protocol for writers of multiple choice items is to pilot more than are needed in order to arrive at the most usable ones (Hughes, 2003: 77). However, as can be seen in the prototype version of the Phrasal VST in Appendix 10, the full fifty-item test (10 items per frequency band x 5 levels) cannot realistically be shrunk to less than four A4 pages without compromising the legibility of the items (and therefore potentially creating an unnecessary confounding variable). Nonetheless, as stated in Fulcher (2010: 79), "(w)hen items are piloted it is not necessary that the test takers are presented with a complete test. In piloting it is only essential that sub-tests are used that generally resemble" the envisaged full test.

In coordination with Carol Spoettl, therefore, it was determined that three separate test booklet versions could be distributed, which would allow test sub-sections to be piloted that together would produce more items than necessary for one full test. The researcher therefore re-utilized items already prototyped in the earlier trials, and added five more per frequency band, for a total of 75 (50 [prototype PVST] + 25 [new]) multiple choice items (Appendix 14). These items, in turn, were subdivided and allocated to three test versions. Version A contained items 1 – 6, Version B 7 – 12, and Version C 1 – 3 (repeated[54]) plus 13-15.

Finally, the test was reformatted in order to be machine-readable, approved by the researcher, and allowed to be included in the University of Innsbruck test booklets.

5.3.2 *Participants and procedure*

A total of 2,204 candidates took the test, with 742 taking Version A, 731 Version B, and 730 Version C. These test-takers were all students starting in the Austrian university system, all at least 18 years of age, and reported by the University of Innsbruck to be at or around the B2 CEFR level of proficiency. Participants were assigned exam codes which they were instructed to write at the top of the test paper – the same code used for taking their other proficiency tests. Once all the tests were returned to the University of Innsbruck, they were run through a test scanner that reads the answers given on the test at a rate of 100 sheets per minute and produces a text file with all the responses coded.

Once the text file data were collected, the researcher entered them into SPSS, including which version of the test was taken, a breakdown of items by frequency band, which option was chosen for each multiple choice item, and then binary values assigned for choosing the key ('1') or a distractor ('0').

[54] It was originally intended that the repeated items could potentially be useful for common item linking (Bond & Fox, 2007), when conducting a Rasch analysis of the items, for example.

5.3.3 Results

The means and standard deviations for the total scores (sum of all frequency bands) on all three test versions are presented in Table 5.7. The breakdown of the individual means by band is shown in Table 5.8.

Table 5.7 Test performance - Innsbruck pilot

Test Version	N	Mean	SD
A	742	22.67	5.30
B	731	22.32	5.76
C	730	19.95	5.59

Table 5.8 Results of Innsbruck pilot by frequency band (Max. = 6 per frequency level)

Freq.	Version A	M	SD	Version B	M	SD	Version C	M	SD
1K		5.50	0.87		4.78	1.26		4.25	0.97
2K		5.05	1.20		5.17	1.14		4.65	1.41
3K		4.33	1.34		4.63	1.44		4.72	1.59
4K		4.21	1.65		3.52	1.62		4.01	1.56
5K		3.60	1.65		4.22	1.67		2.32	1.63

On the surface at least, it would appear that Version C was the most difficult version (Table 5.8), with a mean of just under 20 points (maximum of 30 possible). Further exploration of the nature of the difficulty on the test will be examined further later in Section 5.3.5. Otherwise, the means on the total scores of all versions are similar, and a one-way ANOVA shows that the difference between them not significant (i.e. not on individual band level nor total test score level).

Reliability estimates were also checked for each version (Table 5.9), but these are only relevant as a frame of reference for now, as the items that end up on the revised version will be those that that exhibit the best overall performance (operationalized in the following section), and therefore the reliability is likely to change.

Table 5.9 Reliability statistics for piloted versions

Version	Chronbach's Alpha
A	.869
B	.854
C	.879

5.3.4 Item analysis of Phrasal VST

There are a number of a ways to select items for further trialling once a test has been piloted, in part predicated on how performance on the test is meant to be interpreted. For example, if the information desired is where a particular test-taker places in relation to other candidates taking the same test (e.g. in the top 10 percent of the candidature), then that test can be called *norm-referenced* (Hughes, 2003: 20).

On the other hand, if what is desired is not knowing how an individual's performance compares with other test-takers, but rather the extent of the knowledge and/or language-related ability of that candidate, then the test can be considered *criterion-referenced* (Hughes, ibid.). In the former, one key criterion for selecting good items is their ability to discriminate between stronger and weaker candidates in order to produce scores that are distributed on a normal curve (Fulcher & Davidson, 2007). While discrimination is also important in criterion-referenced tests, what is perhaps more important is the degree to which the items can be said to be representative of the construct and therefore provide meaningful information about what the test-taker knows (Fulcher, 2010).

However, it is not entirely straightforward to categorize tests like the VLT and VST as purely norm-referenced *or* criterion-referenced:

> The Vocabulary Levels Test ... is currently widely used to determine whether learners need to focus on high frequency words, academic words, or low frequency words. It is a diagnostic test that looks at separate slices of a learner's vocabulary (by sampling from frequency bands). The Vocabulary Size Test has a different purpose. It is not a diagnostic measure like the Vocabulary Levels Test, but is a proficiency measure used to determine how much vocabulary learners know. (Nation & Beglar, 2007: 10)

Nation and Beglar make the preceding assertion based on the fact that while both tests sample from different frequency bands, the VLT does so only from the 2nd, 3rd, 5th and 10th 1000 bands (plus the AWL), and the VST "fills in the gaps" (ibid.), thus providing a better estimate of vocabulary size, according to the authors. The score of a test of vocabulary size, therefore, can be used "to determine an individual's standing in relation to other examinees (i.e., a norm-referenced interpretation)" (Beglar, 2009: 17), and also to assess "whether an individual as achieved specific abilities, levels or knowledge (i.e., criterion-referenced interpretations)" (ibid.).

On the other hand, it can also be argued that vocabulary learning is really item-based, with each vocabulary item "addressing a separate construct" (Schmitt, 2010: 185). Put another way, it would be questionable to claim that knowing any given lexical item implies knowing any other given lexical item. One needs to exercise caution, therefore, when asserting that a test like the VST is criterion-referenced as any claims of 'mastery' or 'non-mastery' (Bachman, 2004: 131) of a construct as broad as 'vocabulary' seems too broad a proverbial brushstroke. The best approach seems to be frequency sampling in terms of being able to draw extrapolations based on knowledge demonstrated on vocabulary tests (Schmitt, 2010: 185), which is precisely the case in the VST test format.

In summary, in the light of the fact that performance on the PVST can have both criterion and norm-referenced interpretations (as in the VST (Beglar, 2009)), the following features will be considered for item selection:

- **representativeness**[55] (*Does the phrase tested represent the construct?*)
- **difficulty** (*Is the item so easy as to be of limited value on the test?*)
- **validity** (*Is the item measuring, as much as possible, what it is intended to measure without evidence of extraneous or otherwise unintended linguistic or non-linguistic influences?*)
- **discrimination** (*Does the item show that it can separate masters from non-masters and/or stronger candidates from weaker candidates?*)

Traditional methods of determining discrimination, such as those used on norm-referenced tests using classical item analysis techniques, would therefore prove to only be of limited value when determining which items should ultimately appear on an interim final version of the PVST. For example, one way of determining the extent to which items discriminate between test takers is to run item-total score correlations, obtained by calculating a dichotomous item score (1 or 0) and its correlation with a total

[55] In the case of the PVST, care needs to be taken to include both adverbial and verb phrasal expressions, with a stronger representation from the former, reflecting the grammatical features of the items on the PHRASE List.

test score (Bachman, 2004: 129), also called a point-biserial correlation. Such correlations were executed on each of the PVST versions, and can be found in their entirety in Appendix 15. However, items with the highest correlations – those with greatest indication of discrimination (Fulcher, 2010: 183) – were tabulated for initial investigation (Table 5.10). This exercise revealed that relying on the correlation data alone would not be satisfactory in order arrive at a final test of the fifty best items.

Table 5.10 Highest (top ten) item-total score correlations by frequency band

Frequency band	Item-total correlation*	Version/Item number
K1	.538	B12
	.477	C3
	.467	C1
	.453	A4
	.444	A6
	.427	B10
	.425	C13
	.407	B8
	.398	B11
	.395	B7
K2	.561	C3
	.537	A6
	.529	B9
	.526	C15
	.522	C2
	.512	C13
	.472	B10
	.464	B12
	.445	B8
	.441	C14
K3	.658	C3
	.633	C13
	.598	C14
	.570	C1
	.586	B10
	.564	A4
	.563	C2
	.560	B9
	.549	B8
	.547	C15
K4	.665	C2
	.662	C15

	.650	B9
	.622	A2
	.618	B12
	.612	A6
	.596	A4
	.572	C3
	.552	C14
	.534	C1
	.505	B10
K5	.671	B10
	.661	B7
	.626	B9
	.625	B11
	.620	B12
	.585	A5
	.560	A6
	.501	C2
	.497	C1
	.497	C15

* As a general guideline, Fulcher (2010) suggests that "any value above .250 is acceptable" (p. 185), but the items in Table 5.10 are all well above that value.

Specifically, and of relevance to the present discussion regarding item selection, there were many details in the individual items in Table 5.10 that the broad stroke of the statistic seemed to miss. As can be seen in Table 5.10, the correlations seem to be generally revealing of a broad trend, tending to rise from the highest frequency bands (K1, K2) to the lowest frequency bands (K4, K5). A possible explanation for this is simply that the larger the test score variance, the larger the correlation (Bachman, 2004: 130), and one would naturally expect less variation in scores at the highest frequency levels, where items are more likely to be known by more candidates. In any case, closer inspection of the items with point-biserial correlation values lower than those in Table 5.10 showed that many actually discriminated well but, more importantly, provided insight into the thinking of the candidates when analyzed qualitatively. Since all items on the PHRASE List are intended carry some degree of semantic opacity and therefore possible misinterpretation, it is this last characteristic – related to *representativeness*, *validity* and *difficulty* – that ended up being a key decisive factor in item selection.

When calculating item difficulty, one focus in classical item analysis is the number of candidates that choose the key, which renders what is known as a *p*-value (proportion of dichotomously scaled items). The p-value is calculated as follows (Bachman, 2004: 125):

$p_i = R_T/N$

where R_T is the total number of test takers who answered the item correctly; and N is the total number of test takers. (In distractor analysis, the p-value is calculated using the same equation, but with each distractor instead of the item as a dichotomous whole.)

An example of the type of analysis that was carried out for the present study is shown in Table 5.11:

Table 5.11 Analysis of K1, B8 (or 'item 8 on Version B of the First 1000')

K1, Item B8 (item-total correlation (.407)

8.	at all: I don't like it **at all**.		Facility	Upper	Lower	D
	a.	all the time	.07	.01	.13	-.12
	b.	in any way	. 91	.99	.82	.07
	c.	first	.01	.00	.02	-.02
	d.	sometimes	.00	.00	.02	-.02
		No attempt		0 (0%)	4 (2%)	

In Table 5.11, the p_i-value, or the proportion of candidates that chose each distractor, is listed under the column 'Facility'. The item facility value is the number in the shaded line, reflecting the key of that item. In the case of K1, B8 in Table 5.11, the item facility value of .91 indicates that a fairly large number of the candidate pool answered that item correctly[56]. This datum alone, however, is of limited value without a more qualitative analysis of the item in general, reflecting on the criteria of representativeness, facility, validity and discrimination as a whole. First of all, the phrase at all does reflect the criteria of semantic opacity, in that a reading of each word individually will not indicate the meaning of the expression. Therefore, the phrase can be said to be representative of the construct. Further, a look at the distractors does not reveal any salient issue in terms of their validity. Each fits both semantically and grammatically into the stem. Although none of the distractors seems to be drawing very much relative to the key, this fact in itself does not indicate the item is faulty. As suggested in Schmitt (2010: 186-187), "The point that matters is whether that test item reflects testees' knowledge of the lexical item, and not whether it matches the results from other test items." The fact that the facility value is so high, therefore, may simply be an indication that the phrase is so common that the proficiency of the students who took the test was such that it posed little difficulty for them. In such a case, when choosing items for the next field test of the PVST, if two items displayed similar characteristics of roughly equal representativeness, difficulty and validity, the next criterion was discrimination.

In Table 5.11, the value in the last column, 'D', represents the measure of discrimination, or the 'discrimination index'. Summary statistics for all items are presented in Table 5.12.

[56] There is no hard and fast rule regarding what constitutes an acceptable facility value. Bachman (2004), for example, suggests a "rule of thumb" range of between .20 and .80 for norm-referenced tests (p. 138), and discrimination indices for criterion-referenced tests that "maximize the capability of the test for categorizing test takers into groups (e.g. mastery/non-mastery)" (ibid.).

Table 5.12 Summary statistics for all items (Innsbruck field test)

Pilot Item	Phrasal expression	Facility	Discrimination
K1, A1, C1	lead to	.94/.92	.14/.21
K1, A2, C2	have to	.97/.96	.05/.02
K1, A3, C3	a number of	.98/.98	.03/.05
K1, A4	go on	.97	.06
K1, A5	a bit	.97	.06
K1, A6	likely to	.68	.55
K1, B7	deal with	.66	.46
K1, B8	at all	.91	.07
K1, B9	is to	.61	.41
K1, B10	a lot	.97	.06
K1, B11	I mean	.90	.20
K1, B12	at least	.72	.57
K1, C13	so that	.93	.13
K1, C14	used to	.26	.48
K1, C15	rather than	.22	.26
K2, A1, C1	as soon as	.74/.84	.37/.19
K2, A2, C2	find out	.99/98	.01/.05
K2, A3, C3	so far	.90/.91	.25/.27
K2, A4	to do with	.89	.22
K2, A5	for instance	.82	.31
K2, A6	take over	.83	.35
K2, B7	a range of	.97	.10
K2, B8	as a result	.78	.34
K2, B9	take place	.94	.14
K2, B10	and so on	.94	.15
K2, B11	carried out	.74	.43
K2, B12	each other	.91	.19

K2, C13	in particular	.82	.29
K2, C14	expected to	.56	.57
K2, C15	about to	.73	.50
K3, A1, C1	it takes	.95/.93	.15/.19
K3, A2, C2	other than	.85/84	.41/.56
K3, A3, C3	carry on	.92/.88	.24/.18
K3, A4	all over	.78	.50
K3, A5	turn out	.84	.32
K3, A6	in time	.25	.23
K3, B7	feel like	.85	.37
K3, B8	or so	.96	.17
K3, B9	shake your head	.90	.22
K3, B10	whether or not	.90	.25
K3, B11	get to	.82	.34
K3, B12	at once	.43	.65
K3, C13	give up	.87	.35
K3, C14	in touch	.78	.53
K3, C15	get rid of	.58	.69
K4, A1, C1	as yet	.90/.88	.20/.24
K4, A2, C2	prove to be	.89/.89	.33/.36
K4, A3, C3	in effect	.53/.71	.41/.43
K4, A4	happen to	.47	.81
K4, A5	by no means	.73	.30
K4, A6	take advantage	.84	.44
K4, B7	in the light of	.75	.46
K4, B8	give rise to	.42	.41
K4, B9	no matter	.94	.16
K4, B10	come across	.63	.55
K4, B11	even so	.38	.81

K4, B12	run out	.86	.38
K4, C13	might as well	.16	.00
K4, C14	next door	.86	.24
K4, C15	on the other hand	.87	.11
K5, A1, C1	take for granted	.54/.55	.65/.57
K5, A2, C2	as of	.57/.51	.58/.65
K5, A3, C3	would appear	.69/.73	.22/.29
K5, A4	to blame	.55	.57
K5, A5	stand for	.96	.10
K5, A6	by far	.88	.29
K5, B7	keep on	.95	.07
K5, B8	over time	.27	.51
K5, B9	come up to	.87	.40
K5, B10	straight away	.80	.66
K5, B11	shut up	.97	.09
K5, B12	a handful of	.93	.17
K5, C13	you can tell	.28	.57
K5, C14	under way	.25	.03
K5, C15	turn down	.57	.59

As described in Hughes (2003), a discrimination index is simply "an indicator of how well an item discriminates between weak candidates and strong candidates" (p. 226). In classical item analysis, the index is created by indentifying and isolating the upper and lower 27 percent of scorers and comparing how well they performed on the same items (Bachman, 2004: 123). According to Bachman, "we want the average D to be maximal, as this will increase both the variance and internal consistency reliability" (p. 138), particularly important in norm-referencing. However, it should also be noted that an item with a low discrimination index does not mean it should be discarded without further consideration (Hughes, 2003: 227). Items that are very easy or difficult will generally have lower discrimination indices, and there may be reasons for retaining them since, for example, they can allow candidates to ease into a test, and more importantly, they may represent the construct well and provide meaningful information about candidates' knowledge.

As seen in the item in Table 5.11 (*'at all'*, p. 175), also included in each analysis are the percentages of those candidates that skipped the item altogether. While that percentage on the test as a whole may not be of much use beyond a very general idea of difficulty when compared with the same percentages on other items, it does become a potentially useful datum when broken down into upper and lower groups, as shown in Table 5.11, juxtaposed, and compared. When a comparatively large percentage of lower group participants did not attempt an item while almost the entire cohort of higher scoring candidates did, it can be an additional indication of that item's relative difficulty. On the other hand, evidence of a comparatively high percentage of candidates skipping an item even from the upper group, relative the lower group, may suggest an awareness of non-compositionality that the lower group as a whole did not have as much of.

The analysis and inclusion of the item omission data was not straightforward, however. At the lower frequency bands – K4 and K5 in particular – there was some attrition among the lower group, which in turn makes the percentage of test takers who apparently did skipped that item appear inflated in the raw dataset output. In actuality, there is no way to know how well or poorly participants who made it only partially through the test (for example, due to running out of time, or simply giving up) would have done on each item had they progressed that far. At the same time, it of course cannot be assumed that all candidates who did not progress to a certain band would have omitted all the items in that band. Therefore, in order to arrive at a more representative estimate of candidates who skipped a given item in the lower frequency bands especially, the highest number of total test takers shown to have attempted each frequency band was taken to be the total possible attempts for that band. The 'No attempt' figure, therefore, as far as possible, reflects how many candidates could have attempted that item but chose not to.

Taken together, therefore, although the item in Table 5.11 (*'at all'*, p. 175) was judged to be acceptable for inclusion on the final version of the test on the strength of all selection criteria, other items in the same band were found to be superior. Consider the following item as an example:

1.	lead to: No one knows what it will **lead to.**		Facility	Upper	Lower	D
	a.	want	.01	.00	.04	-.04
	b.	have inside	. 02	.00	.06	-.06
	c.	cause in the future	.94	1.00	.86	.14
	d.	find	.01	.00	.04	-.04
		No attempt		0 (%)	9 (4.5%)	

The above item, which became the first item on the test, exhibits characteristics broadly similar to the one in Table 5.11: the distractors do not appear to be drawing very much, the facility value is relatively high, and the validity of the item in terms of representativeness seems to be fine. However, the 'D' is higher (.14 versus .07), and the number of candidates who did not attempt the item is also higher (4.5% versus 2%). All else being equal, therefore, the above item was judged to be of slightly greater overall value in the test.

5.3.5 *Results and discussion of item analysis*

As discussed in the previous section, the facility values and discrimination indices were calculated for each item, and for each distractor in each item. The full results of this analysis is provided in Appendix 15, but in the present section a summary of the results will be presented through exemplification of items that were selected or rejected. In addition, the analysis proved interesting for evidence of some of the issues regarding formulaic language discussed in Chapters 2 and 3, and these insights will also be presented here.

Following the types of analysis discussed in the previous section, all items were first scrutinized for overall representativeness and validity, and then those with the best evidence of relative difficulty and discrimination were identified as the most likely potentials for inclusion on the test. For example, K2, C14 can be considered fairly representative of items that were considered acceptable for inclusion:

K2, C14 (item-total correlation .441)

14.	be expected to: We **are expected to** do it.		Facility	Upper	Lower	D
	a.	are waiting	.18	.04	.27	-.13
	b.	hoping to	.16	.05	.21	-.16
	c.	must	.56	.88	.31	.57
	d.	are able to	.10	.03	.21	-.18
		No attempt		0 (0%)	20 (12%)	

The analysis of K2, C14 reveals no issues in terms of representativeness and validity. Moreover, there is evidence of difficulty, and the p-value on the distractors indicates that they are all contributing to the item. There is some draw even among the upper group on the distractors, but a semantic analysis of them does not show any anomalies (e.g. any one distractor drawing a noticeable and disproportionate of responses away from the key), and in any case the values are very low when compared with the lower group's response behaviour on the same distractors. The D is relatively high, indicating that the item is discriminating fairly well (especially desirable in the case of norm-referencing), and there is evidence in the item omission data that the lower group found it difficult while the upper group did not. This item was therefore selected and can be seen in the present study's final version of the test (Appendix 8).

It was found to always be important, however, not to simply pay attention to the statistical values in each item analysis alone (e.g. p- and D values), but to also carefully observe the differential response patterns from the upper and lower groups and take care not to include items that may have higher D values, for example, but which contain evidence of problematic distractors (i.e., issues with validity). An example of the importance of this more qualitative analysis can be found in K1, C15:

K1, Item C15 (item-total correlation .240)

15.	rather than: I'll cook **rather than** eat.		Facility	Upper	Lower	D
	a.	or maybe	.01	.01	.03	-.02
	b.	but I prefer to	.55	.41	.63	-.22
	c.	before I	.22	.20	.23	-.03
	d.	and not	.22	.38	.12	.26
		No attempt		5 (2.6%)	14 (7.2%)	

Not unlike the item discussed prior to this one (K2, C14), item K1, C15 also exhibits evidence of difficulty (facility value of .22), and does appear to discriminate adequately (.26). What is more, there is strong evidence in the item that distractor b ('but I prefer to') is appropriately drawing many candidates from both groups, providing evidence of relying on the word 'rather' to decode the meaning of the item. This would be interesting (and indeed is) and would have potentially helped to contribute to making this item a good one for the test, if not for the other distractor which seem to be getting much attention in this item, 15c ('before I'). The very fact that this item's p-value was so high (.20) among the higher scorers drew the interest of the researcher in the analysis. Unlike 15b ('prefer'), there is a possibility in 15c that 'before I' might have drawn responses from key because it actually is legitimately interpretable as a correct answer: 'I'll cook before I eat' can be thought of as similar in meaning to 'I'll cook rather than eat'. There is of course no post-hoc way of determining if this was the case among every candidate who chose this distractor, but semantic analysis of it is enough to raise the question. If distractor 15c, therefore, spuriously drew responses away from the key, this behaviour also affected the surrounding choices, and the validity of the entire item is undermined. Hence, even though the item was arguable statistically viable as a choice for the test, it was discarded.

The type of careful quantitative and qualitative analysis described above took place for each item (Appendix 15). Finally, those items which best met the criteria of representativeness, difficulty, validity and discrimination were sequenced in order of descending facility value (i.e. increasing difficulty) for each band into which the items were inserted. The full test with the all chosen items can be seen in Appendix 8, but Table 5.13 provides a summary list of them, presented in the order in which they appear on the rest.

Table 5.13 List of items chosen for revised test

No. on new test	Pilot Item	Phrasal expression	Facility	Discrimination
1.	K1, A4	go on	.97	.06
2.	K1, A1	lead to	.96	.14
3.	K1, C13	so that	.93	.13
4.	K1, B8	at all	.91	.07
5.	K1, B11	I mean	.90	.20
6.	K1, B12	at least	.72	.57
7.	K1, A6	is likely to	.68	.55
8.	K1, B7	deal with	.66	.46
9.	K1, B9	is to	.61	.41
10.	K1, C14	used to	.26	.48
1.	K2, C3	so far	.91	.27
2.	K2, A4	to do with	.89	.22
3.	K2, A6	take over	.83	.35

4.	K2, C13	in particular	.82	.29
5.	K2, A5	for instance	.82	.31
6.	K2, B8	as a result	.78	.34
7.	K2, A1	as soon as	.74	.37
8.	K2, B11	carry out	.74	.43
9.	K2, C15	am about to	.73	.50
10.	K2, C14	are expected to	.56	.57
1.	K3, C13	give up	.87	.35
2.	K3, B7	feel like	.85	.37
3.	K3, A5	turned out	.84	.32
4.	K3, C2	other than	.84	.56
5.	K3, B11	got to	.82	.34
6.	K3, A4	all over	.78	.50
7.	K3, C14	in touch	.78	.53
8.	K3, C15	got rid of	.58	.69
9.	K3, B12	at once	.43	.65
10.	K3, A6	in time	.25	.23
1.	K4, C2	proved to be	.89	.36
2.	K4, C14	next door	.86	.24
3.	K4, B12	ran out	.86	.38
4.	K4, A6	take advantage	.84	.44
5.	K4, C3	in effect	.81	.43
6.	K4, B7	in the light of	.75	.46
7.	K4, A5	by no means	.73	.30
8.	K4, B10	came across	.63	.55
9.	K4, A4	happen to	.47	.81
10.	K4, B11	even so	.38	.81
1.	K5, A6	by far	.88	.29

2.	K5, B9	came up to	.87	.40
3.	K5, B10	straight away	.80	.66
4.	K5, A3	would appear	.69	.22
5.	K5, C15	turned down	.57	.59
6.	K5, A4	to blame	.55	.57
7.	K5, A1	took it for granted	.54	.65
8.	K5, C2	as of	.51	.65
9.	K5, C13	you can tell	.28	.57
10.	K5, B8	over time	.27	.51

In general, there seems to be a trend of increasing discrimination as the frequency decreases, which is what one would generally expect. Some of the items that proved to have the highest facility values and/or discrimination indices were also interesting from the point of view of the issues and research that has been described in previous chapters in this thesis. Consider the last item on the test, (K5, B8) as an example:

K5, Item B8 (item-total correlation .378)

8.	over time: **Over time** it was cheaper.		Facility	Upper	Lower	D
	a.	long ago	.61	.38	.73	-.35
	b.	eventually	.27	.57	.06	.51
	c.	when it was too late	.08	.03	.13	-.10
	d.	at the perfect moment	.04	.03	.08	-.05
		No attempt		12 (6%)	10 (7%)	

Although this item has a very low facility value when compared to other items, it does discriminate well (.51). It is interesting to note that while the majority of the higher scoring students were able to answer the item correctly, those that did not (38 percent of them) chose distractor 'long ago'. It is likely the case that the test takers in these cases simply read 'over' as meaning 'abundant' or 'too much' and combined it with 'time' to try to arrive at a meaning. There were a number of other items that exhibited evidence of a focus on individual words in distractors which caused choosing the wrong answer, such as K2, A3 and K3, B12:

K2, Item A3 (item-total correlation .488)

3.	so: It's good **so far**.		Facility	Upper	Lower	D
	a.	until now	.90	1.00	.75	.25
	b.	but not really	.04	.00	.08	-.08
	c.	sometimes	.01	.00	.02	-.02
	d.	from a distance	.05	.00	.15	-.15
		No attempt		0 (0%)	12(5%)	

K3, Item B12 (item-total correlation .503)

12.	at once: I did it **at once**.		Facility	Upper	Lower	D
	a.	one time	.47	.16	.78	-.62
	b.	many times	.00	.00	.00	.00
	c.	early	.02	.00	.06	-.06
	d.	immediately	.43	.81	.16	.65
		No attempt		4 (2%)	29 (16%)	

In K2, A3, there can be little doubt that the 15 percent of the candidates that chose option 3d ('from a distance') is because they associated it with the word 'far' in the stem. Likewise, the word 'one' in item 12 (K3) seems to have even drawn 16 percent of the upper group, let alone the vast majority of the lower group. This focus on the individual words when a phrase is not known/recognized in turn would seem consistent with studies like the one described in Chapter 2 (Martinez & Murphy, 2011), showing the negative repercussions of decoding individual words when they are parts of phrasal expressions. Most importantly in terms of the purposes of the item analysis, such items were found to be especially representative of the construct being examined.

It was also sometimes surprising to see how even very common expressions were missed by the test takers, sometimes even the upper group:

K1, Item C14 (item-total correlation .389)

14.	used to: I **used to** go.		Facility	Upper	Lower	D
	a.	want to	.12	.01	.29	-.28
	b.	did before	.26	.55	.07	.48
	c.	usually	.56	.40	.54	-.14
	d.	always	.07	.05	.09	-.04
		No attempt		1 (0.5%)	14 (7.2%)	

In the item above (K1, C14), there is evidence that both upper and lower groups do not understand what 'used to' means, and that they read it as 'usually' or 'am used to'. It is even likely that, especially among the upper group candidates, they in fact have been taught the 'used to' (past) form, but what this item suggests is that a sizeable proportion of them still may misinterpret that item in context. Considering that 'used to' is one of the most common phrasal expressions in the English language according to the

PHRASE List, this perhaps deserved further investigation[57]. (Further discussion the pedagogical implications of the results of this study can also be found in Chapter 6.)

5.3.6 *Validation of final version of the Phrasal Vocabulary Size Test*

The version of the PVST which was arrived at through the item analysis detailed in Section 5.3.4 was submitted to a check of the extent to which test-takers actually know the phrases tested in the items they answer correctly. An online version of the test was created (Appendix 16), with a pre-test translation task of each phrase tested, thus allowing for further confirmation of actual knowledge of the phrase (cf. Shillaw, 2009). A link was sent to a number of Brazilian teachers of English known to the researcher in a professional capacity (the researcher is highly proficient in Brazilian Portuguese). Nearly all the teachers to whom the test was sent responded (n = 39), however, only the data of those who completed all five frequency levels (n = 31) were entered into SPSS for analysis. (Scores are shown in Table 5.14.)

Table 5.14 Central tendency for Brazil PVST validation study (Max. poss. = 10)

Band	N	Minimum	Maximum	Mean	SD
K1 Score	31	6	10	9.45	0.96
K2 Score	31	8	10	9.71	0.64
K3 Score	31	7	10	9.10	0.98
K4 Score	31	6	10	9.39	1.02
K5 Score	31	4	10	8.77	1.52

[57] In fact, during the validation of the English Profile Wordlist (Chapter 4), one of the points of discussion that was raised when apparent discrepancies in level were found was what to do in the case of 'used to'. It is clearly frequent enough to belong to the A1/A2 lists, but currently is still listed at B1 because of the Cambridge Learner Corpus data which shows that below that level it is still relatively rare in student production.

Although the scores were calculated for all participants (Table 5.14), the main focus of the study was to determine, through written translations, how accurately the answers given by the test takers actually reflected their demonstrable knowledge of the phrases tested. Unfortunately, of the 31 test respondents, only 16 completed the translation task. These translations were coded into SPSS as follows:

'0' = Incorrect answer and translation
'1' = Correct answer and translation
'2' = Incorrect answer, correct translation
'3' = Correct answer, incorrect translation

Therefore, only codes '0' and '1' above reflect answers that are consistent with actual knowledge; codes '2' and '3' would reflect some kind of knowledge discrepancy. The frequencies of these codes are provided in Table 5.15.

Table 5.15 Breakdown of consistent and discrepant answer behavior on PVST

<table>
<tr><th></th><th>Answer type totals</th><th>Combined totals*</th></tr>
<tr><td>Answer type (consistent)</td><td rowspan="2">33</td><td rowspan="3">740 (consistent)</td></tr>
<tr><td>'0' = Incorrect answer and translation</td></tr>
<tr><td>'1' = Correct answer and translation</td><td>707</td></tr>
<tr><td>Answer type (discrepant)</td><td rowspan="2">6</td><td rowspan="3">8 (discrepant)</td></tr>
<tr><td>'2' = Incorrect answer, correct translation</td></tr>
<tr><td>'3' = Correct answer, incorrect translation</td><td>2</td></tr>
</table>

* 800 total possible (52 translations missing)

As shown in Table 5.15, the respondents demonstrated that they were consistent in their answering of items on the PVST (i.e. answers on test reflected actual knowledge) 98.93% of the time, showing that – at least in this small-scale study – the answers that participants gave on the test mostly reflected their true knowledge of the phrases. Moreover, although there were a total of 6 answers of code type '2', these were relatively rare, and never occurred more than once on any given item. Code type '3', which might be indicative of guessing, only occurred twice (on different items)[58]. Although the sample size focused on here (16 participants) is relatively small, on the basis of these data it can be concluded that no one particular item stands out as requiring revision.

5.3.7 Scoring and interpretation of scores of the Phrasal Vocabulary Size Test

As mentioned earlier in the present chapter, scoring has grown to be thought of as a key of component of test validity among language testing researchers (Fulcher, 2010). In fact, Weir (2005) has suggested that "validity resides in test scores" (p. 12). Scoring of the Phrasal Vocabulary Size Test needs to have its basis in empirical data and be meaningful. The items on the PVST are of course written based on the PHRASE List, which contains 505 phrasal expressions. However, it would not be valid to provide one composite score alone for the PVST that would somehow suggest the test taker knows a certain proportion of the entire PHRASE List. The phrasal expression in the PVST are sampled from frequency levels that are equivalent to the same frequency levels of single words, and the scores from the PVST should therefore be frequency-sensitive. Moreover, the standard error of measurement (SEM) should also be taken into account for the new version of the PVST which as resulted from the present study, but as the SEM uses reliability statistics to inform the calculation, and the items on the new version of the

[58] The 52 missing translations are items that were not attempted.

test have not yet been checked for internal consistency, any guesses would be merely that. Nonetheless, in theory the scoring would be indicative of a proportion of phrases known per frequency level on the PHRASE List, and therefore the scoring is suggested as being calculated as follows:

$$KS = T \times N_e/NK_i$$

where KS is the score on the individual 1000 word frequency level;
T is the total number of phrasal expressions answered correctly on each level;
N_e is the total number of expressions per level on the PHRASE List;
K_i is the total number of items per level on the test.

In practice (e.g. for a student or teacher), it would look something like the following:

K1 SCORE = T x 3.2
K2 SCORE = T x 8.4
K3 SCORE = T x 12.9
K4 SCORE = T x 15.7
K5 SCORE = T x 10.3

Hence, in the case of the hypothetical student in Table 5.16, the score sheet would appear as a series of estimates.

Table 5.16 Sample score sheet for Phrasal Vocabulary Size Test

PVST Level	**No. correct**	**Estimated knowledge of PHRASE List expressions**	**Total possible**
First 1000	9	28	32
Second 1000	8	67	84
Third 1000	5	65	129
Fourth 1000	2	31	157
Fifth 1000	1	10	103
	Estimated total	201	505

It is felt that the construct of the phrasal expression is sound, as was the compilation of the list (Chapters 2 and 3). If the list is valid, then as long as the test items have been carefully written and chosen – as is believed to have been the case in the research discussed in this chapter – then the scores on the test are in theory reflective of a valid construct. Furthermore, the scores can only be indicative of a certain degree of knowledge on each frequency band of the PHRASE List, and since the number of total expressions in each band is different, the representativeness of the sample size of each of the bands on the PVST is also different. (10 items represent 31 percent of the list at K1, for example, but just over 6 percent at K4.) Therefore, the degree of accuracy of the scoring estimate provided will be less at the less frequent levels, and this is a factor which should be considered and addressed as the test continues to evolve along with the scoring. However, in the case of the student's score sheet in Table 5.16, the estimate of 201 does not seem unreasonable. All that can be affirmed, however – and this needs to be made clear to the taker of the PVST – is that the candidate in question has demonstrated a certain degree of recognition of the phrasal expressions tested, and by no means does that mean that the same person can use those expressions productively, or even readily recognize them in speech. Moreover, the student should be made aware of the type of expression being tested, and that the PHRASE List does not claim to represent all the most important formulaic sequences in English, but just a special subset of them, to a certain frequency threshold, derived from the BNC.

Finally, no claims can or should be made regarding what the knowledge demonstrated on the PVST yet means. It may be that the scores also happen to correlate with reading comprehension measures, for example, but this is a practical element of the test, and its consequential validity, that will need to be explored more in the future. Also, as seen in the previous chapter, some general alignment between the PHRASE List and the EPW can be found, and there is therefore the temptation to suggest that, for example, a student who demonstrates a mastery of all levels to at least the K4 threshold can be said to be at least B1. However, such a claim would at the moment be only based on the EPW CEFR alignment, which itself has not been fully validated. Nonetheless, these, as well as other possible future directions for the research, will be taken up again in the next chapter.

5.4 Conclusion

The research question posed at the beginning of this chapter asked how a test of phrasal vocabulary could be developed. In order to address that question, first existing tests of vocabulary were turned to as it was envisaged that since phrasal expressions – as has been argued throughout this thesis – are really an important part of vocabulary and not a separate construct, that any new test involving the assessment of phrasal expressions should complement and be usable with instruments already widely employed. A comparison between the response behavior on the Vocabulary Levels Test and Vocabulary Size Test formats revealed that the cognitive processes participants reported when taking the former implied a threat to its validity, due to phrases presented in isolation rather than in sentences. The Vocabulary Size Test, therefore, which allows for the expression to be embedded in a sentence, was considered a better format. The piloting revealed that the format both more accurately reflected true knowledge on the items tested, and was generally preferred for its relative lack of interference of metacognition that posed threats to validity. Following a field test of what came to be called the Phrasal Vocabulary Size Test, items were identified that exhibited desirable testing properties such as good representativeness of the construct and the ability to discriminate between stronger and weaker candidates, and those items now form the full version of the PVST as it now stands.

Although the preceding paragraph to some extent answers the original question as posed, there are still many questions that need to be addressed regarding the test. For example, although some generalizations about the response behaviour in the items can made, they can currently only be made about the particular group that took the test, namely Austrian German speaking learners of English. Further field testing should now take place among a more heterogeneous sample of candidates. The items on this version of the test have been analyzed through classical item analysis, which was intentional and useful for this first round of rewrites to arrive at an understanding of the behaviour of the particular group that took the test (Stewart & White, 2011), but classical item analysis has its limitations. One of the principal limitations is that the item and score statistics discussed in this chapter are essentially "sample-based descriptive statistics" (Bachman, 2004: 139), from which generalizations regarding the items and how they might behave outside the sample discussed are simply of limited validity. For that reason, now that the test is complete (Appendix 8), latent trait item response theory models (IRT), such as Rasch, should be applied to the test as the item difficulty estimates that emerge from such models are theoretically independent of the sample who take the test (Crocker & Algina, 1986: 363). Indeed, the Nation and Beglar VST was in recent years submitted to such an exercise and achieved positive results (Beglar, 2009). The ability of IRT models to test for unidimensionality[59], for example, would perhaps be even more relevant for a test that purports to test multiword expressions as if they were single lexical items[60]. Until such investigations take place, the results of the test and the scores it produces need to be weighed with some degree of reservation. Nevertheless, it is hoped that the test that resulted from the research presented here helps provides the groundwork for such investigation.

[59] In a test, the focus on "one attribute or dimension at a time" (Bond & Fox, 2007: 34)
[60] The test results from the present field test study were submitted, however, to a principal components factor analysis in SPSS which identified loading on only one factor.

Chapter 6 – Summary and conclusion

The present chapter aims to provide a summary of the research presented in the thesis, discussing both its limitations and applications, and concludes by discussing possible future directions for the work.

6.1 The story so far

As illustrated in the 'phraseological iceberg' in Chapter 1 (Figure 1.1) of this thesis, it has long been clear that multiword expressions are an important part of the lexicon that need to be considered when teaching and testing language. Chapter 2 presented the need for a list of formulaic sequences that could serve pedagogical functions comparable to those served by frequency lists of words, such as the GSL. Chapter 3 then described in detail the development of the PHRASE List, and the practical applicability of that list was illustrated in the projects presented in Chapters 3 and 4. However, although to some extent the research presented in the thesis is product-oriented in the sense that it is hoped that practical outcomes have emerged from it, it is also hoped that neither the applications of the list, nor the research from which it was derived, end with this thesis. Indeed, as with probably any research endeavour, while some questions have perhaps been answered during the development of the PHRASE List and related projects extending from it, still further questions arose during and at the conclusion of the work – the answers to which lie beyond the scope of this thesis – yet nonetheless are deserving of attention. Moreover, the projects described in the thesis serve as examples of ways in which the PHRASE List may be employed in research and language pedagogy, but by no means represent the full extent of its potential.

6.2 Limitations

There are clear limitations to the main product of the research carried out and presented in this thesis, the PHRASE List. One of the main limitations is the fact that the list is derived from one corpus, the BNC. While the BNC is a large corpus and has been employed extensively in various areas of applied linguistic research over the years, it has not been updated since the 1990s, and only contains 10% spoken discourse. The extent to which the same phrasal expressions would emerge from other corpora – and how comparable they would be in terms of frequency – is an unknown at the time of writing. It would be useful to apply the same criteria used in selecting expressions for the PHRASE List to other, non-BNC corpora – particularly non-British corpora.

Another limitation of the PHRASE List, and therefore any research or applications of research that have emerged or will extend from it, is that the selection criteria are themselves restrictive. The items chosen for inclusion in the PHRASE List are not intended to represent the entirety of all phraseology in English (if such a feat is even possible), but rather a particular type of expression with particular attention to interpretability from the standpoint of an L2 learner. While the criteria were shown to be applicable reliably and consistently, it cannot and should not be the claim here that they are the only criteria that can be applied when compiling a list of the most common expressions in English. Indeed, the Shin and Nation (2008) study is an example of how there are other approaches that may be taken in selecting multiword expressions for a pedagogically-relevant list of phrases, and therefore the PHRASE List should not be taken to exist in lieu of other lists, but instead as a complement and/or alternative to them.

6.3 Applications in research and pedagogy

The studies detailed in Chapters 4 and 5 exemplify the type of applications in language research and pedagogy in which the PHRASE List may be usefully employed. However, while still unexplored, there seem to be other clear potential uses for the PHRASE List. One use which stands out in particular is in materials design and development. For example, currently one can enter a text into a lexical profiler which will check that text against a frequency list of single words. What results is a breakdown of all the words in the text into frequency bands, which can potentially help inform judgements about the relative difficulty of that text, for example. An obvious limitation in the light of the research presented in this thesis is that there are no multiword expressions represented in the lists against which texts are currently checked, and therefore there may be a degree of difficulty which may not be accounted for in current profiling methods. Figure 6.1 and Tables 6.1 and 6.2 serve to exemplify how methods of text profiling that do not account for opaque phraseology risk underestimating the lexical complexity of a text.

in Organization

Richard H. Axelrod, Emily Axelrod, Rob

Seventy percent of organizational change efforts fail or fall short of achieving their intended objectives. This figure is even more astounding when you take account of a recent Oxford University study on change, which found that over 68% of employees welcomed meaningful involvement in change. Clearly, we are missing the boat when it comes to effective change, and our clients are not getting what they paid for.

We believe that there is a relationship between the failure rate of

The Fe

As the
velopir
way ou
emerge

• • • • •

Heed organizati

Source: Axelrod, et al. (2006).

Figure 6.1. *Introduction from an authentic academic text (phrases underlined)*

Table 6.1 Lexical profile of text in Figure 6.1 counting only single words

Frequency Band	Words (types)	Text coverage (tokens)
0-1000	*a account and are boat change clearly comes effective efforts employees even fail fall figure for found getting in is it meaningful missing more not number of on or our over paid recent seventy short study take that the their they this to university we welcomed what when which you*	86.57%
1001-2000	*intended*	1.49%
Academic Word List	*achieving involvement percent*	4.48%
Off list	*astounding clients objectives organizational Oxford*	7.46%
	Words in Top 2,000:	88.06%
	+ AWL words:	4.48%
	Total text coverage:	**92.54%**

Table 6.2 Lexical profile of text in Figure 6.1, phrases accounted for

Frequency Band	Words and Phrases (types)	Text coverage (tokens)
0-1000	*a and are change clearly effective efforts employees even fail figure for found getting in is meaningful more not number of on or our over paid recent seventy study that their they this university we welcomed what when which you*	67.16%
1001-2000	*intended*	1.49%
Academic Word List	*achieving involvement percent*	4.48%
Off list	*astounding clients fall short of missing the boat objectives organizational Oxford take account of when it comes to*	26.87%
	Words in Top 2,000:	68.65%
	+ AWL words:	4.48%
	Total text coverage:	**73.13%**

The number of words "off list" in the text in Figure 6.1 rises from a relatively manageable 7.46% (Table 6.1) to a much more onerous 26.87% when multiword expressions are accounted for, with their respective frequencies (Table 6.2). Therefore, assuming a learner knows only words within the 2000 word family level, and none in the AWL, that coverage drops from 88.06% to a much more challenging 68.65%[8].

An analysis of the type illustrated above could be of value to such people as foreign language textbook writers, editors and authors of graded readers[61], and test item writers (a person who, for example, finds and adapts texts for inclusion in a test). At the present time, the texts that are adapted or specifically written for input in textbooks, graded readers, and even tests of reading skill are generally submitted to an analysis of grammatical and single-word vocabulary analysis in order to determine their relative

[61] An original or adapted work of fiction or non-fiction designed to be of language that is 'graded' to the level of the intended reader, usually for the purposes of language proficiency maintenance and/or development.

difficulty – if analyzed linguistically at all. Indeed, in many cases, selection of a text and decisions regarding its fitness for a given level are left to the subjective judgments of individual authors, test writers, and so on. While knowledge and experience doubtlessly positively contribute to generally good evaluation of the overall linguistic suitability of a text, it is also the experience of this researcher – who also has extensive experience designing and writing textbooks designed for learners of English around the world – that such decisions often come down to such things as 'salience' (i.e. how much a lexical item 'stands out' in a text as useful, for example), which also means that, inevitably, less salient items such as very common (but potentially useful) phrasal expression composed of highly frequent words get missed.

One can therefore envisage how a test item writer or textbook author might find an analysis like the one generated in Table 6.2 useful. Moreover, such a frequency-sensitive breakdown could potentially help more accurately discriminate between levels that are still under-specified, such as the difference between what characterizes the difference in proficiency between a CEFR C1 and C2 level of mastery. It would be unsurprising to find that if the texts contained in current popular examinations of proficiency (such as the Cambridge First Certificate Examination) were subjected to such a phraseological analysis that many items that could help discriminate are not targeted for assessment in the instrument, and that still others are included as part of the test that perhaps should not be because they are too difficult (or even easy) for the intended level of candidate. At the time of writing, the software to carry out such an automated analysis using the PHRASE List is under development (Bax, 2011), so it is not unreasonable to expect that such a tool may become available in the near future.

6.4 Just the tip of the phraseological iceberg (or what to do with those 4.2 million n-grams)

To conclude the chapter and the thesis, it is interesting to return to the metaphor of the iceberg in Chapter 1 (Figure 1.1). To a great extent, the PHRASE List itself can be considered just the tip of a phraseological iceberg – and there is still much to learn about the submerged portion of that iceberg. Returning to the n-gram data presented in Chapter 3, there is ample evidence that the iceberg may run deep indeed. The original extraction from the BNC rendered a raw unfiltered list of over 4 million items, of which less than 15,000 were ultimately analyzed closely for consideration as candidates for inclusion in the PHRASE List. One obvious direction forward, therefore, would be to explore the phraseological extent of phrasal expressions beyond the 5000 word-frequency threshold represented in the PHRASE List. Vocabulary tests could then be devised, for example, that extended into the lower frequency ranges and included phrasal expressions.

However, the researcher also believes that there may be still more profound implications for those 4.2 million unanalyzed n-grams with regard to our current estimates of how many words the average adult native speaker of English is purported to know. Current estimates of around 20,000 word families (e.g. Zechmeister et al., 1995) may show themselves to be too modest if one also takes into account the type of lexical items included on the PHRASE List. Estimates would merely be conjecture at the present stage of the research, but even if just 1% of the unanalyzed n-gram list can be considered as vocabulary items that should be included in word knowledge estimates, the current estimates of adult native-speaker lexicon size increase to around 60,0000. Of course, the only way in which this conjecture can be empirically confirmed would be to replicate a study such as the Zechmeister et al. research, and in order to do that the full n-gram remaining list needs to be analyzed.

Finally, if current estimates of the size of the adult lexicon in English are currently too modest, and if research does indeed show that the size may be at least twice what it is claimed to be in the relevant literature, this fact in itself perhaps raises an additional question: When and how are these items acquired? Wray and Perkins (2000) have proposed that children may go through an initial holistic mode of processing which lasts until a certain age in early childhood, and then proceed to acquire language more analytically or atomistically, to finally revert to a more holistic processing mode, which in turn coincides with expanded social roles. Such a proposal can be interpreted to suggest that there may be a kind of second 'vocabulary burst' (e.g. Nelson, 1973) in childhood vocabulary development, one involving phrases; however, little to nothing is yet known as to when this phenomenon occurs, or even if indeed it does occur. A longitudinal or even cross-sectional study involving a sampling from items extracted from the 4.2 million n-grams still to be analyzed might be able to contribute to a research instrument to help detect the existence of such a burst.

References

Aijmer, K. (1996). *Conversational routines in English: Convention and creativity.* London: Longman.

Aitchison, J. (1987). *Words in the mind: An introduction the mental lexicon.* Malden, MA: Blackwell Publishing.

Alderson, J.C. (2005). *Diagnosing language proficiency: The interface between learning and assessment.* London: Continuum.

Alderson, J.C. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383-409.

Alexander, R.J. (1987). Problems in understanding and teaching idiomaticity in English. *Anglistiik und Englischunterricht*, 32, 105-122.

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Anglin, J.M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development* Serial No. 238, vol. 58, No. 10. Chicago, IL: Society for Research in Child Development.

Arnaud, P.J.L. and Savignon, S.J. (1997). Rare words, complex lexical units and the advanced learner. In: J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition* (157-173). Cambridge: Cambridge University Press.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67-82.

Axelrod, R.H., Axelrod, E., Jacobs, R.W., and Beedon, J. (2006). Beat the odds and succeed in organizational change. *Consulting to Management*, 17(2), 1-4.

Bachman, L.F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Barfield, A. and Gyllstad, H. (2009). *Researching collocations in another language: Multiple Interpretations.* Basingstoke: Pelgrave Macmillan.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.

Bax, S. (2011). Automated text analysis: analyzing concreteness, multiword expressions and metadiscourse markers in text corpora. Presented at the English Profile Research Seminar, 11 February, University Centre, Cambridge (UK).

Beglar, D. (2009). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 26(4), 1-22.

Beglar, D. and Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*, 16(1), 131-162.

Bensoussan, M., & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7, 15-32.

Bertram, R., Baayen, H.R. and Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, 42, 390-405.

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286.

Biber, D., Conrad, S., and Cortes, V. (2004). *If you look at ...:* Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Longman.

Bishop, H. (2004). Noticing formulaic sequences: A problem of measuring the subjective. *LSO Working Papers in Linguistics, 4*, 15-19.

Blackwell, S. (1987). Syntax versus orthography: Problems in the automatic parsing of idioms. In R. Garside, G. Leech, and G. Sampson, (Eds.), *The Computational Analysis of English* (pp. 110-119). Harlow: Longman.

Boers, F. and Lindstromberg, S. (2009). *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Basingstoke: Pelgrave Macmillan.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., and Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research*, 10(3), 245-261.

Bond, T.G. and Fox, C.M. (2007). *Applying the Rasch model.* Mahway, NJ: Lawrence Erlbaum and Associates.

Bransford, D. B., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 717-726.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, 34, 15-35.

Brians, P. (2008). *Common errors in English usage.* Sherwood, OR: Franklin, Beedle & Associates.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. In B.D. Joseph and R.D. Janda (Eds.), *The handbook of historical linguistics* (602-623). Oxford: Blackwell.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4), 711-733.

Campion, M. and Elley, W. (1971). An academic vocabulary list. Wellington, New Zealand: Council for Educational Research.

Capel, A. (2010). A1-B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(e3), 1-11.

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior*, 26(4), 413-437.

Channel, J. (1994). *Vague language*. Oxford: Oxford University Press.

Chronbach, L.J. and Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In: J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition* (225-237). Cambridge: Cambridge University Press.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, 11(3), 38-63.

Cobb, T. (n.d.). Web VocabProfile [Online program]. Retrieved March 17, 2008 from http://www.lextutor.ca/vp/, [an adaption of A. Heatley & P. Nation, *Range* (1994)].

Cobb, T., (n.d.), Why and how to use frequency lists to learn words. Available: http://www.lextutor.ca/research/ [March 22, 2008].

Conklin, K. and Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1), 72-89.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In: M. Lewis (Ed.), *Teaching collocation: Further developments in the Lexical Approach* (70-86). Hove: LTP.

Cook, G. (2001). The uses of computerized language corpora: A reply to Ronald Carter. In: D.R. Hall and A. Hewings (Eds.), *Innovation in English Language Teaching* (64-70). London: Routledge.

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly, 33*(2), 233-262.

Corrigan, R., Moravcsik, E.A., Ouali, H. and Wheatley, K.M. (Eds.) (2009a). *Formulaic language volume 1: Distribution and historical change*. Amsterdam: John Benjamins Publishing Company.

Corrigan, R., Moravcsik, E.A., Ouali, H. and Wheatley, K.M. (Eds.) (2009b). *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations.* Amsterdam: John Benjamins Publishing Company.

Corsini, R.J. (1999). *The dictionary of psychology.* New York: Routledge.

Coulmas, F. (1991). *Conversational routine.* Berlin: Mouton De Gruyter.

Cowie, A.P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics,* 2(3), 223-235.

Cowie, A.P. (1993). Getting to grips with phrasal verbs. *English Today,* 9(4), 38-41.

Cowie, A.P. (1998). *Phraseology: Theory, analysis, and applications.* Oxford: Oxford University Press.

Cowie, A.P., Mackin, R., and Mcaig, I.R. (1983). *Oxford dictionary of English idioms.* Oxford: Oxford University Pres.

Coxhead, A.J. (1998). *An academic word list* (English Language Institute Occasional Publication No. 18). Wellington, New Zealand: Victoria University of Wellington.

Coxhead, A.J. (2000). A new academic word list. *TESOL Quarterly,* 34, 213-238.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory.* San Diego: Harcourt Brace.

Crow, J.T. (1986). Receptive vocabulary acquisition for reading comprehension. *The Modern Language Journal,* 70(3), 242-250.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics,* 14(1), 29-59.

Dahlmann, I. and Adolphs, S. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)? *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions,* June 2007, Association for Computational Linguistics, 49-56.

Daller, H., Milton, J., and Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

Davies, M., and Gardner, D. (2010). *A frequency dictionary of contemporary American English.* New York: Routledge.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair and M. Hundt (Eds.), *Corpus linguistics and linguistic theory: Papers from ICAME 20 1999* (51-68). Amsterdam: Rodopi.

De Jong, J. (2010). *Symposium 1: General Language Proficiency: Academic review.* Delivered at the Language Testing Forum, 20 November 2010, Lancaster (UK)

Deignan, A. (2009). Searching for metaphorical patterns in corpora. In P. Baker, (Ed.), *Contemporary Corpus Linguistics* (9-31). London: Continuum.

Dorgeloh, H. and Wanner, A. (2009). Formulaic argumentation in scientific discourse. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (523-544). Amsterdam: John Benjamins Publishing Company.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Elley, W., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly, 19*, 53-67.

Elley, W.B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, 24(2), 174-187.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier and S. Granger (Eds.), *Phraseology in language learning and teaching* (1-13). Amsterdam: John Benjamins.

Ellis, N.C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91-126.

Ellis, N.C. (2008). Phraseology: The periphery and the heart of language. In: F. Meunier and S. Granger (Eds.), *Phraseology in language learning and teaching* (1-13). Amsterdam: John Benjamins.

Ellis, N.C. and Frey, E. (2009). The psycholinguistic reality of collocation and semantic prosody (2): Affective priming. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (473-498). Amsterdam: John Benjamins Publishing Company.

Ellis, N.C. and Robinson, P. (2008). An introduction to cognitive linguistics, second language acquisition, and language instruction. In: P. Robinson and N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (3-24). New York: Routledge

Ellis, N.C. and Sinclair, S.G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology*, 49A(1), 234-250.

Ellis, N.C., Simpson-Vlach, R. and Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3), 375-396.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics*, 12(1), 25-53.

Erman, B. (2009). Formulaic language from a learner perspective: What the learner needs to know. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (323-346). Amsterdam: John Benjamins Publishing Company.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29-62.

Estill, R.B. and Kemper, S. (1982). Interpreting idioms. *Journal of Psycholinguistic Research*, 11(6), 559-568.

Eykmas, J., Van de Velde, H., van Hout, R. and Boers, F. (2007). Learners' response behaviour in Yes/No vocabulary tests. In H. Daller, J. Milton and J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (59-76). Cambridge: Cambridge University Press.

Fan, M.Y. (2000). The Dictionary Look-up Behavior of Hong Kong Students: A Large-Scale Survey. *Education Journal*, 28(1), 123-138.

Farghal, M. and Obiedat, H. (1995). Collocations: A neglected variable in EFL. *International Review of Applied Linguistics*, 33(4), 315-331.

Fernando, C. and Flavell, R. (1981). *On idiom: Critical views and perspectives*. Exeter: University of Exeter.

Flower, J. and Martinez, R. (1995). *American business vocabulary*. Hove: LTP.

Fogliata, A., Rizzo, S., Reati, F., Miniussi, C., Oliveri, M., and Papagno, C. (2007). The time course of idiom processing. *Neuropsychologia*, 45(14), 3215-3222.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan and M. Swain (Eds.) *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 75-94). Harlow: Longman.

Fox, G. (1987). The case for examples. In J.M. Sinclair (Ed.) *Looking up: An account of the COBUILD Project in lexical computing* (137-149). London: Harper Collins.

Francis, N.W. and Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment*. New York: Routledge.

Gardner, D. (2007). Validating the construct of *Word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241-265.

Gardner, D. and Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339-359.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson (Eds.), *The Computational Analysis of English* (30-41). London: Longman.

Garside, R., Leech, G., and McEnery, A. (1997). *Corpus annotation: Linguistic information from computer text corpora.* New York: Longman.

Gibbs, R., Bogadanovich, J., Sykes, J., and Barr, D. (1997). Metaphor in idiom comprehension. *Journal of Memory and Language*, 37, 141-154.

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language*, 23(1), 65-83.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 8(3), 183-206.

Giora, R. (1999). On the priority of salient meanings: Studies of literal and figurative language. *Journal of Pragmatics*, 31(7), 919-929.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory and Cognition*, *10*(6), 597-602.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11(4), 341-363.

Gouverneur, C. (2008). Phraseology in foreign language learning and teaching: Thirst for efficient metalanguage. Paper presented at the *Formulaic Language Research Network (FLaRN) Third International Postgraduate Conference*, 20 June, 2008, Nottingham, UK.

Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-drived approach to EFL textbook analysis. In F. Meunier and S. Granger (Eds.) *Phraseology in foreign language learning and teaching*, (223-243). Amsterdam: John Benjamins Publishing Company.

Grabe, W. (2002). Dilemmas for the development of second language reading abilities. In: J.C. Richards and W.A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (276-286). Cambridge: Cambridge University Press.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics*, 24, 44-69.

Grabe, W. and Stoller, F.L. (1997). Reading and vocabulary development in a second language: A case study. In: J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition* (98-121). Cambridge: Cambridge University Press.

Granger, S. and Meunier, F. (2008). *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing Company.

Granger, S. and Paquot, M. (2008). Disentangling the phraseological web. In S. Granger and F. Meunier (Eds.). *Phraseology: An interdisciplinary perspective* (27-50). Amsterdam: John Benjamins Publishing Company.

Grant, L. and Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics*, 25(1), 38-61.

Grant, L., and Nation, P. (2006). How many idioms are there in English? *ITL - International Journal of Applied Linguistics*, 151, 1-14.

Greenall, S. (1997). *Reward elementary*. Oxford: Macmillan.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, 39(6), 1041-1070.

Haynes, M. (1993). Patterns and perils of guessing in second language reading. In T. Huckin, M. Haynes and J. Coady (Eds.) *Second Language Reading and Vocabulary Learning* (46-66). Norwood: Ablex Publishing Corporation.

Heatley, A. and Nation, P. (1994). Range. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Heibeck, T.H. and Markman, E.M. (1987). Word learning in children: An examination of fast mapping. *Child Development*, 58(4), 1021-1034.

Hill, D.R. (1997). Survey review: graded readers. *ELT Journal*, 51(1), 57-81.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: Further developments in the Lexical Approach* (47-69). Hove, UK: LTP.

Himmelmann, N.P. (2006). The challenges of segmenting spoken language. In J. Gippert, N.P. Himmelmann, and U. Mosel (Eds.), *Essentials of Language Documentation*, (253-274). Berlin: Mouton de Gruyter.

Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

Hirsh, D., and Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Oxford: Routledge.

Hoftland, K. and Johansson, S. (1982). *Word frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Howarth, P. (1998a). The phraseology of learners' academic writing. In Cowie, A. (Ed.), *Phraseology: Theory, analysis and applications* (161-186). Oxford: Oxford University Press.

Howarth, P. (1998b). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24-44.

Hsu, J.T. (2008). Role of the multiword lexical units in currect EFL/ESL textbooks. *US-China Foreign Language*, 27-39.

Hu, H.M. and Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-429.

Hudson, R. A. (2007). *Language Networks: The new word grammar.* Oxford: Oxford University Press.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Hulstijn, J.H. (1993). When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables. *Modern Language Journal*, 77, 139-147.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4-21.

Hymes, D. (1971). *On communicative competence.* Philadelphia: University of Philadelphia Press.

Jiang, N. and Nekrasova, T.M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, 91(3), 433-445.

Jung, E. H. (2003). The effects of organization markers on ESL learners' text understanding. *TESOL Quarterly*, *37*(4), 749-759.

Kaivanpanah, S. and Alavi, M. (2008). Deriving unknown word meaning from context: Is it reliable? *RELC Journal*, 39(1), 77-95.

Kapatsinski, V. and Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (499-520). Amsterdam: John Benjamins Publishing Company.

Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. *ACL 2006: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 12-19.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135-155.

Kirkpatrick, A. (2007). *World Englishes: Implications for international communication and English language teaching.* Cambridge: Cambridge University Press.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal*, 78(3), 285-299.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, 59(4), 322-332.

Korkontzelos, I. and Manandhar, S. (2009). Detecting compositionality in multi-word expressions. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 65-68.

Krashen, S. (1993). *The power of reading: Insights from research*. Englewood, CO: Libraries Unlimited.

Kučera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuiper, K. (1996). *Smooth talkers*. Hillsdale, NJ: Lawrence Erlbaum.

Kuiper, K. and Haggo, D. (1984). Livestock auctions, oral poetry, and ordinary language. *Language in Society*, 13, 205-234.

Lado, R. (1955). Patterns of difficulty in vocabulary. *Language Learning*, 6(1), 23-41.

Lado, R. (1961). *Language testing*. London: Longman.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Laufer, B. (1989a). What percentage of lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (316-323). London: Multilingual Matters.

Laufer, B. (1989b). A factor of difficulty in vocabulary learning: Deceptive transparency. *AILA Review, 6*, 10-20.

Laufer, B. (1991). How much lexis is necessary for reading comprehension? In: Arnaud, P. J. L. and H. Béjoint (Eds.), *Vocabulary and applied linguistics* (126-132). Oxford: Macmillan.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (20-34). Cambridge: Cambridge University Press.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In: J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition* (20-34). Cambridge: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19(2), 255-271.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307-22.

Laufer, B. and Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33-51.

Lee, D. (2002). Notes to accompany the BNC Word Edition (bibliographical) index. Available at http://www.uow.edu.au/~dlee/home/BNCWIndexNotes.pdf (accessed 18 November 2009).

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus.* Harlow: Longman.

Lessard-Clouston, M. (2006). Breadth and depth of specialized vocabulary learning in theology among native and non-native English speakers. *The Canadian Modern Language Review,* 63(2), 175-198.

Lewis, M. (1993). *The Lexical Approach: The state of ELT and a way forward.* Hove, UK: LTP.

Lewis, M. (1997). *Implementing the Lexical Approach.* Hove, UK: LTP.

Lewis, M. (2000). *Teaching collocation.* Hove: Language Teaching Publications.

Lewis, M. (2008). The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency and proficiency. Unpublished doctoral thesis, Stockholm: Stockholm University.

Liontas, J. I. (2003). Killing two birds with one stone: Understanding Spanish VP idioms in and out of context. *Hispania, 86* (2), 289-301.

Liontas, J. I. (2007). The eye never sees what the brain understands: Making sense of idioms in second languages. *Lingua et Linguistica, 1* (2), 25-44.

Liontas, J.I. (2002). Context and idiom understanding in second languages. *EUROSLA Yearbook, 2,* 155-185.

Littlemore, J. (2001). Metaphoric competence: A language learning strength of students with a holistic cognitive style? *TESOL Quarterly,* 33(3), 459-491.

Littlemore, J. and Low, G. (2006). Metaphoric competence: Secong language learning, and communicative language abilitiy. *Applied Linguistics,* 27(2), 268-294.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly,* 37(4), 671-700.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. Schwarz & L. Bennett (Eds.), *Applied metacognition* (39-67). Cambridge: Cambridge University Press.

Manning, C.D. and Schutze, H. (1999). *Foundations of statistical national language processing.* Cambridge, MA: MIT Press.

Martinez, R. and Murphy, V. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly,* 45(2), 267-290.

Matthews, P. (1997). *The concise Oxford dictionary of linguistics.* Oxford: Oxford University Press.

McCarthy, M., McCarten, J. and Sandiford, H. (2005). *Touchstone*. Cambridge: Cambridge University Press.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, and J. Williams (Eds.), *Performance and competence in second language acquisition* (35-53). Cambridge: Cambridge University Press.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, 18(4), 393-407.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition.* Amsterdam: John Benjamins Publishing Company.

Meara, P. and Milton, J. (2003). *X_Lex, The Swansea Levels Test*. Newbury: Express.

Meara, P., and Jones, G. (1990). Eurocentres Vocabulary Size Test 10Ka (version E1 1/K10 MSDOS), Zurich Eurocentres.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

Messick, S. (1989). Validity. In R.L. Lin (Ed.), *Educational measurement* (13-103). New York: American Council on Education.

Meunier, F. and Granger, S. (Eds.) (2008). *Phraseology in foreign language learning and teaching.* Amsterdam: John Benjamins Publishing Company.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Cambridge ESOL Research Notes*, 37, 2-5.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, 32(2), 129-148.

Miller, G.A. (1999). On knowing a word. *Annual Review of Psychology*, 50(1), 1-19.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Oxford University Press.

Moon, R. (1987). The analysis of meaning. In J.M. Sinclair (Ed.), *Looking Up: An account of the COBUILD Project in lexical computing* (86-103). London: Harper Collins.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds), *Vocabulary: Description, acquisition and pedagogy* (40-63). Cambridge: Cambridge University Press.

Moore, D., Lin-Agler, L. M., & Zabrucky, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology*, 26(3), 251-265.

Morrison, L. (2004). Comprehension monitoring in first and second language reading. *The Canadian Modern Language Review,* 61(1), 77-106.

Nagy, W.E., Anderson, R.C., Schommer, M., Scott, J.A. and Stallman, J. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly,* 24(3), 262-282.

Nagy, W.E., Herman, P.A., and Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly,* 20(2), 233-253.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines*, 5, 12-25.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I.S.P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher,* 31(7), 9-13.

Nation, I.S.P. (2004). A study of the most frequent word families in the British National Corpus. In B. Laufer and P. Bogaards (Eds.), Vocabulary in a Second Language: Selection, Acquisition, and Tesing, (3-13). Amsterdam: John Benjamins Publishing Company.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review,* 63(1), 59-82.

Nation, I.S.P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (6-19). Cambridge: Cambridge University Press.

Nation, P. (1997). The language learning benefits of extensive reading. *Language Teacher,* 21, 13-16.

Nation, P. and Wang, K.M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language,* 12(2), 355-380.

Nattinger, J.R. and DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, 38(1/2), 1-135.

Nesselhauf, N. (2004). What are collocations? In D.J. Allerton, N. Nesselhauf, and P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (1-21). Basel: Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam: John Benjamins Publishing Company.

O'Dell, F. and McCarthy, M. (2008). *English collocations in use.* Cambridge: Cambridge University Press.

O'Keeffe, A., McCarthy, M., and Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

O'Sullivan, B. and Weir, C.J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (13-32). Basingstoke: Pelgrave Macmillan.

Oh, S. Y. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly*, 35(1), 69-96.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 2: Acquisition, loss, psychological reality, and functional explanations* (375-386). Amsterdam: John Benjamins Publishing Company.

Ostyn, P. and Godin, P. (1985). RALEX: An alternative approach to language teaching. *Modern Language Journal*, 69(4), 347-355.

Pareto, V. (1935). *The mind and society*. New York: Harcourt Brace.

Paribakht, T.S. and Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for Pedagogy* (174-199). New York: Cambridge University Press.

Paul, P.V., Stallman, A.C. and O'Rourke, J.P. (1990). Using three test formats to assess good and poor readers' word knowledge. Center for the Study of Reading Technical Report No. 509. Champaign, IL: University of Illinois.

Pawley, A. (2009). Grammarians' languages versus humanists' languages and the place of speech act formulas in models of linguistic competence. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 1: Distribution and historical change* (3-26). Amsterdam: John Benjamins Publishing Company.

Pawley, A. and Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C Richards and R.W. Schmidt (Eds.) *Language and communication* (191-225). London: Longman.

Pellicer-Sanchez, A., and Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, 22(1), 31-55.

Pennycook, A. (2001). *Critical applied linguistics: A critical introduction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1-28.

Pinker, S. (2007). *The stuff of thought*. New York: Penguin.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, 56(2), 282-307.

Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: As assessment perspective. *Language Learning*, 52(3), 513-536.

Rayson, P. (2008). Software demonstration: Identification of multi-word expressions with WMatrix. Paper presented at the *Formulaic Language Research Network (FLaRN) Third International Postgraduate Conference*, 19 June 2008, Nottingham, UK.

Read, J. (1988). Measuring vocabulary knowledge of second language learners. *RELC Journal*, 19(2), 12-25.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, 24, 146-161.

Read, J. and Nation, P. (2004). Measurement of formulaic sequences. In Schmitt, N. (Ed.), *Formulaic sequences* (23-36). Amsterdam: John Benjamins Publishing Company.

Renandya, W.A. and Jacobs, G.M. (2002). Extensive reading: Why aren't we all doing it? In: J.C. Richards and W.A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 295-302). Cambridge: Cambridge University Press.

Richards, J.C. (2003). *Interchange Level 1*. Cambridge: Cambridge University Press.

Richards, J.C. and Schmidt, R. (2002). *Dictionary of language teaching and applied linguistics*. London: Longman.

Rott, S. (1999). The effect of exposure frequency on intermediate language learner's incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 21(4), 589-619.

Salamoura, A. (2008). Aligning English Profile research data to the CEFR. *Cambridge ESOL Research Notes*, 33, 5-7.

Sarac, S., & Tarhan, B. (2009). Calibration of comprehension and performance in L2 reading. *International Electronic Journal of Elementary Education*, *2*(1), 167-179.

Saville, N. and Hawkey, R. (2010). The English Profile Programme – the first three years. *English Profile Journal*, 1(e7), 1-14.

Schmidt, R. and Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In: R.R. DAY, (Ed.), *Talking to learn conversation in second language acquisition* (pp. 237-322). Rowley, MA: Newbury House.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2004). *Formulaic sequences.* Amsterdam: John Benjamins.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research,* 12(3), 329-363.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Schmitt, N. and Carter, R. (2004). Formulaic sequences in action: An introduction. In Schmitt, N. (Ed.) *Formulaic Sequences* (1-22). Amsterdam: John Benjamins.

Schmitt, N. and Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research,* 15(4), 389-411.

Schmitt, N. and Meara, P. (1997). Researching vocabulary through a word knowledge framework: word associations and verbal suffixes. *Studies in Second Language Acquisition,* 20, 17-36.

Schmitt, N., Dörnyei, Z., Adolphs, S. and Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In: N. Schmitt, (Ed.), *Formulaic sequences* (55-86). Amsterdam: John Benjamins.

Schmitt, N., Jiang, X., and Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal,* 95, 26-43.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Schonell, F.J., Meddleton, I.G., Shaw, B.A., Routh, M. (1956). *A study of the oral vocabulary of adults.* Brisbane: University of Queensland Press.

Segalowitz, N. (2003). Automaticity in second languages. In C.J. Doughty and M.H. Long (Eds.), *The handbook of second language acquisition* (382-408). Malden, MA: Blackwell Publishing.

Segalowitz, S.J., and Korri, C.L. (2000). Lexical access of function versus content words. *Brain and Language,* 75(3), 376-389.

Seidelhofer, B. (2005). English as a lingua franca. *ELT Journal,* 59(4), 139-141.

Shillaw, J. (1996). The application of Rasch modeling to Yes/No vocabulary tests. Available at http://www.lognostics.co.uk/vlibrary/index.htm

Shillaw, J. (2009). Putting Yes/No tests in context. In T. Fitzpatrick and A. Barfield (Eds.) *Lexical processing in second language learners* (13-25). Bristol: Multilingual Matters.

Shin, D. and Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, 62(4), 339-348.

Shiotsu, T. and Weir, C.J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, 24(1), 99-128.

Simpson-Vlach, R. and Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31,* 487-512.

Sinclair, J. (1987). Looking Up: An account of the COBUILD Project in lexical computing. London: Harper Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers . *Second Language Research, 27,* 1-22.

Sonbul, S. and Schmitt, N. (2010). Direct teaching of vocabulary after reading: Is it worth the effort? *Applied Linguistics*, 64(3), 253-260.

Sorhus, H. (1977). To hear ourselves—Implications for teaching English as a second language. *ELT Journal*, 31, 211-21.

Sosa, A.V. and MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, 83, 227-236.

Spöttl, C., & McCarthy, M. (2003). Formulaic utterances in the multi-lingual context. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *The multilingual lexicon* (133-151). Dordrecht, Netherlands: Kluwer.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Stewart, J. and White, D.A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, 45(2), 370-380.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type descriminators: A quantitative method and some findings. *Functions of Language,* 10(1), 61-104.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In: N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (156-180). Cambridge: Cambridge University Press.

Tannen, D. and Oztek, P. (1981) Health to our mouths: Formulaic expressions in Turkish and Greek. In F. Coulmas (Ed.), *Conversational Routine: Explorations in standardized communication situations and prepatterned speech* (31-48). The Hague: Mouton De Gruyter.

Taylor, J.R. (2006). Polysemy and the lexicon. In G. Kristiansen, M. Achard, R. Dirven, and F.J. Ruiz de Mendoza Ibanez (Eds.), *Cognitive linguistics: Current applications and future perspectives* (51-80). Berlin: Mouton de Gruyter.

Thorndike, E.L. (1921). Word knowledge in the elementary school. *Teacher College Record*, 22(4), 334-370.

Thorndike, E.L. (1931). *A teacher's word book of the 20,000 words found most frequently and widely in general reading for children and young people.* New York: Teachers College, Columbia University.

Thorndike, E.L. and Lorge, L. (1944). *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Tremblay, A., Baayen, R.H., Libben, G. and Derwing, B. (2008). Lexical bundles and working memory: an ERP study. Paper presented at the *Formulaic Language Research Network (FLaRN) Third International Postgraduate Conference,* 19 June, 2008, Nottingham, UK.

Underwood, G., Schmitt, N. amd Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (153-173). Amsterdam: John Benjamins.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22(2), 217-234.

Walter, E., Woodford, K, & Good, M. (Eds.). (2008). *Cambridge advanced learner's dictionary* (3rd ed.). Cambridge, UK: Cambridge University Press.

Waring, R. and Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 51(2), 130-163.

Waring, R., 2008. The missing piece of the puzzle. Paper presented at the *11th Braz-Tesol National Convention,* July 17 2008, Fortaleza, Brazil.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33-52.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1). 46-65.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(1), 79-95.

Weinert, R. (1995). The role of formulaic language in second language acquisition: A review. *Applied Linguistics,* 16(2), 180-205.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York: Palgrave Macmillan.

West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.

West, M. (1954). Vocabulary selection and the minimum adequate vocabulary. *ELT Journal*, VIII(4), 121-126.

Wilkins, D. (1976). *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development.* Oxford: Oxford University Press.

Woolard, G. (2000). Collocation - encouraging learner independence. In: M. Lewis (Ed.), *Teaching collocation: Further developments in the Lexical Approach* (28-46). Hove: LTP.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics,* 21(4), 463-489.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 1: Distribution and historical change* (27-51). Amsterdam: John Benjamins Publishing Company.

Wray, A. and Namba, K. (2003). Formulaic language in a Japanese-English bilingual child: A practical approach to data analysis. *Japan Journal for Multilingualism and Multiculturalism,* 9(1), 24-51.

Wray, A. and Perkins, M.R. (2000). The functions of formulaic language: an integrated model. *Language & Communication,* 20, 1-28.

Wulff, S. (2008). *Rethinking idiomaticity.* London: Continuum.

Wulff, S., Swales, J.M., and Keller, K. (2009). "We have about seven minutes for questions": The discussion sessions from a specialized conference. *English for Academic Purposes,* 28, 79-92.

Xue, G. and Nation, I.S.P. (1984). A University Word List. *Language Learning and Communication,* 3, 215-229.

Zahar, R., Cobb, T. and Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review,* 57(4), 541-572.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research,* 23(2), 123-153.

Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A. and Healy, N.A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior,* 27, 201-212.

Zempleni, M., Haverkort, M., Renken, R., and Stowe, L.A. (2007). Evidence for bilateral involvement in idiom comprehension: An fMRI study. *NeuroImage*, 34(3), 1280-1291.

Zgusta, L. (1967). Multiword lexical units. *Word*, 23, 578-587.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Oxford: Addison-Wesley.

OTHER REFERENCES

Barlow, M. (1999). MonoConc Pro. Houston: Athelstan.

British National Corpus (version 3) (BNC XML edition) [CD-ROM]. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/

Cobb, T. Web Vocabprofile. [accessed 17 March, 2008 from http://www.lextutor.ca/vp/], an adaption of A. Heatley and P. Nation, Range (1994).

Appendix 1 – The PHRASE List (final version)

Phrasal expressions divided to match 1K frequency bands of the most common word families in the BNC. The 'Integrated List Rank' represents where each item falls when both lists (individual and phrase lists) are merged together. There are three categories of genre with frequency information to help the list user discern the appropriateness/usefulness of each phrase. The frequency information breaks down as follows:

* * *	=	phrase most common in this genre (or as common)
* *	=	phrase less common in this genre
*	=	phrase infrequent in this genre
X	=	phrase rare or non-existent in this genre

Integrated List Rank	Phrase	Frequency (per 100 million)	Spoken general	Written general	Written academic	Example
107	**HAVE TO**	83092	* * *	* *	*	I exercise because **I have to.**
165	**THERE IS/ARE**	59833	* * *	* * *	* *	**There are** some problems.
415	**SUCH AS**	30857	*	* * *	* * *	We have questions, **such as** how it happened.
463	**GOING TO (FUTURE)**	28259	* * *	* *	X	I'm **going to** think about it.
483	**OF COURSE**	26966	* * *	* *	*	He said he'd come **of course.**
489	**A FEW**	26451	* * *	* *	*	After **a few** drinks, she started to dance.
518	**AT LEAST**	25034	* * *	* *	* *	Well, you could email me **at least.**
551	**SUCH A(N)**	23894	* * *	* *	*	She had **such a** strange sense of humor.
556	**I MEAN**	23616	* * *	X	X	It's fine, but, **I mean,** is it worth the price?
598	**A LOT**	22332	* * *	*	X	They go camping **a lot** in the summer.

631	**RATHER THAN**	21085	* * *	* * *	* * *	Children, **rather than** adults, tend to learn quickly.
635	**SO THAT**	20966	* * *	* *	*	Park it **so that** the wheels are curbed.
655	**A LITTLE**	20296	* * *	* *	*	I like to work out **a little** before dinner.
674	**A BIT (OF)**	19618	* * *	* *	x	There was **a bit** of drama today at the office.
717	**AS WELL AS**	18041	* * *	* * *	* * *	She jogs **as well as** swims.
803	**IN FACT**	15983	* * *	* * *	* *	The researchers tried several approaches, **in fact.**
807	**(BE) LIKELY TO**	15854	* * *	* * *	* * *	To be honest, I'm **likely to** forget.
825	**GO ON**	15610	* * *	* * *	*	He **went on** for a while before stopping for lunch.
845	**IS TO**	15232	x	* * *	x	Obama **is to** address the media this afternoon.
854	**A NUMBER OF**	15090	* *	* * *	* * *	**A number of** concerns were raised.
879	**AT ALL**	14650	* * *	* * *	* *	Do you have any kids **at all?**
888	**AS IF**	14470	* *	* * *	x	They walked together **as if** no time had passed.
892	**USED TO (PAST)**	14411	* * *	*	x	It **used to** snow much more often.
894	**WAS TO**	14366	x	* * *	* *	The message **was to** be transmitted worldwide.
908	**NOT ONLY**	14110	* * *	* * *	* * *	**Not only** was it cheap, it was delicious.
913	**THOSE WHO**	13951	x	* * *	* * *	He would defend **those who** had no voice.
934	**DEAL WITH**	13634	* * *	* * *	* *	The police had several issues to **deal with.**
939	**LEAD TO ('CAUSE')**	13555	*	* * *	* *	Excessive smoking can **lead to** heart disease.
951	**SORT OF**	13361	* * *	x	x	It's **sort of** why I'm here.
974	**THE FOLLOWING**	12963	x	*	* * *	He made **the following** remarks.
984	**IN ORDER TO**	12762	* *	* * *	* * *	We shared a room **in order to** reduce costs
988	**HAVE GOT (+NP)**	12734	* * *	*	x	I don't know what he **has got** planned.

---end of 1k

1033	**HAVE GOT TO**	12270	* * *	* *	x	You **have got to** try this salad.

1079	**SET UP**	11560	* *	* * *	* *	The whole thing was **set up** beforehand.
1082	**AS TO**	11535	*	* * *	* * *	There was some confusion **as to** its whereabouts.
1083	**AS WELL**	11519	* * *	* * *	* * *	I like it **as well.**
1088	**BASED ON**	11440	* *	* * *	* * *	**Based on** the reports, it seems it was an accident.
1146	**CARRY OUT**	10753	*	* *	* * *	It was not as easy to **carry out** without funding.
1169	**TAKE PLACE**	10556	* * *	* * *	* *	No one was sure exactly why it **took place** there.
1175	**TEND TO**	10504	* * *	* * *	* * *	**I tend to** think it's actually a political matter.
1183	**DUE TO**	10454	* * *	* * *	* * *	Many people believe it is **due to** global warming.
1204	**FAIL TO**	10263	* * *	* * *	* * *	**I fail to** see the humor in all this.
1209	**EACH OTHER**	10160	* * *	* *	*	The good thing is that they still have **each other.**
1234	**IN TERMS OF**	9881	* * *	* * *	* * *	What is your limit, **in terms of** price?
1262	**NO ONE**	9597	* * *	* *	*	I will speak to **no one** about this.
1299	**PICK UP**	9252	* * *	* *	X	She dropped by to **pick up** her friend.
1352	**UP TO (MAXIMUM)**	8733	* * *	* * *	* * *	**Up to** twenty people may be interviewed for the job.
1356	**A SINGLE ('ANY')**	8710	* * *	* *	*	Not **a single** idea emerged from the meeting.
1371	**NO LONGER**	8556	* *	* * *	* *	Fortunately, it is **no longer** a concern now.
1387	**LOOK FOR**	8377	* * *	* *	*	Bargains can be found if you know what to **look for.**
1427	**LAST NIGHT**	7992	* * *	*	X	What did he say **last night?**
1438	**AS A RESULT**	7939	*	* *	* * *	He was tired and **as a result** not very aware.
1453	**IN ADDITION (TO)**	7822	* *	* * *	* * *	The house was well located **in addition.**
1487	**WORK ON**	7600	* * *	* *	*	That kid needs to **work on** his attitude.
1517	**THINK ABOUT**	7243	* * *	* *	X	I'm **thinking about** changing careers.
1529	**FOR INSTANCE**	7138	*	* * *	* * *	That holds true for even the governor **for instance.**
1530	**TOO MUCH**	7123	* * *	* *	*	I don't worry about it **too much.**

1532	**YOU SEE**	7102	***	X	X	And the problem **you see** is with awareness-raising.
1533	**IN PARTICULAR**	7092	**	***	***	There's nothing **in particular** wrong with it.
1542	**A COUPLE OF**	7007	***	**	X	They need **a couple of** minutes still.
1556	**INSTEAD OF**	6907	***	***	**	He can go **instead of** me.
1573	**COME BACK**	6772	***	***	***	But we can **come back** to that later.
1591	**LOOK LIKE**	6595	***	**	X	It's not what it **looks like.**
1605	**FIND OUT**	6499	***	***	**	How did you **find out?**
1638	**POINT OUT**	6325	***	***	***	She **pointed out** that it was getting late.
1647	**APART FROM**	6287	***	***	***	Are there any others, **apart from** him?
1657	**CALL FOR**	6243	*	***	**	Hard times **call for** tough measures.
1661	**MANAGE TO**	6234	***	***	**	Did you **manage to** get in touch at all?
1667	**OR TWO**	6192	***	**	*	He had a thing **or two** to drink, then left.
1681	**A FURTHER ('ANOTHER')**	6121	*	**	***	It's **a further** reason to reconsider the project.
1696	**COME OUT**	6031	***	***	***	It's supposed to **come out** on Friday.
1713	**BE EXPECTED TO**	5964	***	***	**	They can't **be expected to** just sit there and wait.
1716	**SEEK TO**	5937	*	***	***	The new board **sought to** find alternative solutions.
1733	**GO THROUGH**	5857	***	***	**	You can't imagine what I'm **going through.**
1740	**LONG TERM**	5831	***	***	***	It seems to work, but what of the **long term** effects?
1759	**RESULT IN**	5763	*	**	***	Excessive criticism **resulted in** feelings of animosity.
1764	**THAT IS (REPHRASING)**	5737	***	***	***	It's yours, as long as you pay **that is.**
1778	**EVEN THOUGH**	5664	***	***	***	I can't **even though** it looks easy enough.
1779	**A RANGE OF**	5651	***	***	***	Eventually, **a range of** events changed that.
1815	**THE LATTER**	5519	X	**	***	**The latter** may still be a challenge.

	(anaphor)					
1818	**MAKE SURE**	5510	* * *	* * *	* *	I had to **make sure** before I opened the door.
1843	**TAKE OVER**	5394	* * *	* * *	* *	You can't just let the kids **take over.**
1850	**CONSIST OF**	5362	*	* * *	* * *	What does it **consist of?**
1861	**AS SOON AS**	5323	* * *	* * *	* *	**As soon as** he can he will call you back.
1873	**AT THE TIME ('WHEN THIS HAPPENED')**	5282	* * *	* * *	* *	I was busy **at the time.**
1877	**ON THE OTHER HAND**	5267	* * *	* * *	* *	**On the other hand,** the business did make a profit.
1886	**ON ONE'S OWN**	5240	* * *	* * *	*	You did this **on your own?**
1894	**ALL RIGHT**	5230	* * *	* *	x	If it's **all right** I think I'll head off to bed.
1900	**SUBJECT TO**	5218	x	*	* * *	All baggage is **subject to** inspection.
1904	**AFTER ALL (adv.)**	5197	* * *	* * *	*	He is only six years old, **after all.**
1908	**IN FRONT OF ('BEFORE')**	5190	* * *	* * *	* *	She did not want to say anything **in front of** the kids.
1910	**TO DO WITH**	5184	* * *	*	*	I think it has something **to do with** physics.
1912	**GO OUT**	5173	* * *	* * *	*	Brochures **went out** to prospective buyers.
1920	**A GOOD/GREAT DEAL ('MUCH')**	5126	* * *	* * *	*	She means **a great deal** to me.
1929	**ON THE WAY**	5085	* * *	* *	*	We can stop for lunch **on the way.**
1931	**AS LONG AS**	5084	* * *	* * *	*	It makes no difference **as long as** it's done.
1951	**SO FAR ('UNTIL NOW')**	5018	* * *	* * *	* * *	Any questions **so far?**
1958	**OUGHT TO**	5002	* * *	*	x	She hasn't but she **ought to.**
1959	**AT THE MOMENT**	5001	* * *	* * *	* *	They have a lot going on **at the moment.**
1967	**AS THOUGH**	4988	*	* * *	* *	He smiled then, **as though** remembering a joke.
1970	**COME TO ('EVOLVE TO')**	4970	* *	* * *	*	We **came to** see it for what it was.
1974	**ALONG WITH**	4948	* *	* * *	* *	**Along with** his dog, the cat slowly stopped eating.

1982	**MAY WELL ('COULD')**	4931	* * *	* * *	* * *	He hasn't yet but **may well** try before long.
2001	**GET OUT**	4858	* * *	* *	x	There was no way to **get out** of work that Friday.
2013	**FOLLOWED BY**	4816	*	* * *	* * *	The workshop will be **followed by** time for questions.
2018	**IN (THE SENSE) THAT**	4805	* * *	* * *	* * *	It's great **in that** there are so many restaurants.
2022	**THE CASE ('TRUE')**	4794	* *	* * *	* * *	It's not **the case** that I don't love him.
2038	**TAKE UP**	4717	* * *	* *	*	I don't wish to **take up** more time than is necessary.
2060	**ACCOUNT FOR**	4642	*	* * *	* * *	They'll need to **account for** their actions.
2064	**SET OUT**	4624	*	* * *	* *	She accomplished what she **set out** to do.
2067	**AS FAR AS**	4619	* * *	* * *	* *	What are you thinking, **as far as** feasibility on this?
2068	**CONCERNED WITH**	4619	*	* * *	* * *	They spoke on issues **concerned with** culture.
2075	**ABOUT TO**	4600	* * *	* * *	*	Things were **about to** change.
2086	**SUPPOSED TO**	4586	* * *	* *	* *	I didn't go, but I was **supposed to.**
2087	**AND SO ON**	4584	* * *	* * *	* * *	My parents arrived, then John, his wife, **and so on.**
2105	**COME ON**	4519	* * *	*	x	**Come on,** think about it!

--end of 2k

2124	**TAKE ON**	4451	* *	* * *	*	You'd better think twice before **taking on** more work.
2131	**WORK OUT**	4432	* * *	* *	* *	I'm trying to **work out** what it stands for.

2135	**ALL OVER ('EVERYWHERE')**	4420	* * *	* * *	*	The news was **all over** the web.
2149	**OTHER THAN**	4380	* *	* * *	* * *	No one **other than** you could have come up with it.
2156	**OUT OF ('IN'/'FROM')**	4361	*	* * *	* * *	In terms of colour, three **out of** five were silver.
2157	**TURN OUT**	4354	* * *	* * *	* *	How did it **turn out?**
2164	**LOOK AFTER**	4332	* * *	* *	**X**	She did more than just **look after** the finances.
2173	**AT LAST**	4306	* *	* * *	*	**At last** we met over coffee last week.
2179	**A VARIETY OF**	4283	* *	* * *	* * *	And a **variety of** issues seemed to stall the process.
2180	**AT FIRST**	4275	* * *	* * *	* * *	Not **at first,** but just the other day there was one.
2216	**OR SO**	4164	* * *	* * *	* * *	A day **or so** later he called me back on my mobile.
2231	**IN FAVOUR**	4073	* * *	* * *	* * *	I don't know about you but I'm **in favour.**
2258	**GIVE UP**	3997	* * *	* * *	*	You can't just **give up.**
2263	**GET TO ('ARRIVE AT')**	3979	* * *	* *	*	When he **got to** the end he started to weep.
2280	**FIND ONESELF**	3917	* *	* * *	*	**I found myself** driving home.
2298	**GET UP**	3857	* * *	* *	**X**	It was easy **getting up** but the commute was awful.
2345	**CARRY ON**	3759	* * *	* * *	* *	He could simply not **carry on.**
2361	**GO BACK**	3722	* * *	* * *	* * *	If I could I'd **go back** and do it all over again.
2371	**FOCUS ON**	3703	* *	* * *	* * *	But today I'd like to **focus on** something different.
2379	**AT ONCE**	3684	* *	* * *	* *	I did it **at once.**
2384	**IT TAKES**	3670	* * *	* * *	* * *	No matter what **it takes.**
2388	**GET ON/OFF (TRANs)**	3656	* * *	* *	**X**	We waited for hours in the queue to **get on.**
2404	**AS A WHOLE**	3615	* *	* * *	* * *	**As a whole** it's OK.
2407	**IN PRACTICE**	3609	*	* * *	* * *	There was nothing she could do **in practice.**

2408	**BY THE TIME**	3607	* * *	* * *	*	**By the time** dinner started there were none left.
2409	**LOTS OF**	3605	* * *	* *	X	**Lots of** them do travel during the winter months.
2417	**SAID TO BE**	3586	X	* * *	*	The priest was **said to be** missing.
2425	**IN TIME**	3566	* *	* * *	* *	You will **in time.**
2429	**IN TURN**	3558	*	* *	* * *	The supermarket **in turn** will donate seven thousand.
2437	**ONCE AGAIN**	3532	* * *	* * *	* * *	**Once again**, this was completely unforeseen.
2439	**ALL THE TIME**	3527	* * *	* *	*	I go there **all the time.**
2442	**ON THE BASIS (OF)**	3515	*	* * *	* * *	She came **on the basis** that it would help her.
2445	**KIND OF**	3510	* * *	*	X	The windows are **kind of** fogged up.
2446	**GET INTO**	3508	* * *	* *	*	I really **got into** what he was talking about.
2454	**RELY ON**	3488	* * *	* * *	* * *	He has no one else to **rely on.**
2483	**GO FOR**	3421	* * *	*	X	I could really **go for** a hamburger right now.
2484	**AIM TO**	3415	*	* * *	* * *	They **aim to** complete the project by the spring.
2498	**MAKE UP ('COMPRISE')**	3394	*	* * *	* * *	Hispanics **make up** a large part of the population.
2531	**APPEAL TO**	3299	* *	* * *	* *	It **appeals to** younger learners.
2536	**END UP**	3285	* * *	* * *	* *	We **ended up** going anyway.
2548	**SHAKE ONE'S HEAD**	3250	*	* * *	X	He just **shook his head** and laughed.
2557	**NO MORE THAN ('ONLY')**	3226	*	* *	* * *	**No more than** five of them had auto insurance.
2575	**GET BACK**	3178	* * *	* *	*	What time will we **get back?**
2584	**WHAT ABOUT**	3160	* * *	* *	X	But **what about** tomorrow?
2586	**IN OTHER WORDS**	3159	* * *	* * *	* * *	The groups are **in other words** conservative.
2588	**AS FOR**	3157	* *	* * *	* *	**As for** the promotion, there may be other options.

2599	**NOT EVEN**	3128	* * *	* * *	* * *	No one cared, **not even** when they saw the photos.
2606	**ENTITLED TO**	3113	*	* * *	* * *	You're **entitled to** your opinion.
2608	**PRIOR TO**	3110	*	* * *	* * *	**Prior to** the event, the organizers called the band.
2616	**CHOOSE TO**	3099	* *	* * *	* * *	You may **choose to** believe what the papers say.
2622	**SOMETHING LIKE ('AROUND')**	3092	* * *	*	x	She makes **something like** five grand a month.
2624	**KNOWN TO**	3091	x	* * *	* * *	He's **known to** be like that.
2636	**IN TOUCH (WITH)**	3060	* * *	* * *	*	Are you two at all **in touch?**
2637	**IN THE END**	3050	* * *	* * *	* *	**In the end** it's like what they say about horses.
2654	**IN THE WAY**	3013	* * *	* *	x	I won't stand **in the way.**
2657	**CARE FOR**	3004	* * *	* *	*	He doesn't **care for** it very much in his tea.
2658	**IN THE EVENT (OF)**	2998	x	* * *	* * *	**In the event** you change your mind, let me know.
2675	**THEY SAY**	2962	* * *	*	x	**They say** it tastes like chicken.
2688	**SO CALLED**	2944	* *	* * *	* * *	Any **so called** rumors are as easily their fault.
2702	**TAKE INTO ACCOUNT**	2921	*	* * *	* * *	You must also **take into account** the rush hour.
2709	**IN RESPECT OF**	2909	x	* * *	* * *	There was nothing **in respect of** drink.
2711	**OUT OF ('USING')**	2907	* * *	* * *	* *	The walls are made **out of** wood.
2719	**AT THE SAME TIME ('CONVERSELY')**	2892	* *	* * *	* * *	**At the same time** it may be worth it.
2724	**NEXT TO**	2882	* * *	* * *	*	**Next to** his sister, he was the best looking of them.
2734	**TURN UP**	2865	* * *	*	x	The money **turned up** later.
2735	**POINT OF VIEW**	2864	* *	* * *	* * *	They all had a different **point of view.**
2744	**AT PRESENT**	2847	*	* * *	* * *	There's nothing to do **at present.**

2752	**USED TO (ACCUSTOMED)**	2831	* * *	* * *	*	It may take a while to get **used to.**
2759	**WHETHER OR NOT**	2824	*	* * *	* * *	It depends on **whether or not** he comes on time.
2767	**IN PLACE**	2805	*	* * *	* * *	There are systems **in place** to handle that.
2776	**NO DOUBT ('SURELY')**	2791	* *	* * *	*	She can and will **no doubt** improve over time.
2785	**FULL TIME**	2761	* * *	* * *	*	I'm a **full time** fan of his music.
2817	**SORT OUT**	2696	* * *	* *	* *	They managed to **sort out** everything on their own.
2823	**IN A WAY**	2684	* * *	* * *	*	It's funny **in a way.**
2826	**OR SOMETHING ('PERHAPS')**	2683	* * *	*	X	The sales clerk seemed distracted **or something.**
2827	**ON BEHALF OF**	2679	X	*	* * *	I'd like to apologize **on behalf of** the committee.
2828	**OVER THERE**	2678	* * *	*	X	He put it way **over there.**
2829	**IN SPITE OF**	2676	*	* * *	* * *	**In spite of** all the work there were few alterations.
2831	**THAT'S IT**	2674	* * *	*	X	**That's it** for today, I'm afraid.
2848	**IN PART**	2652	X	* * *	* * *	It is **in part** the reason people come here.
2857	**OH NO**	2642	* * *	X	X	**Oh no** not again.
2859	**IN MIND**	2638	* * *	* * *	* *	Do you have anything special **in mind?**
2863	**(WITH) REGARD TO**	2630	*	* * *	* * *	There was some confusion **with regard to** payment.
2868	**ONE ANOTHER**	2623	*	* * *	*	They looked at **one another** for a few minutes.
2871	**AS FOLLOWS**	2620	X	* *	* * *	The plan is **as follows.**
2879	**THE ABOVE**	2608	X	* * *	* * *	**The above** only underscores strength of the data.
2888	**TO DATE**	2600	*	* * *	* * *	**To date** there have been over nine instances.
2892	**GO INTO**	2595	* * *	* *	*	I won't **go into** what he said.
2898	**TOO MANY**	2587	* * *	* * *	* * *	You don't see **too many** like that one, I'll tell you.

2899	**IN THE COURSE OF**	2585	*	* * *	* * *	**In the course of** the discussion the manager left.
2903	**MORE OR LESS**	2579	* * *	* * *	* * *	It's **more or less** what we imagined I suppose.
2906	**SHORT TERM**	2574	* * *	* * *	* * *	There are many **short term** gains to be had.
2907	**AIMED AT**	2573	*	* * *	* * *	The study is **aimed at** exploring how people use it.
2919	**GO OFF**	2551	*	* * *	X	She **went off** without even saying goodbye.
2926	**IN CASE**	2536	* * *	* * *	* * *	**In case** you're wondering, this isn't my natural color.
2941	**OUT THERE**	2513	* * *	* *	X	There are simply no jobs **out there** right now.
2942	**LED BY**	2511	X	* * *	* *	He was **led by** his competitive drive above all.
2967	**MORE AND MORE**	2468	* * *	* * *	* *	**More and more** it's about the customer.
2970	**HAVE A LOOK**	2464	* * *	*	X	Why don't you **have a look?**
2981	**BELIEVE IN**	2450	* * *	* * *	x	I don't **believe in** corporal punishment.
2983	**PUT IT ('SAY')**	2449	* * *	* *	*	There's no better way to **put it,** I think.
2992	**THESE DAYS**	2440	* * *	* * *	*	It's what **these days** the media call 'viral'.
2999	**IN CHARGE**	2432	* * *	* * *	* * *	He was able to stay **in charge** while she was away.
3000	**FEEL LIKE**	2431	* * *	*	X	**I feel like** eating out.
3016	**UP TO (UNTIL)**	2409	* * *	* * *	* *	I walked **up to** the window and paused.
3022	**HEARD OF**	2403	* * *	* * *	*	Haven't you ever **heard of** manners?
3044	**TAKE PART IN**	2374	*	* * *	* * *	It's something we all wanted to **take part in.**
3059	**IN SO FAR AS**	2344	*	* *	* * *	The plan was fine **in so far as** time and transport.
3061	**PART TIME**	2343	* * *	* * *	*	She can take it on **part time.**
3067	**LOOK FORWARD TO**	2331	* *	* * *	X	It's something we can all **look forward to.**
3094	**AS SUCH**	2290	*	* * *	* * *	The film was not a horror **as such.**
3104	**BOUND TO**	2278	* * *	* * *	* * *	It's **bound to** be better next time.
3113	**TURN ON**	2266	* * *	* * *	*	It won't **turn on.**

3121	**SET TO**	2258	* * *	* * *	*	He's **set to** arrive today.
3134	**MOVE ON**	2245	* * *	* * *	*	I'd like to **move on** if I may.
3149	**IN CONTRAST (TO)**	2229	*	* * *	* * *	The inside was amazing **in contrast.**
3152	**THIS STAGE**	2223	* * *	* * *	* * *	We can't at **this stage.**
3157	**ALL BUT**	2214	* * *	* *	*	She **all but** sent him chocolates and flowers.
3160	**ABOVE ALL**	2212	* *	* * *	* * *	It is **above all** what people care most about.
3162	**RID OF**	2212	* * *	* * *	*	She was happy to be **rid of** it.
3197	**IN ANY CASE**	2159	* * *	* * *	* * *	It's not due till tomorrow **in any case.**
3199	**THANKS TO**	2159	* * *	* * *	* * *	And it's **thanks to** her research that we know that.
3205	**GO AWAY**	2150	* * *	* *	*	The problem won't just **go away.**
3207	**ONCE MORE**	2146	* *	* * *	* *	I call on you **once more** my fellow citizens.
3220	**OH WELL**	2129	* * *	x	x	It was due yesterday? **Oh well.**
3222	**FOLLOW UP**	2128	*	* *	* * *	I'd like to **follow up** on what that gentleman said.
3224	**WOULD SAY**	2128	* * *	* * *	* * *	I don't know, but I **would say** at least twenty.
3235	**FOUND TO**	2104	x	* *	* * *	It was **found to** be of little use in battle.
3240	**MEANT TO**	2098	* * *	* * *	* *	I haven't seen it but it's **meant to** be his best so far.

---end of 3k

3263	**HANG ON**	2074	* * *	* *	x	If you can **hang on** for just a second.
3265	**TURN INTO**	2072	* * *	* * *	*	The idea **turned into** something amazing.
3268	**SOMETHING ABOUT**	2071	* * *	* *	x	There's **something about** her I find

						interesting.
3295	BY NOW	2044	* * *	* * *	*	I thought they'd be here **by now.**
3308	THINK SO	2033	* * *	*	*	Do you **think so?**
3320	GO AHEAD	2023	* * *	* *	*	Why don't you **go ahead** and call him again?
3322	BRING ABOUT	2022	x	* * *	* *	It was expected that he would **bring about** change.
3323	HAD BETTER	2022	* * *	* * *	*	They **had better** listen.
3327	IN ACCORDANCE WITH	2017	x	* *	* * *	The plan was execute **in accordance with** municipal guidelines.
3328	CALL ON	2016	*	* * *	*	He **called on** his members of staff for advice.
3331	AT TIMES	2014	* * *	* * *	* * *	It can be slightly frustrating **at times.**
3340	ALL THE WAY	2007	* * *	* * *	* * *	I'm with you **all the way.**
3349	IN EFFECT	1995	* *	* * *	* * *	**In effect** it was like being in space.
3355	AFFORD TO	1989	* * *	* * *	*	It's something you can't **afford to** miss.
3357	SIGHT OF	1988	* * *	* * *	* *	I can't stand the **sight of** it, to tell you the truth.
3362	IN ADVANCE	1983	* * *	* * *	* * *	It's something you should do **in advance.**
3379	ON THE PART OF	1965	*	* * *	* * *	There are no barriers **on the part of** the government.
3387	BRING UP	1958	* * *	* *	* *	Did you manage to **bring up** the holiday pay?
3388	TAKE OFF	1957	* * *	* * *	*	I was hungry by the time we **took off.**
3390	SO AS TO	1954	* *	* * *	* * *	He cluttered his desk **so as to** appear busy.
3396	TAKE ADVANTAGE	1949	* * *	* * *	* * *	Some people just want to **take advantage.**
3399	SHORT OF	1944	*	* * *	* * *	**Short of** calling the doctor, I don't know what to do.
3402	OVER THE YEARS	1942	* * *	* * *	* * *	She has **over the years** visited many cities.
3406	SWITCH ON	1935	* * *	* *	*	It's plugged in but it won't **switch on.**
3417	BY NO MEANS	1925	* *	* * *	* * *	We have **by no means** concluded the matter.
3431	COULD HARDLY	1911	* * *	* * *	*	I **could hardly** wait for the weekend.
3446	COME UP WITH	1898	* * *	* *	*	Is that the best you could **come up with?**

3448	**IN QUESTION**	1898	x	*	* * *	The book **in question** was published last year.
3450	**IN THE FIRST PLACE**	1897	* * *	* *	* *	He couldn't see how it got there **in the first place.**
3456	**PROVE TO BE**	1893	*	* * *	* * *	The bicycle **proved to be** of immense value.
3457	**IN COMMON**	1890	* * *	* * *	* * *	What they have **in common** is their stamina.
3461	**NO MATTER**	1888	* * *	* * *	*	**No matter** what they say, it's down to you now.
3465	**AT THIS POINT**	1884	* *	* * *	* * *	**At this point** there's no telling who might win.
3491	**IN ITSELF**	1855	* *	* * *	* * *	The car **in itself** wasn't enough incentive.
3502	**THE FORMER**	1842	x	* * *	* * *	**The former** can actually require much more money.
3511	**IF ONLY**	1830	* *	* * *	*	**If only** I'd known sooner.
3525	**YET TO**	1818	* *	* * *	* * *	The package is **yet to** arrive.
3534	**UP TO (DECISIONS)**	1807	* * *	* * *	* *	I'll leave it **up to** you if that's all right.
3536	**OR WHATEVER**	1806	* * *	x	x	The challenge **or whatever** is doing it under budget.
3548	**HAND OVER**	1798	* * *	* *	*	The responsibility was expected to be **handed over.**
3554	**IN THE LIGHT OF**	1789	*	* * *	* * *	**In the light of** the results, we delayed the study.
3560	**IN THE SAME WAY**	1784	* *	* * *	* * *	**In the same way,** Apple appeals to the young.
3567	**THAT MUCH**	1781	* * *	*	x	I don't go out **that much.**
3601	**THE EXTENT TO WHICH**	1751	x	*	* * *	**The extent to which** the preceding can be asserted is largely dependent on what emerges in the study.
3632	**FOR SOME TIME**	1725	* *	* * *	* *	He's actually been studying **for some time** now.
3640	**IN RETURN (FOR)**	1720	* * *	* * *	* * *	**In return** I'd like to present you with this gift.
3650	**TO DEATH**	1714	* * *	* *	x	The children were bored **to death.**
3652	**ON THE GROUNDS**	1713	* *	* * *	* * *	The decision was made **on the grounds** of

						safety.
3670	**OH DEAR**	1700	* * *	X	X	He didn't? **Oh dear.**
3672	**IN FULL**	1698	X	* *	* * *	She paid **in full.**
3673	**ON BOARD**	1698	* * *	* * *	*	Are you **on board** with the decision?
3686	**TO SOME EXTENT**	1688	* * *	* * *	* * *	It is **to some extent** what is considered standard.
3695	**SOME KIND OF**	1678	* * *	* *	X	What are you, **some kind of** genius?
3700	**KEEP UP**	1674	* * *	* * *	*	**Keep up** if you can.
3709	**NO IDEA**	1670	* * *	*	X	You have **no idea.**
3712	**GREATER THAN**	1666	*	* *	* * *	Anything **greater than** five is fine.
3715	**HAPPEN TO (BE)**	1664	* * *	* *	X	This **happens to** be my first conference.
3718	**HELD THAT ('BELIEVED')**	1661	X	* *	* * *	The congress **held that** it was acceptable.
3722	**FACED WITH**	1658	*	* * *	*	**Faced with** defeat, he changed tactics.
3733	**DO(ING) SO**	1647	*	* * *	* * *	Unfortunately, **doing so** also meant facing traffic.
3738	**SET OFF**	1644	* *	* * *	X	We **set off** at noon.
3740	**PUT FORWARD**	1640	*	* *	* * *	The group **put forward** several other proposals.
3760	**FROM TIME TO TIME**	1627	* * *	* *	*	He does come round **from time to time.**
3791	**THE MEANS**	1626	*	* * *	* * *	He hasn't **the means** to get there.
3762	**EVER SINCE**	1622	* * *	* * *	*	And they haven't been the same **ever since.**
3763	**JUST ABOUT**	1622	* * *	*	X	Am I finished? **Just about.**
3771	**AS OPPOSED TO**	1615	* *	* * *	* * *	Better now, as opposed to what?
3774	**GIVE RISE TO**	1613	*	* * *	* * *	The protests **gave rise to** new violence.
3780	**LARGE SCALE**	1610	*	* * *	* * *	They were thinking on a **large scale.**
3782	**MAKE SENSE**	1608	* * *	* *	*	Doesn't it **make sense** to do it that way?
3789	**BY MEANS OF**	1602	*	* * *	* * *	It was possible to achieve **by means of** coercion.
3790	**IN SHORT**	1602	X	* *	* * *	The employees were **in short** tired of it.

3791	**A BIT OF A**	1599	* * *	*	x	He's **a bit of a** film fanatic.
3796	**BREAK UP**	1595	* * *	* *	x	They did keep in touch after they **broke up.**
3803	**BUT THEN (AGAIN)**	1589	* * *	* *	x	I went over budget, **but then** I expected I would.
3830	**ALL TOO**	1571	* * *	* * *	x	Why is this **all too** familiar?
3831	**PUT UP**	1571	* * *	* * *	x	The authorities **put up** a sign to notify the public.
3847	**GOOD AT**	1562	* * *	*	x	You need to find something you're **good at.**
3851	**A LONG WAY**	1557	* * *	* * *	*	A little bit of compassion can go **a long way.**
3854	**AMOUNT TO**	1556	* *	* * *	* * *	Unfortunately, the player never **amounted to** much.
3861	**FOR LONG**	1553	* * *	* * *	*	He did attend university, but not **for long.**
3883	**(BE) RUN BY**	1540	* *	* * *	* *	The operation is **run by** two college students.
3887	**SOME MORE**	1539	* * *	*	*	I'll have to think **some more** before deciding.
3888	**IN THE ABSENCE OF**	1538	x	* *	* * *	**In the absence of** truth, there is only conjecture.
3895	**ALL SORTS OF**	1535	* * *	* * *	*	We have **all sorts of** time left.
3896	**GET ON WITH**	1535	* * *	* * *	x	I need to **get on with** my work.
3902	**NO GOOD**	1528	* * *	* *	x	It's **no good** sitting there – you have to type.
3939	**YET ANOTHER**	1500	* *	* * *	* * *	That's **yet another** day gone by with nothing done.
3940	**KEY TO**	1499	* *	* * *	* * *	Her input was **key to** the project's success.
3947	**I'M AFRAID**	1495	* * *	*	x	It hasn't arrived**, I'm afraid.**
3955	**THAT WHICH**	1491	x	* * *	* * *	You cannot undo **that which** has already been done.
3960	**IF SO**	1488	* *	* * *	* * *	**If so**, you may want to consider Greece.
3973	**RIGHT NOW**	1482	* * *	* *	x	The best time is **right now.**
3979	**IN VIEW OF**	1477	x	* *	* * *	She reconsidered **in view of** her children.
3984	**IN DETAIL**	1473	* *	* * *	* * *	I can explain **in detail.**
3989	**REFLECTED IN**	1471	* *	* * *	* * *	The artist's angst is **reflected in** her painting.

3992	**NO SUCH**	1470	**	***	***	He said **no such** thing.
4004	**NOTHING BUT**	1466	***	***	*	I have **nothing but** respect for her.
4020	**IN THE FACE OF**	1457	*	***	***	They quit **in the face of** the media scrutiny.
4023	**SUCH THAT**	1454	*	***	***	It can be developed **such that** it does not interfere.
4025	**OUT OF ('DUE TO')**	1453	***	***	*	Most commuters drive **out of** pure necessity.
4029	**NEXT DOOR**	1449	***	***	X	The people **next door** didn't seem to mind.
4032	**TO THE POINT**	1447	***	***	***	It got **to the point** I couldn't breathe.
4035	**MAKE ITS/ONE'S WAY**	1446	*	***	X	The mouse **made its way** across the field.
4037	**IN HAND**	1443	**	***	X	It was well **in hand** by then.
4052	**GET TO ('OPPORTUNITY')**	1437	***	***	X	Did you **get to** try the cheese?
4058	**BY THE WAY**	1433	***	**	X	**By the way**, what day is the wedding?
4061	**BY CONTRAST**	1430	**	***	***	The weather was **by contrast** a pleasant surprise.
4062	**RUN OUT (OF)**	1430	***	**	X	You need to get milk before you **run out**.
4063	**IN PRINCIPLE**	1429	*	***	***	You can **in principle**.
4071	**ADD TO**	1424	**	***	***	It does nothing to **add to** the scenery.
4072	**AS YET**	1423	*	***	***	I'm not convinced **as yet**.
4077	**AT RISK**	1419	*	***	***	There's no reason to put passers by **at risk.**
4092	**A MERE**	1410	*	***	***	She made **a mere** penny per shirt.
4093	**SHOWN TO**	1409	X	**	***	Dogs have been **shown to** help.
4095	**ON THE ONE HAND**	1406	***	**	*	**On the one hand**, it's cheap.
4099	**BY WAY OF**	1404	*	***	**	It was only **by way of** bribery that they managed.
4107	**ON THE ROAD**	1400	**	***	X	I can't wait to get **on the road** again.
4109	**BEAR IN MIND**	1398	***	***	***	It was designed **bearing in mind** environmental impact.
4113	**OLD FASHIONED**	1393	***	***	X	He's just a bit **old fashioned** sometimes.

4136	**FOR SALE**	1379	*	***	X	The house is **for sale.**
4138	**OR ANYTHING**	1379	***	*	X	I don't want any trouble **or anything**.
4143	**MOST LIKELY**	1376	***	***	***	It is **most likely** the reason for the warming.
4153	**PROVIDE FOR**	1370	X	**	***	The other variables have been **provided for.**
4165	**EVEN SO**	1363	***	***	**	**Even so,** it's no way to treat a guest.
4167	**COME ACROSS**	1362	**	***	*	They happened to **come across** a map.
4182	**FIRST OF ALL**	1352	***	*	*	It is **first of all** what most people prefer.
4195	**MIGHT AS WELL**	1348	***	**	X	We're here now so we **might as well.**
4199	**LIMITED TO**	1346	X	**	***	The dinner was **limited to** seniors.
4201	**TO ME ('IN MY OPINION')**	1345	***	*	*	**To me** what matters is that it gets done.
4207	**MIND YOU**	1342	***	*	X	**Mind you,** that's one reason why I go there.
4210	**AT A TIME ('SIMULTANEOUSLY')**	1340	**	***	***	They went in two **at a time.**
4232	**HALF PAST**	1325	**	***	X	I make it **half past.**
4233	**WITH RESPECT TO**	1325	*	***	***	She had nothing to add **with respect to** the report.
4242	**CONSISTENT WITH**	1315	X	*	***	The results are **consistent with** our hypothesis.
4248	**WAY OUT**	1309	***	***	X	He lives **way out** in the country.
4250	**THIRD PARTY**	1307	X	X	***	They work with **third party** software too.
4253	**CONTRARY TO**	1303	*	**	***	The city is clean, **contrary to** popular belief.
4257	**WORTH OF**	1299	***	**	X	Two dollars **worth of** wood was all he needed.
4259	**A GOOD ('AT LEAST')**	1298	***	**	*	It took him **a good** six or seven years.
4261	**ACT ON**	1296	***	***	**	She never **acted on** that rage.
4262	**EXCEPT THAT**	1296	***	***	*	It seemed fine **except that** it was pink.
4278	**DAY TO DAY**	1289	***	***	*	They carried on with their **day to day** tasks.
4282	**AS USUAL**	1287	***	***	*	**As usual** he left it until the last minute.
4290	**LONG BEFORE**	1282	***	***	***	**Long before I** could quit **I** was offered a

						promotion.
4307	**LONG AGO**	1273	* * *	* * *	* * *	I gave up on that idea **long ago.**
4314	**IN CONJUNCTION WITH**	1270	x	* * *	* * *	The police**, in conjunction with** the fire department, managed to keep the crowd under control.
4319	**UP TO DATE**	1268	* * *	* * *	* * *	I try to keep **up to date.**
4333	**LET AL.ONE**	1261	* * *	* * *	* * *	I can't run **let al.one** walk.
4337	**QUITE A LOT**	1260	* * *	* *	x	I do swim **quite a lot.**
4342	**IF YOU LIKE**	1256	* * *	*	x	It was her 'chutzpah' **if you like** that impressed.
4349	**TO THE EXTENT**	1253	x	* *	* * *	He's interesting **to the extent** that he knows a lot.
4356	**SO FAR AS**	1251	* *	* * *	*	**So far as** I know it's not the first time either.
4362	**GIVEN THAT**	1247	* *	* * *	* * *	It was no surprise **given that** it was her job.
4372	**IN LINE WITH**	1240	* *	* * *	* * *	The decision is **in line with** the department's wishes.
4374	**ON THE WHOLE**	1238	* * *	* * *	* * *	It was acceptable **on the whole.**
4396	**CARE TO**	1223	* * *	* * *	x	Do you **care to** comment at all?
4408	**TAKE ACCOUNT OF**	1217	*	* *	* * *	We did not **take account of** the paperwork.

--end of 4k

4461	**SOMETHING LIKE THAT**	1192	* * *	x	x	He said he had a meeting to go to or **something like that.**
4462	**MAKE USE OF**	1191	* * *	* * *	* * *	We **made use of** the stopover.
4465	**WHEN IT COMES TO**	1188	* * *	* * *	* * *	I'm hopeless **when it comes to** goodbyes.
4466	**FILL IN**	1187	* * *	* * *	* * *	The victim **filled in** what was stolen.
4471	**FOR ALL ('CONSIDERING')**	1182	* * *	* * *	*	**For all** his money, he does not seem very happy.

4482	**A QUESTION OF**	1174	* * *	* * *	* * *	It's **a question of** trust.
4484	**FOR LIFE**	1172	* * *	* * *	* *	And he stayed there **for life.**
4497	**GET AWAY**	1165	* * *	* * *	X	They managed to **get away.**
4501	**IN THE MEANTIME**	1161	* * *	* * *	* *	I'll keep working on it **in the meantime.**
4522	**SOMETHING OF A**	1154	* * *	* *	X	She's **something of a** mathematics wiz.
4523	**THE ODD**	1154	* * *	* * *	*	I do play **the odd** jazz tune.
4556	**LITTLE MORE THAN**	1133	*	* * *	* * *	It's **little more than** pageantry if you ask me.
4558	**WOULD YOU LIKE**	1133	* * *	*	X	**Would you like** tea?
4584	**IN NEED**	1122	*	* * *	* * *	We want to help children **in need.**
4587	**TAKE FOR GRANTED**	1120	* * *	* * *	*	It's just something I've always **taken for granted.**
4605	**IN THIS RESPECT**	1110	X	* * *	* * *	**In this respect,** our study refutes earlier research.
4606	**PROVIDED THAT**	1110	*	* * *	* * *	It's OK **provided that** he come up with the goods.
4614	**ALLOW FOR ('CALCULATE IN')**	1105	*	* * *	* * *	Even if you **allow for** inflation, the price increase still seems exhorbitant.
4634	**CATCH UP**	1095	* * *	* * *	X	I took extra classes just to **catch up.**
4639	**A GO ('ATTEMPT')**	1093	* * *	* *	X	Why don't you have **a go?**
4649	**FOR THE MOMENT**	1092	* * *	* * *	*	**For the moment** things seem stable.
4656	**AT THE EXPENSE OF**	1086	*	* * *	* * *	They were laughing **at the expense of** the photo.
4670	**PUT TOGETHER**	1082	* * *	* * *	*	He **put together** the proposal in a matter of days.
4671	**THINGS LIKE THAT**	1082	* * *	*	X	He does karate and **things like that,** you see.
4682	**OF LITTLE**	1079	* *	* * *	* *	It was **of little** relevance, really.
4683	**SHUT UP**	1079	* * *	* *	X	I just couldn't **shut up.**
4703	**AS OF**	1069	* * *	* * *	* * *	It was completed **as of** June.
4704	**OVER TIME**	1069	* *	* * *	* * *	It might change **over time.**
4710	**WOULD APPEAR**	1068	* * *	* * *	* *	They people have spoken, it **would appear.**

4711	**THE OTHER DAY**	1066	* * *	*	X	I thought of you **the other day.**
4712	**IN THEORY**	1065	* * *	* * *	* * *	**In theory** that is why he went there.
4713	**THOUGHT OF (AS)**	1065	* * *	* * *	* * *	Beaches are **thought of** as vacation spots.
4721	**FOR GOOD**	1061	* * *	*	X	It's gone **for good.**
4726	**OPPOSED TO**	1059	*	* * *	* *	He's strangely **opposed to** gun control.
4749	**COMMON SENSE**	1049	* * *	* * *	* * *	To me it's **common sense.**
4754	**BOTHER TO**	1046	* * *	* *	X	He never **bothered to** reply.
4759	**AS GOOD AS ('LIKE')**	1043	* * *	* * *	*	It's **as good as** gone.
4761	**BACK UP**	1042	* * *	* * *	*	I have proof to **back up** my story.
4791	**TAKE CARE OF**	1034	* * *	* * *	*	I'll **take care of** this customer.
4816	**THE SIGHT OF**	1024	*	* * *	X	I can't stand **the sight of** it.
4819	**GO ROUND**	1022	* * *	* * *	X	I'll **go round** asking if anyone wants any more.
4834	**THE WHOLE THING**	1015	* * *	*	X	Why don't we just forget **the whole thing?**
4837	**AT ONE TIME**	1014	* *	* * *	* *	**At one time** I thought that too.
4839	**HEAD TO**	1014	* * *	* * *	X	Meanwhile, he decided to **head to** the hospital.
4840	**IN A SENSE**	1014	* * *	* *	* *	The food was **in a sense** only part of the experience.
4841	**ON AVERAGE**	1014	* *	* * *	*	How much does he make **on average?**
4846	**WAY ROUND**	1013	* * *	* * *	X	Do you know your **way round?**
4850	**CAN TELL**	1011	* * *	*	X	You **can tell** from the markings.
4896	**FREE FROM**	991	X	* * *	* * *	I dream of a life **free from** stress.
4901	**AND ALL THAT**	989	* * *	X	X	I like art **and all that.**
4909	**AS IT WERE**	985	* * *	* *	*	The party was a smash **as it were.**
4915	**WHAT IF**	984	* * *	*	*	**What if** we could fly?
4917	**TOUCH OF**	982	* * *	* * *	X	There was a **touch of** sadness in her voice.
4919	**BETTER OFF**	980	* * *	* *	*	You're **better off** without him.

4932	STAND FOR	977	*	* * *	*	That symbol used to **stand for** something.
4944	TO BLAME	973	* * *	* * *	* * *	I think the government's **to blame**, really.
4950	THE BULK OF	968	* * *	* * *	* * *	**The bulk of** it was done by noon.
4955	A HANDFUL OF	965	* * *	* * *	* * *	Only **a handful of** them were actually awake.
4959	(AT) THE OUTSET	963	* * *	* * *	* * *	It went well, but not **at the outset**.
4991	BY VIRTUE OF	954	* *	* * *	* * *	She won **by virtue of** her superior intellect.
5001	TURN DOWN	951	* * *	* * *	X	I won't **turn down** such a great offer.
5010	GET ON ('RELATE')	944	* * *	*	X	My parent don't really **get on** very well.
5022	UNDER WAY	939	X	* * *	* * *	Changes are already **under way**.
5025	IN THE INTEREST OF	938	*	* *	* * *	They stopped **in the interest of** time.
5051	ON THE MARKET	928	*	* * *	* *	How long has it been **on the market**?
5060	BY FAR	925	* * *	* * *	* *	The rise in unemployment was **by far** the highest.
5072	A DEGREE OF	921	*	* * *	* * *	There is **a degree of** irony in the story.
5082	NEVER MIND	915	* * *	*	X	**Never mind** that it's my birthday today.
5134	UP AND DOWN	898	* *	* * *	X	He was **up and down** the hall worrying about it.
5156	IN ONE'S OWN RIGHT	891	* * *	* * *	* * *	Gesture is a form of expression **in its own right**.
5169	A CASE OF	888	* * *	* * *	* * *	Now it's just **a case of** getting them to subscribe.
5171	MORE SO	887	* * *	* * *	* * *	Boston is walkable, and San Francisco **more so**.
5193	COME UP TO	881	* * *	* * *	X	Men always **come up to** me first.
5216	IN WHICH CASE	872	* *	* * *	* * *	That means it's midnight, **in which case** it's closed.
5222	NO SIGN OF	871	* * *	* * *	X	Still **no sign of** him?
5231	JUST AS ('WHEN')	869	* * *	* * *	X	I saw him **just as** he left.
5244	FOR THE SAKE OF	865	* * *	* * *	*	He exaggerated his points **for the sake of** effect.

5247	**IN A POSITION TO**	864	***	***	***	I'm not **in a position to** comment right now.
5255	**TO COME (FUTURE)**	862	***	***	***	We'll see how things go in the weeks **to come.**
5285	**BACKED BY**	850	x	***	x	The group is **backed by** the drug trade.
5303	**AT BEST**	844	***	***	*	He's **at best** average.
5306	**WEALTH OF**	844	*	***	***	The library holds a **wealth of** knowledge.
5309	**THAT SORT OF THING**	843	***	x	x	He drinks a lot and **that sort of thing.**
5311	**MAKE OUT**	842	*	***	x	He could barely **make out** her signature.
5323	**COME TO TERMS WITH**	839	*	***	*	They were finally able to **come to terms with** the change in weather.
5348	**FOND OF**	831	***	***	x	I'm not very **fond of** it, to be honest.
5354	**WITH A VIEW TO**	829	**	***	***	It was written **with a view to** inspiring hope.
5361	**TURN BACK**	827	*	***	*	There was no **turning back** now.
5366	**GET AWAY WITH**	824	***	***	x	Somehow they managed to **get away with** it.
5384	**NO WONDER**	820	***	**	x	**No wonder** he's ranked first.
5438	**WELL BEING**	809	***	***	*	It was good for their overall **well being.**
5447	**HOW ABOUT**	805	***	*	x	**How about** tomorrow?
5463	**TO GO ('REMAINING')**	801	***	**	x	Two more **to go.**
5479	**STRAIGHT AWAY**	794	***	***	*	They'll get to it **straight away.**
5485	**OWING TO**	793	**	***	***	**Owing to** his honesty, he was held in high esteem.
5491	**HOLD UP**	791	***	***	x	Sorry – we got **held up** on the M25.
5492	**LOOK TO**	790	**	***	***	He constantly **looked to** us for advice and support.
5496	**LAY OUT**	789	***	***	***	The countries **laid out** plans for a peace agreement.
5497	**THE LOT**	789	***	*	x	They tood the jewellery, the laptops, **the lot.**
5501	**KEEP ON**	788	***	**	x	Freeways **kept on** being built.
5502	**MAKE UP ONE'S**	788	***	**	x	You'd better **make up your mind.**

	MIND					
5503	**AT WORK**	787	**x**	* * *	* * *	There were strange forces **at work.**
5504	**COME ABOUT**	787	* *	* * *	*	It all **came about** through a meeting back in April.

--**end of 5k**

TOTAL: 505

Appendix 2 – First version of the PHRASE List (sample)

(For reasons of space, only the first two 1000 levels are shown here. The full PHRASE List is found in Appendix 1.)

MULTIWORD ITEMS LEMMATIZED AND DIVIDED INTO 1K FREQUENCY BANDS

HAVE TO	83092
THERE ARE	59833
SUCH AS	30857
YOU KNOW	28577
GOING TO (FUTURE)	28259
OF COURSE	26966
A FEW	26451
AT LEAST	25034
SUCH A	23894
I MEAN	23616
A LOT	22332
RATHER THAN	21085
SO THAT	20966
A LITTLE	20296
A BIT	19618
AS WELL AS	18041
IN FACT	15983
THE TWO	15689
SO MUCH	15643
GO ON	15610
ACCORDING TO	15432
IS TO ('WILL')	15232
A NUMBER OF	15090
AT ALL	14650
AS IF	14470
USED TO (PAST)	14411
WAS TO ('WOULD')	14366
NOT ONLY	14110
THOSE WHO	13951
DEAL WITH	13634
LEAD TO ('CAUSE')	13555
SORT OF	13361
THE FOLLOWING	12963
IN ORDER TO	12762
HAVE GOT A	12734

1k

HAVE GOT TO ('MUST')	12270
SET UP	11560

AS TO	11535
AS WELL	11519
BASED ON	11440
CARRY OUT	10753
TAKE PLACE	10556
TEND TO	10504
DUE TO	10454
FAIL TO	10263
EACH OTHER	10160
IN TERMS OF	9881
NO ONE	9597
PICK UP	9252
UP TO (MAXIMUM)	8733
A SINGLE ('ANY')	8710
NO LONGER	8556
LOOK FOR	8377
BE HELD ('TAKE PLACE')	8043
LAST NIGHT	7992
AS A RESULT	7939
IN ADDITION	7822
WORK ON	7600
THINK ABOUT	7243
MAKE UP	7176
FOR INSTANCE	7138
TOO MUCH	7123
YOU SEE	7102
IN PARTICULAR	7092
A COUPLE OF	7007
INSTEAD OF	6907
COME BACK	6772
ON BEHALF OF	6734
LOOK LIKE	6595
FIND OUT	6499
AS __ AS POSSIBLE	6440
COME UP	6405
POINT OUT	6325
APART FROM	6287
CALL FOR	6243
OR TWO	6192
A FURTHER	6121
COME OUT	6031
BE EXPECTED TO	5964
SEEK TO	5937
GO THROUGH	5857
LONG TERM	5831
RESULT IN	5763
THAT IS (REPHRASING)	5737
EVEN THOUGH	5664
A RANGE OF	5651
FULL OF	5549
AS FAR AS	5538
THE LATTER (anaphor)	5519
MAKE SURE	5510
TAKE OVER	5394
CONSIST OF	5362
AS SOON AS	5323
AT THE TIME ('WHEN THIS HAPPENED')	5282
ON THE OTHER HAND	5267
ON ONE'S OWN	5240

ALL RIGHT	5230
SUBJECT TO	5218
AFTER ALL (adv.)	5197
IN FRONT OF ('BEFORE')	5190
TO DO WITH	5184
GO OUT	5173
A GREAT DEAL ('MUCH')	5126
ON THE WAY	5085
AS LONG AS	5084
SO FAR ('UNTIL NOW')	5018
OUGHT TO	5002
AT THE MOMENT	5001
AS THOUGH	4988
COME TO ('EVOLVE TO')	4970
ALONG WITH	4948
MAY WELL ('COULD')	4931
PUT ON	4927
WHAT IF	4915
GET OUT	4858

2k

Appendix 3 – The PHRASE List criteria validation rating sheet

(The phrases are highlighted here for purposes of illustration only.)

TICK ALL THE NGRAMS THAT MATCH THE CRITERIA. IF YOU ARE UNSURE, YOU CAN MAKE NOTES ABOUT THE UNCERTAINTY IN THE RIGHT HAND COLUMN. (THE FIRST 50 ARE FOR TRAINING PURPOSES.)

	1.	AND HE	25,723	
	2.	TO TAKE	25,684	
	3.	FROM A	25,672	
	4.	DO YOU	25,441	
	5.	AT A	25,372	
	6.	THE WAY	25,203	
	7.	BUT IT	25,143	
	8.	AT LEAST	25,034	
	9.	AND TO	24,198	
	10.	NOT BE	24,050	
	11.	IT WOULD	24,013	
	12.	THE LAST	23,985	
	13.	SUCH A	23,894	
	14.	BETWEEN THE	23,887	
	15.	MUST BE	23,874	
	16.	OF THEM	23,692	
	17.	I MEAN	23,616	
	18.	THE WORLD	23,601	
	19.	BUT I	23,422	
	20.	FOR EXAMPLE	23,417	
	21.	DURING THE	23,385	
	22.	Â€™ THE	23,295	
	23.	OF THESE	23,278	
	24.	A LOT	22,332	
	25.	OF AN	22,038	
	26.	NEED TO	21,966	
	27.	OF HER	21,809	
	28.	A GOOD	21,485	
	29.	AFTER THE	21,373	
	30.	THE NEXT	21,217	
	31.	THAT WAS	21,184	
	32.	YOU HAVE	21,168	
	33.	WHICH THE	21,090	
	34.	RATHER THAN	21,085	
	35.	THE END OF	21,073	
	36.	DO NOT	21,002	
	37.	SO THAT	20,966	
	38.	IN THEIR	20,774	
	39.	THERE WERE	20,654	
	40.	BACK TO	20,638	
	41.	A LITTLE	20,296	
	42.	WHEN HE	20,179	
	43.	OF ITS	20,070	

	44. LIKE A	19,980	
	45. UNDER THE	19,946	
	46. YOU ARE	19,884	
	47. THAT I	19,785	
	48. UP THE	19,785	
	49. THE SECOND	19,772	
	50. Â€ AND	19,764	

	51. DOES NOT	19,612	
	52. AND HIS	19,561	
	53. I HAD	19,522	
	54. SHE SAID	19,488	
	55. WE ARE	19,449	
	56. THE GOVERNMENT	19,370	
	57. BE THE	19,299	
	58. A VERY	19,252	
	59. I AM	19,167	
	60. THAT A	19,145	
	61. ARE NOT	19,100	
	62. THE ONLY	19,043	
	63. INTO A	19,018	
	64. WHERE THE	18,966	
	65. OF ALL	18,952	
	66. NOT TO	18,814	
	67. WELL AS	18,779	
	68. IT HAS	18,746	
	69. THEY HAD	18,672	
	70. THE HOUSE	18,545	
	71. THAT WE	18,495	
	72. WITHIN THE	18,495	
	73. THE WHOLE	18,373	
	74. OF IT	18,236	
	75. AS WELL AS	18,198	
	76. THAT THERE	18,198	
	77. AND THEY	18,190	
	78. AS HE	18,188	
	79. TO SAY	18,157	
	80. TO HIS	17,936	
	81. THE BEST	17,813	
	82. THE RIGHT	17,792	
	83. USE OF	17,508	
	84. BECAUSE OF	17,480	
	85. IN AN	17,477	
	86. Â€ THE	17,456	
	87. MOST OF	17,416	
	88. A BIT	19,618	
	89. HE IS	17,317	
	90. PART OF THE	17,306	
	91. THIS WAS	17,285	

	92. THEIR OWN	17,246	
	93. THAT YOU	17,146	
	94. WAS IN	17,076	
	95. TO HER	16,955	
	96. THEY HAVE	16,864	
	97. THAN THE	16,678	
	98. AGAINST THE	16,593	
	99. WHO HAD	16,528	
	100. IT WILL	16,469	
	101. Â€™ HE SAID	16,424	
	102. I KNOW	16,338	
	103. ON TO	16,261	
	104. AND SO	16,250	
	105. TO GIVE	16,161	
	106. BEFORE THE	16,088	
	107. AS AN	15,993	
	108. IN FACT	15,983	
	109. A NUMBER	15,935	
	110. IS AN	15,931	
	111. LIKELY TO	15,854	
	112. IS NO	15,843	
	113. LOT OF	15,821	
	114. DON'T KNOW	15,774	
	115. AS IT	15,751	
	116. AND SHE	15,734	
	117. OUT OF THE	15,702	
	118. THE TWO	15,689	
	119. SO MUCH	15,643	
	120. GO ON	15,610	
	121. THE MAIN	15,609	
	122. ARE THE	15,547	
	123. TRYING TO	15,518	
	124. MIGHT BE	15,481	
	125. WHICH WAS	15,463	
	126. WHEN I	15,447	
	127. ACCORDING TO	15,432	
	128. THE OLD	15,298	
	129. IS TO (WILL)	15,232	
	130. A NUMBER OF	15,090	
	131. SOME OF THE	15,089	
	132. OR THE	15,069	
	133. WHAT IS	15,006	
	134. HAS A	15,000	
	135. THERE IS A	14,872	
	136. TO FIND	14,867	
	137. ON HIS	14,828	
	138. THE FACT	14,796	
	139. THAT SHE	14,732	
	140. AS THEY	14,701	
	141. AT ALL	14,650	

	142. WE CAN	14,640	
	143. IN HER	14,558	
	144. I CAN	14,530	
	145. IF THEY	14,483	
	146. LOOK AT	14,475	
	147. AS IF	14,470	
	148. AND WE	14,439	
	149. HIS OWN	14,436	
	150. IS IT	14,431	
	151. A LOT OF	14,430	
	152. USED TO	14,411	
	153. OF THAT	14,362	
	154. BUT HE	14,355	
	155. IN ANY	14,319	
	156. I WOULD	14,316	
	157. A SMALL	14,265	
	158. FACT THAT	14,242	
	159. IT TO	14,219	
	160. A LONG	14,196	
	161. IS IN	14,173	
	162. I SAID	14,169	
	163. NOT ONLY	14,110	
	164. THERE WAS A	14,050	
	165. THOSE WHO	13,951	
	166. UP AND	13,935	
	167. IF IT	13,934	
	168. NOT A	13,900	
	169. AND YOU	13,876	
	170. WHICH ARE	13,860	
	171. DOWN THE	13,825	
	172. HE WOULD	13,805	
	173. IT HAD	13,804	
	174. IN ORDER	13,784	
	175. AND OTHER	13,700	
	176. AND WAS	13,677	
	177. THE COMPANY	13,666	
	178. DEAL WITH	13,634	
	179. BE ABLE	13,597	
	180. OF OUR	13,572	
	181. END OF THE	13,560	
	182. HAVE THE	13,558	
	183. THAT IN	13,557	
	184. WITH HIS	13,425	
	185. BE ABLE TO	13,383	
	186. SORT OF	13,361	
	187. IF HE	13,240	
	188. OF WHICH	13,229	
	189. IT'S A	13,197	
	190. AND THERE	13,194	
	191. WE WERE	13,175	

	192. TO KEEP	13,121	
	193. THE NATIONAL	13,064	
	194. A GREAT	13,060	
	195. COULD NOT	12,997	
	196. IT WAS A	12,988	
	197. SAID THAT	12,985	
	198. THE FOLLOWING	12,963	
	199. THE FACT THAT	12,954	
	200. OF THOSE	12,943	
	201. THE BRITISH	12,930	
	202. HE HAS	12,924	
	203. THEM TO	12,911	
	204. MANY OF	12,864	
	205. WHEN THEY	12,826	
	206. AND IS	12,799	
	207. LIKE THE	12,798	
	208. IN ORDER TO	12,762	
	209. HAVE GOT A	12,734	
	210. LIKE TO	12,722	
	211. BEEN A	12,699	
	212. TERMS OF	12,686	
	213. ACROSS THE	12,639	
	214. ORDER TO	12,634	
	215. RANGE OF	12,619	
	216. AS I	12,618	
	217. LIKE THAT	12,608	
	218. YOU WANT	12,604	
	219. NOT THE	12,560	
	220. Â€™ Â€˜I	12,559	
	221. WHEN YOU	12,532	
	222. THE PAST	12,531	
	223. THAT THIS	12,472	
	224. WAS NO	12,398	
	225. IN ITS	12,379	
	226. TO USE	12,377	
	227. IT AND	12,351	
	228. THE TOP	12,335	
	229. THE ONE	12,272	
	230. HAVE GOT TO	12,270	
	231. THE DOOR	12,255	
	232. WHO WAS	12,246	
	233. MEMBERS OF	12,236	
	234. IN SOME	12,203	
	235. OFF THE	12,188	
	236. IF I	12,180	
	237. HE COULD	12,170	
	238. ARE YOU	12,148	
	239. FOR BNC	12,147	
	240. WANTED TO	12,124	
	241. IS ALSO	12,113	

	242. TO WORK	12,060	
	243. CENT OF	12,049	
	244. PER CENT OF	12,037	
	245. WOULD NOT	12,029	
	246. OUT THE	11,986	
	247. IT IN	11,908	
	248. THE UNITED	11,900	
	249. TIME TO	11,899	
	250. GO TO	11,878	
	251. A LARGE	11,861	
	252. AS SHE	11,850	
	253. HIM TO	11,839	
	254. TO BE A	11,826	
	255. AFTER A	11,821	
	256. TO HIM	11,818	
	257. FORM OF	11,780	
	258. I THOUGHT	11,779	
	259. GOT A	11,709	
	260. THE NUMBER	11,690	
	261. THE DAY	11,663	
	262. THE REST	11,647	
	263. YOU WILL	11,624	
	264. Â€™ I	11,604	
	265. SET UP	11,560	
	266. IN MY	11,548	
	267. AS TO	11,535	
	268. AS WELL	11,519	
	269. AND THEIR	11,511	
	270. TO THEIR	11,475	
	271. BASED ON	11,440	
	272. WITH AN	11,418	
	273. AWAY FROM	11,411	
	274. SINCE THE	11,376	
	275. TO HELP	11,369	
	276. AROUND THE	11,351	
	277. I COULD	11,347	
	278. THE COUNTRY	11,266	
	279. OR A	11,228	
	280. A MAN	11,186	
	281. WENT TO	11,184	
	282. THERE IS NO	11,182	
	283. WHAT I	11,156	
	284. TO KNOW	11,135	
	285. THAT OF	11,132	
	286. SO I	11,119	
	287. WILL HAVE	11,119	
	288. OF WHAT	11,110	
	289. AND THIS	11,085	
	290. THE MORE	11,081	
	291. MAKE A	11,076	

	292. WAY OF	11,068	
	293. AND IF	11,052	
	294. OF MY	11,025	
	295. MAY HAVE	11,017	
	296. THE LOCAL	10,973	
	297. TO AN	10,928	
	298. DO IT	10,865	
	299. WILL NOT	10,848	
	300. WAS THAT	10,834	
	301. SEEMED TO	10,830	
	302. IT MAY	10,819	
	303. FOR THIS	10,764	
	304. IF WE	10,758	
	305. MAY NOT	10,752	
	306. ABOUT IT	10,748	
	307. WITH HER	10,658	
	308. TO THIS	10,638	
	309. YOU DON'T	10,584	
	310. THE END OF THE	10,575	
	311. TO LOOK	10,569	
	312. BEGAN TO	10,552	
	313. WAY TO	10,549	
	314. WHICH HE	10,540	
	315. AT THE END	10,526	
	316. YOU TO	10,509	
	317. BUT THEY	10,483	
	318. Â€ A	10,458	
	319. DUE TO	10,454	
	320. KNOW WHAT	10,450	
	321. UP IN	10,445	
	322. IN TERMS	10,429	
	323. SEE THE	10,396	
	324. FOR HIS	10,389	
	325. YEAR OLD	10,383	
	326. OF YOUR	10,381	
	327. THE BACK	10,381	
	328. BE USED	10,372	
	329. SIDE OF	10,343	
	330. THE PUBLIC	10,336	
	331. IS NOW	10,328	
	332. THE YEAR	10,291	
	333. A YEAR	10,290	
	334. AND AS	10,288	
	335. ALL OF	10,275	
	336. IT WOULD BE	10,273	
	337. AND ITS	10,215	
	338. TO PUT	10,214	
	339. OUT TO	10,193	
	340. TOWARDS THE	10,192	
	341. Â€™ SHE SAID	10,180	

	342. THE UK	10,180	
	343. THAN A	10,170	
	344. EACH OTHER	10,160	
	345. WHAT THE	10,138	
	346. AND FOR	10,125	
	347. YEARS AGO	10,123	
	348. WAS AN	10,112	
	349. BECAUSE THE	10,074	
	350. DOWN TO	10,062	
	351. SEEMS TO	10,048	
	352. PEOPLE WHO	10,035	
	353. MUST HAVE	10,034	
	354. IT IS NOT	10,016	
	355. OF ANY	9,999	
	356. AND OF	9,996	
	357. IT'S NOT	9,970	
	358. I DON'T KNOW	9,969	
	359. THE PEOPLE	9,947	
	360. BUT IN	9,934	
	361. TO PROVIDE	9,910	
	362. THE PROBLEM	9,900	
	363. HAD THE	9,899	
	364. WHEN IT	9,894	
	365. DIFFICULT TO	9,887	
	366. THE COURT	9,883	
	367. IN TERMS OF	9,881	
	368. HAS TO	9,863	
	369. HAVE TO	9,861	
	370. THE EARLY	9,848	
	371. A MORE	9,842	
	372. TO YOU	9,835	
	373. SAY THAT	9,825	
	374. BE IN	9,813	
	375. DID YOU	9,793	
	376. AS WE	9,792	
	377. AT THIS	9,788	
	378. THE WORK	9,779	
	379. WENT TO	9,771	
	380. THE MAN	9,759	
	381. UP A	9,753	
	382. COULD HAVE	9,749	
	383. I DO	9,728	
	384. WHO ARE	9,707	
	385. WHEN SHE	9,705	
	386. THE NUMBER OF	9,704	
	387. THE USE	9,660	
	388. WE HAD	9,626	
	389. WHICH HAD	9,626	
	390. AND NOT	9,615	
	391. THE GREAT	9,614	

	392. WHAT YOU	9,599	
	393. NO ONE	9,597	
	394. ONLY A	9,595	
	395. UP WITH	9,591	
	396. HAD NOT	9,576	
	397. AND ALL	9,559	
	398. THIS TIME	9,545	
	399. FOR THEIR	9,541	
	400. THEY WOULD	9,540	
	401. WERE NOT	9,521	
	402. NOT HAVE	9,517	
	403. TRIED TO	9,484	
	404. HOW TO	9,481	
	405. THEY CAN	9,432	
	406. SENSE OF	9,431	
	407. BE MADE	9,429	
	408. SEEM TO	9,427	
	409. THE POLICE	9,403	
	410. MUCH OF	9,400	
	411. ME TO	9,364	
	412. IT IS A	9,356	
	413. YOU THINK	9,340	
	414. AT THE END OF	9,305	
	415. IT CAN	9,265	
	416. WHAT WAS	9,248	
	417. FOR AN	9,209	
	418. WHICH HAS	9,209	
	419. THE NORTH	9,201	
	420. YOU COULD	9,191	
	421. BUT THERE	9,189	
	422. WHILE THE	9,156	
	423. IS ONE	9,155	
	424. RESULT OF	9,153	
	425. IN ONE	9,137	
	426. CAME TO	9,134	
	427. ON HER	9,107	
	428. SAID THE	9,107	
	429. TO PAY	9,089	
	430. THE USE OF	9,086	
	431. SHE COULD	9,083	
	432. I DIDN'T	9,061	
	433. OF US	9,060	
	434. THE FUTURE	9,056	
	435. ITS OWN	9,052	
	436. THEN THE	9,047	
	437. YOU WERE	9,046	
	438. LESS THAN	9,015	
	439. ON THIS	8,973	
	440. AT HER	8,967	
	441. ON THEIR	8,957	

	442. THAT IF	8,926	
	443. SAID TO	8,921	
	444. THE GENERAL	8,913	
	445. INVOLVED IN	8,911	
	446. THERE WAS NO	8,902	
	447. LAST YEAR	8,898	
	448. THE STATE	8,896	
	449. WELL I	8,850	
	450. BUT NOT	8,842	
	451. YOU DO	8,839	
	452. THE MIDDLE	8,829	
	453. REST OF	8,797	
	454. THE CITY	8,786	
	455. OF PEOPLE	8,785	
	456. AND MORE	8,769	
	457. THEY WILL	8,749	
	458. UP TO	8,733	
	459. AMONG THE	8,717	
	460. A SINGLE	8,710	
	461. IN OUR	8,703	
	462. A RESULT	8,693	
	463. UNTIL THE	8,675	
	464. THE IDEA	8,673	
	465. TO THAT	8,670	
	466. THE REST OF	8,665	
	467. ALONG THE	8,655	
	468. THE ROAD	8,654	
	469. AMOUNT OF	8,649	
	470. TRY TO	8,649	
	471. SHOULD HAVE	8,621	
	472. TO WHICH	8,615	
	473. BECAUSE IT	8,612	
	474. WHAT THEY	8,607	
	475. I CAN'T	8,594	
	476. LEVEL OF	8,582	
	477. KNOW THAT	8,579	
	478. AND ON	8,578	
	479. DEVELOPMENT OF	8,571	
	480. MOST OF THE	8,568	
	481. LACK OF	8,566	
	482. NO LONGER	8,556	
	483. SEE P	8,555	
	484. HAVE YOU	8,552	
	485. IT Â€™	8,549	
	486. WHAT HE	8,540	
	487. FIRST TIME	8,521	
	488. FOR YOU	8,503	
	489. THE QUESTION	8,478	
	490. I'M NOT	8,453	
	491. THE MOMENT	8,442	

	492. ON THE OTHER	8,429	
	493. EVEN IF	8,401	
	494. MIGHT HAVE	8,382	
	495. THE VERY	8,381	
	496. AS YOU	8,366	
	497. TIME AND	8,362	
	498. THANK YOU	8,357	
	499. THAT ONE	8,350	
	500. WITH THEIR	8,341	
	501. YOU'VE GOT	8,334	
	502. AND WITH	8,323	
	503. WHICH THEY	8,320	
	504. OF STATE	8,282	
	505. TO IT	8,280	
	506. AND ER	8,265	
	507. THIS YEAR	8,250	
	508. S UNITS	8,236	
	509. WHEN WE	8,227	
	510. FROM HIS	8,216	
	511. HIM AND	8,204	
	512. WAS ALSO	8,204	
	513. HE SAYS	8,202	
	514. GET A	8,197	
	515. OUT IN	8,197	
	516. THAT IT IS	8,195	
	517. THE US	8,181	
	518. AND WHEN	8,180	
	519. HE DID	8,163	
	520. FOR HER	8,145	
	521. A WEEK	8,143	
	522. IN THE FIRST	8,143	
	523. COUPLE OF	8,135	
	524. FAR AS	8,110	
	525. FOR IT	8,106	
	526. THE BNC	8,099	
	527. I WANT	8,081	
	528. NATURE OF	8,080	
	529. IN YOUR	8,079	
	530. AND HAD	8,072	
	531. SERIES OF	8,057	
	532. YOU GET	8,056	
	533. IN OTHER	8,051	
	534. ARE A	8,047	
	535. DON'T THINK	8,024	
	536. THE WAR	8,023	
	537. THESE ARE	8,010	
	538. THE FAMILY	7,997	
	539. HAD NO	7,993	
	540. LAST NIGHT	7,992	
	541. AND ONE	7,990	

	542. AT THE SAME	7,988	
	543. MUCH MORE	7,981	
	544. MEMBERS OF THE	7,973	
	545. WERE THE	7,962	
	546. THE NEED	7,958	
	547. ARE IN	7,953	
	548. AND AN	7,942	
	549. AS A RESULT	7,939	
	550. SHE WOULD	7,939	

Appendix 4 – An integrated list of words and phrasal expressions

(Only a sample of 600 total items is provided here in order to conserve space.)

475.	BLACK	27181
476.	EITHER	27010
477.	OF COURSE	26966
478.	AMONG	26742
479.	SOCIETY	26641
480.	NORTH	26618
481.	COMMUNITY	26558
482.	PATIENT	26501
483.	A FEW	26451
484.	VOICE	26400
485.	SPEND	26365
486.	DISCUSS	26350
487.	WHITE	26344
488.	DOCTOR	26343
489.	LEARN	26317
490.	CITY	26312
491.	COMPUTE	26259
492.	ELECT	26144
493.	SELL	26000
494.	TREAT	25978
495.	TEST	25842
496.	TAX	25824
497.	MARK	25805
498.	EXPLAIN	25748
499.	PAST	25726
500.	DECISION	25710
501.	ATTEND	25656
502.	REMEMBER	25648
503.	UNIVERSITY	25629
504.	REDUCE	25626
505.	REFER	25612
506.	NECESSARY	25535
507.	SCIENCE	25471
508.	ASSOCIATE	25460
509.	ESTABLISH	25300
510.	PASS	25157
511.	ABOVE	25111
512.	AT LEAST	25034
513.	RECEIVE	25031
514.	PERFORM	25009
515.	PRACTISE	25005
516.	SCOTLAND	24991
517.	STRUCTURE	24929
518.	MATERIAL	24895
519.	RANGE	24883
520.	SECTION	24868

521.	SIMILAR	24748
522.	DECIDE	24684
523.	SENSE	24615
524.	TOTAL	24604
525.	ENTER	24597
526.	ELSE	24571
527.	SEND	24570
528.	STRONG	24528
529.	ARM	24492
530.	DETAIL	24446
531.	DISTRIBUTE	24409
532.	REACH	24400
533.	SECURE	24316
534.	UNION	24316
535.	ENVIRONMENT	24224
536.	BED	24200
537.	WEST	24183
538.	MOMENT	24175
539.	CHARGE	24166
540.	CONTAIN	24055
541.	ACROSS	24022
542.	EXPRESS	24011
543.	APPROACH	23959
544.	STAFF	23907
545.	SUCH A	23894
546.	BEAR	23743
547.	FUTURE	23694
548.	PROTECT	23669
549.	I MEAN	23616
550.	STANDARD	23586
551.	TWENTY	23584
552.	ENGLISH	23489
553.	ALONG	23412
554.	HUMAN	23406
555.	CHURCH	23400
556.	SOUND	23340
557.	SEVERAL	23322
558.	PROJECT	23265
559.	LIMIT	23233
560.	SERVE	23216
561.	GERMANY	23211
562.	PAPER	23184
563.	TABLE	23135
564.	TEAM	23109
565.	ANSWER	23102
566.	GIRL	23068
567.	VARY	23044
568.	UPON	22963
569.	GROUND	22954
570.	THEREFORE	22929
571.	COLLECT	22913
572.	PLEASE	22889
573.	FIRM	22881
574.	RECOGNIZE	22870
575.	UNITE	22812
576.	WAIT	22748
577.	COMMITTEE	22727
578.	FRONT	22717
579.	DEPARTMENT	22670
580.	DIE	22665

581.	DEMAND	22569
582.	TODAY	22563
583.	WATCH	22522
584.	COMPARE	22505
585.	DATA	22493
586.	TEN	22470
587.	BEHIND	22420
588.	DEATH	22381
589.	INTRODUCE	22334
590.	A LOT	22332
591.	CHARACTER	22327
592.	STAGE	22307
593.	FIELD	22305
594.	CURRENT	22304
595.	PROPOSE	22264
596.	PAGE	22259
597.	MARRY	22257
598.	DEEP	22249
599.	ATTEMPT	22217
600.	LANGUAGE	22187
601.	STUDENT	22140
602.	ANALYSE	21981
603.	BENEFIT	21968
604.	COLOUR	21824
605.	RESPONSIBLE	21630
606.	ACHIEVE	21587
607.	IDENTIFY	21570
608.	PRESS	21535
609.	REGION	21523
610.	SINGLE	21502
611.	MEASURE	21476
612.	PARENT	21463
613.	RAISE	21396
614.	SOON	21374
615.	EVIDENCE	21334
616.	SIGNIFICANT	21330
617.	LETTER	21244
618.	EAST	21237
619.	TOWN	21200
620.	CHOOSE	21183
621.	PROCEED	21182
622.	RATHER THAN	21085
623.	POST	21035
624.	KING	21011
625.	GOD	20982
626.	SO THAT	20966
627.	FATHER	20961
628.	FOOD	20931
629.	ROLE	20898
630.	AIR	20829
631.	EVENT	20775
632.	SIGN	20773
633.	NORMAL	20768
634.	THEORY	20718
635.	MRS	20670
636.	TELEPHONE	20575
637.	GAME	20536
638.	MORNING	20517
639.	IMPROVE	20509
640.	SITE	20500

641.	SHOP	20429
642.	AWARD	20340
643.	A LITTLE	20296
644.	SUPPLY	20235
645.	AMP	20215
646.	THUS	20159
647.	MISS	20135
648.	MUSIC	20007
649.	MODEL	19989
650.	STEP	19969
651.	CONTRACT	19941
652.	SELF	19901
653.	LIST	19889
654.	LORD	19856
655.	STAY	19849
656.	BOY	19817
657.	SALE	19774
658.	PRESIDENT	19765
659.	INVEST	19761
660.	DOUBT	19710
661.	SEX	19704
662.	REST	19671
663.	A BIT	19618
664.	AMOUNT	19597
665.	INSTITUTE	19544
666.	METHOD	19460
667.	CLUB	19357
668.	WISH	19291
669.	PUBLISH	19272
670.	BOARD	19257
671.	AGO	19231
672.	SHALL	19220
673.	BLOOD	19187
674.	FUND	19168
675.	JUDGE	19126
676.	REGARD	19120
677.	BAD	19060
678.	PRIVATE	19038
679.	WINDOW	19012
680.	WONDER	18951
681.	QUICK	18918
682.	QUALITY	18917
683.	FORMER	18887
684.	SOURCE	18861
685.	PREVIOUS	18843
686.	FAIR	18785
687.	POUND	18771
688.	HUNDRED	18769
689.	KILL	18769
690.	OBVIOUS	18707
691.	SAFE	18671
692.	FIRE	18616
693.	DIVIDE	18583
694.	EDIT	18513
695.	SPECIFIC	18492
696.	DETERMINE	18490
697.	INDEPENDENT	18416
698.	FARM	18306
699.	TRADITION	18299
700.	DARK	18284

701.	PLANT	18193
702.	OFFICIAL	18143
703.	ESPECIAL	18115
704.	INDEED	18084
705.	YESTERDAY	18071
706.	AS WELL AS	18041
707.	SERIOUS	18017
708.	STATES	17999
709.	HAPPY	17996
710.	LEGAL	17975
711.	EQUAL	17974
712.	STORY	17951
713.	LAY	17910
714.	CHAPTER	17903
715.	WALL	17897
716.	SIR	17880
717.	FAVOUR	17873
718.	THANK	17821
719.	HOSPITAL	17790
720.	SUPPOSE	17779
721.	FUNCTION	17763
722.	PREPARE	17760
723.	INDICATE	17711
724.	FIGHT	17666
725.	FOREIGN	17575
726.	CROSS	17526
727.	TELEVISION	17511
728.	ATTACK	17494
729.	SECRETARY	17473
730.	DUE	17449
731.	SEEK	17447
732.	FEATURE	17428
733.	SMILE	17423
734.	DEFINE	17415
735.	WIFE	17397
736.	CULTURE	17394
737.	DRINK	17394
738.	SCHEME	17379
739.	ARRIVE	17306
740.	HONOUR	17293
741.	RED	17248
742.	JOIN	17246
743.	SAVE	17124
744.	AGENT	17122
745.	CHECK	17102
746.	M	17056
747.	PICTURE	17054
748.	SELECT	17040
749.	DEPEND	17011
750.	VOTE	16937
751.	INFLUENCE	16912
752.	EXCEPT	16880
753.	OCCUR	16863
754.	R	16851
755.	PACK	16848
756.	SEPARATE	16807
757.	FISH	16797
758.	FAIL	16765
759.	EXAMINE	16755
760.	FORWARD	16742

761.	STYLE	16710
762.	BILL	16699
763.	ASSUME	16691
764.	OBSERVE	16673
765.	SURPRISE	16663
766.	PAINT	16602
767.	REALISE	16585
768.	PROPERTY	16570
769.	NOTICE	16568
770.	POOR	16567
771.	MAINTAIN	16546
772.	WRONG	16545
773.	LINK	16532
774.	GARDEN	16518
775.	MACHINE	16449
776.	HEART	16441
777.	SIZE	16396
778.	RELATIVE	16298
779.	IMMEDIATE	16221
780.	TECHNOLOGY	16210
781.	PURPOSE	16167
782.	ORIGINAL	16144
783.	ARRANGE	16113
784.	MARCH	16087
785.	STRIKE	16061
786.	MATCH	16044
787.	TRAVEL	16043
788.	SLOW	16022
789.	ENJOY	16019
790.	IN FACT	15983
791.	LICENCE	15964
792.	OXFORD	15915
793.	GREEN	15910
794.	FINE	15884
795.	WEAR	15838
796.	RISK	15837
797.	SON	15835
798.	CHANCE	15824
799.	SUIT	15824
800.	LIE	15810
801.	SEA	15800
802.	SPACE	15771
803.	ROYAL	15761
804.	EAT	15743
805.	OPPORTUNITY	15739
806.	PARK	15704
807.	THE TWO	15689
808.	HAIR	15683
809.	SKILL	15669
810.	VARIOUS	15664
811.	SO MUCH	15643
812.	THOUSAND	15622
813.	GO ON	15610
814.	IRISH	15606
815.	ACCORDING	15563
816.	CONTRIBUTE	15560
817.	FEAR	15530
818.	D	15519
819.	LOSS	15511
820.	ANNOUNCE	15495

821.	CATCH	15495
822.	W	15469
823.	PATTERN	15434
824.	ACCORDING TO	15432
825.	RESPONSE	15404
826.	PICK	15354
827.	CONSULT	15352
828.	G	15295
829.	OCCASION	15294
830.	FINISH	15290
831.	MODERN	15289
832.	MIDDLE	15287
833.	OPPOSE	15287
834.	AWARE	15280
835.	AIM	15266
836.	GRANT	15264
837.	IS TO (WILL)	15232
838.	ASSESS	15230
839.	ADMIT	15220
840.	TEND	15198
841.	PROVE	15143
842.	APRIL	15136
843.	SUDDEN	15123
844.	PRESSURE	15095
845.	A NUMBER OF	15090
846.	SUFFER	15081
847.	DEFENCE	15047
848.	KEY	15008
849.	FACTOR	14963
850.	POPULATION	14954
851.	PIECE	14947
852.	OCTOBER	14939
853.	APPEAL	14938
854.	PROFIT	14934
855.	ANIMAL	14930
856.	FILM	14870
857.	JANUARY	14847
858.	COPY	14844
859.	SHOOT	14807
860.	COUPLE	14806
861.	DE	14781
862.	ENCOURAGE	14781
863.	DAMAGE	14777
864.	COMMISSION	14756
865.	INSIDE	14732
866.	JUNE	14719
867.	RESPECT	14702
868.	HEAVY	14700
869.	ATTRACT	14688
870.	VILLAGE	14656
871.	AT ALL	14650
872.	TREE	14642
873.	KNOWLEDGE	14636
874.	NEWS	14624
875.	Â€˜IT	14621
876.	AFFECT	14596
877.	WOOD	14583
878.	APPOINT	14578
879.	SETTLE	14555
880.	SERIES	14485

881.	AS IF	14470
882.	CONSTRUCT	14447
883.	MILE	14444
884.	SUN	14429
885.	USED TO (PAST)	14411
886.	REFLECT	14399
887.	WAS TO (WOULD)	14366
888.	THREAT	14337
889.	PROFESSIONAL	14324
890.	PARLIAMENT	14322
891.	ENERGY	14317
892.	DANGER	14311
893.	EVENING	14310
894.	BASIS	14292
895.	DESPITE	14256
896.	STATION	14220
897.	DISCOVER	14201
898.	WORTH	14166
899.	CONTACT	14135
900.	INSTEAD	14119
901.	NOT ONLY	14110
902.	BELOW	14078
903.	ENSURE	14022
904.	PULL	14010
905.	DROP	13958
906.	THOSE WHO	13951
907.	CAPITAL	13947
908.	REPLACE	13927
909.	FLY	13912
910.	RAIL	13908
911.	PRISON	13901
912.	NICE	13889
913.	BEHAVIOUR	13884
914.	OKAY	13835
915.	POTENTIAL	13826
916.	READY	13819
917.	INCOME	13766
918.	RACE	13756
919.	APPROPRIATE	13754
920.	CORPORATE	13750
921.	COMMIT	13742
922.	PRINCIPLE	13741
923.	SEPTEMBER	13730
924.	ACCESS	13713
925.	REMOVE	13694
926.	CUP	13683
927.	DEAL WITH	13634
928.	BEAUTY	13624
929.	AID	13589
930.	INVESTIGATE	13578
931.	TOUR	13563
932.	LEAD TO ('CAUSE')	13555
933.	CHOICE	13530
934.	DECEMBER	13525
935.	GENERATE	13522
936.	REVIEW	13470
937.	CELL	13463
938.	ALONE	13448
939.	EFFORT	13415
940.	IMAGINE	13414

941.	COMBINE	13406
942.	CONCENTRATE	13396
943.	RESOURCE	13378
944.	SORT OF	13361
945.	MALE	13235
946.	COLD	13223
947.	RELEASE	13215
948.	CONNECT	13193
949.	REGULAR	13184
950.	HOTEL	13178
951.	TOUCH	13163
952.	LACK	13160
953.	TRUST	13151
954.	ADVANTAGE	13091
955.	BALANCE	13073
956.	CHIEF	13063
957.	SEAT	13050
958.	DEGREE	13048
959.	COMMUNICATE	13029
960.	ASSIST	13010
961.	OBTAIN	13004
962.	FLOOR	12991
963.	COMMENT	12989
964.	NOVEMBER	12980
965.	WEIGH	12974
966.	RING	12969
967.	THE FOLLOWING	12963
968.	SATISFY	12956
969.	AVOID	12903
970.	SEASON	12899
971.	POPULAR	12876
972.	FILL	12805
973.	PROVISION	12792
974.	EXERCISE	12783
975.	CONCEPT	12768
976.	IN ORDER TO	12762
977.	DEAD	12758
978.	LISTEN	12750
979.	PREFER	12744
980.	ELEMENT	12737
981.	HAVE GOT A	12734
982.	COUNTY	12729
983.	PARTNER	12729
984.	CAMPAIGN	12705
985.	TASK	12702
986.	APPARENT	12691
987.	HALL	12674
988.	COLLEGE	12651
989.	SLEEP	12639
990.	CLOSES	12632
991.	PRIMARY	12609
992.	PREVENT	12604
993.	ADVERTISE	12595
994.	SAMPLE	12586
995.	FREQUENT	12560
996.	SHAPE	12559
997.	CONSUME	12543
998.	REFUSE	12527
999.	PROGRESS	12511
1000.	ESSENTIAL	12509

1001.	PRIME	12500
1002.	PRINT	12486
1003.	HIT	12477
1004.	CONFER	12464
1005.	GLASS	12443
1006.	TITLE	12441
1007.	BLUE	12423
1008.	HORSE	12402
1009.	SLIGHT	12401
1010.	WORRY	12379
1011.	FORGET	12371
1012.	GAIN	12366
1013.	REACT	12363
1014.	STORE	12363
1015.	MENTION	12356
1016.	CONSIDERABLE	12340
1017.	CLEAN	12336
1018.	ADVANCE	12317
1019.	THIRTY	12307
1020.	EXACT	12304
1021.	SURVIVE	12300
1022.	TRANSFER	12290
1023.	PHOTOGRAPH	12271
1024.	HAVE GOT TO ('MUST')	12270
1025.	IMAGE	12232
1026.	DOG	12231
1027.	SURVEY	12205
1028.	PROMOTE	12193
1029.	DRESS	12167
1030.	ADDRESS	12148
1031.	STRATEGY	12144
1032.	COMPLEX	12138
1033.	LEG	12136
1034.	DUTY	12129
1035.	PROPER	12129
1036.	CHAIRMAN	12120
1037.	GUIDE	12076
1038.	JULY	12058
1039.	WARM	12016
1040.	STONE	11979
1041.	REVEAL	11969
1042.	OIL	11967
1043.	ESTIMATE	11917
1044.	BROTHER	11912
1045.	TROUBLE	11909
1046.	BOX	11891
1047.	DOCUMENT	11883
1048.	SCALE	11882
1049.	STAR	11853
1050.	ALTERNATIVE	11845
1051.	RIVER	11834
1052.	CORRECT	11826
1053.	MILITARY	11826
1054.	NOR	11825
1055.	ARMY	11821
1056.	DOUBLE	11819
1057.	WORSE	11816
1058.	LADY	11790
1059.	MEMORY	11782
1060.	THROW	11782

1061.	FORTUNE	11769
1062.	GRAND	11765
1063.	FAST	11723
1064.	MANUFACTURE	11722
1065.	CODE	11674
1066.	GOLD	11607
1067.	ASPECT	11574
1068.	SET UP	11560
1069.	PERFECT	11555
1070.	LIBRARY	11540
1071.	AS TO	11535
1072.	AS WELL	11519
1073.	EXTREME	11519
1074.	SECTOR	11503
1075.	DEMONSTRATE	11489

Appendix 5 – The PHRASE List 'User's Guide'

The PHRASal Expressions (PHRASE) List User's Guide

What is the PHRASE List?

The PHRASE List is a listing of the most common multiword expressions in English of a certain type, intended to be used as a complement to existing lists and instruments of second language instruction that use them.

What type of expression does the list contain?

There are if course many different types of multiword formulaic sequences in English (e.g. idioms, phrasal verbs, binomial expressions, collocation), and an exhaustive listing of those is not the aim of the PHRASE List. Instead, what the PHRASE List intends to provide is a full account of the most common expressions in English that also potentially could cause decoding problems is read word-for-word. For example, even beginners might know the verb *take* and the noun *place*, but may not know the phrasal expression *take place*, which of course has its own meaning.

How was the list made?

The list was created using the 100 million word British National Corpus. First, a full list of all recurring word combinations was extracted, and then the phrasal expressions were chosen from among those line by line and checked against selection criteria. The items selected were also confirmed by an external rater who showed the criteria could be applied consistently.

How can I interpret the frequency information in the list?

The PHRASE List is inclusive of phrasal expressions in the British National Corpus (BNC) to a frequency that matches the frequency of the top 5000 words, also from the BNC. This frequency threshold has been identified as having functional significance for single words (e.g. the ability to read and write at good levels of general proficiency), and the items in the PHRASE List provide one missing lexical element that previous wordlists generally had not accounted for.

Each line in the PHRASE List contains frequency information about each item (Figure 1). The first column, 'Integrated List Rank', shows where the item would rank in a list that integrated both single words and phrasal expressions. This information was included to give the user some perspective regarding just how common each expression is.

The column which immediately follows each phrase, entitled 'frequency', shows how many times that item recurs in the BNC. Each item has been lemmatised where appropriate, so the frequency of *take place*, for example, reflects the combined frequencies of *take place, takes place, took place, taking place* and *taken place*. Moreover, when a phrasal expression could also have a different, more transparent meaning (e.g. *I want to see the circus **when it comes to** town* vs. *My wife is a genius **when it comes to** maths*), only the frequency of the less transparent meaning has been listed.

Finally, there are three columns that follow the 'Frequency' column: 'Spoken general', 'Written general', and 'Written academic' - reflecting three different discourse genres identified by the authors as useful for a broad audience of English learners. At least one column of the genre frequency information contains three stars, indicating the genre in which that phrase is most common. If the phrase occurs equally commonly in more than one genre, then the same amount of stars are given (see 'rather than', Figure 1). An 'x' rather than a star is given to a genre in which the expression in question appears very rarely, or even not at all. In most cases, however, there were various levels of frequency, as in the expression 'so that' (Figure 1), which occurs most commonly in general English conversation, less in written general-purpose English, and even less in academic writing.

Figure 1. Sample lines from the PHRASE List

Integrated List Rank	Phrase	Frequency (per 100 million)	Spoken general	Written general	Written academic	Example
631	**RATHER THAN**	21085	* * *	* * *	* * *	Children, **rather than** adults, tend to learn quickly.
635	**SO THAT**	20966	* * *	* *	*	Park it **so that** the wheels are curbed.

What are the practical applications of the list?

Just as there have traditionally been a number of applications for existing wordlists, the PHRASE List will have the same applications and potentially more. Vocabulary tests, for example, often use wordlists with the same type of frequency information that the PHRASE List uses in order to obtain a sampling of the test taker's breadth of vocabulary knowledge. The PHRASE List could also be applied for the same purpose, and even integrated into existing tests.

For language teaching practitioners (e.g. teachers, syllabus designers), the PHRASE List provides one means of prioritising and ensuring the inclusion of important items of vocabulary that perhaps have not been systematically integrated to language instruction previously.

Finally, there are plans already under way to develop a text analyser that would allow the user to download any text and check which phrases at the various frequency levels appear in it, which in turn can be used to help arrive at informed decisions regarding the relative difficulty of the text.

What are the limitations of the PHRASE List?

Like most tools, the PHRASE List should be used appropriately and selectively. First of all, the list is derived from a corpus, and no matter what the size, no corpus is 100% reflective of all language use, but is an artefact of the texts which it comprises and the aims of the compilers of the corpus. Hence, since the PHRASE List is derived from the BNC, and the BNC is reflective of mostly written British English only as modern as the early 1990s, clearly the list would be of very limited use for someone wanting to know which phrases are most common in spoken American casual conversation. However, as long as the user is mindful of these limitations and heeds the genre information provided in the list, the PHRASE List should prove useful for many English language teaching professionals and students alike.

Appendix 6 – Phrases common to both the English Profile Wordlist and the PHRASE List

Phrase	EP Level	BNC Frequency	Frequency Band
a lot	A1	22332	1K
as well (as)	A1	18041	1K
half past	A1	1325	4K
of course	A1	26966	1K
there is/are	A1	59833	1K
Would you like...?	A1	1133	5K
a bit	A2	19618	1K
a variety of	A2	4283	3K
all sorts of	A2	1535	4K
all the time	A2	3527	3K
and so on	A2	4584	2K
as usual	A2	1287	4K
at least	A2	4306	3K
at the moment	A2	5001	2K
by the way	A2	1433	4K
come back	A2	6772	2K
each other	A2	10160	2K
fill in	A2	1187	5K
for sale	A2	1379	4K
get back	A2	3178	3K
get on (sth)	A2	3656	3K
I mean	A2	23616	1K
if you like	A2	1256	4K
I'm afraid	A2	1495	4K
look after	A2	4332	3K
make sure	A2	5510	2K
never mind	A2	915	5K
oh dear	A2	1700	4K
oh no!	A2	2642	3K
or something	A2	2683	3K
quite a lot	A2	1260	4K
such a(n)	A2	23894	1K
such as	A2	30857	1K
these days	A2	2440	3K
too much	A2	7123	2K
What about...?	A2	3160	3K

above all	B1	2212	3K
after all	B1	5197	2K
aim to	B1	3415	3K
all over (places)	B1	4420	3K
apart from	B1	6287	2K
as long as	B1	5084	2K
as soon as	B1	5323	2K
at all	B1	14650	1K
at first	B1	4275	3K
at once	B1	3684	3K
at present	B1	2847	3K
at the same time	B1	2892	3K
break up	B1	1595	4K
carry on	B1	3759	3K
carry out	B1	10753	2K
come on	B1	4519	2K
consist of	B1	5362	2K
deal with	B1	13634	1K
due to	B1	10454	2K
end up	B1	3285	3K
ever since	B1	1622	4K
feel like	B1	2431	3K
find out	B1	6499	2K
for instance	B1	7138	2K
get on (RELATE)	B1	944	5K
give up	B1	3997	3K
go back	B1	3722	3K
go on (CONT.)	B1	10610	2K
hang on	B1	2074	4K
How about...?	B1	805	5K
if only	B1	1830	4K
in addition	B1	7822	2K
in advance	B1	1983	4K
in case	B1	2536	3K
in detail	B1	1473	4K
in fact	B1	15983	1K
in full	B1	1698	4K
in order to	B1	12762	1K
in particular	B1	7092	2K
in spite of	B1	2676	3K
in the end	B1	3050	3K
in time	B1	3566	3K
just about	B1	1622	4K
keep on	B1	788	5K
look forward to	B1	2331	3K
look like	B1	6595	2K
make up one's mind	B1	788	5K
no longer	B1	8556	2K
oh well	B1	2129	3K
on board	B1	1698	4K

on the whole	B1	1238	4K
once again	B1	3532	3K
once more	B1	2146	3K
one another	B1	2623	3K
or so	B1	4164	3K
ought to	B1	5002	2K
over there	B1	2678	3K
rather than	B1	21085	1K
set off	B1	1644	4K
set out	B1	4624	2K
so far	B1	5018	2K
straight away	B1	794	5K
switch on	B1	1935	4K
take care of	B1	1034	5K
take part	B1	2374	3K
take place	B1	10556	2K
that sort of thing	B1	843	5K
the other day	B1	1066	5K
up to (MAX.)	B1	8733	2K
up to date	B1	1268	4K
used to (PAST)	B1	14411	1K
(be) about to	B2	4600	2K
(with) regard to	B2	2630	3K
a good/great deal	B2	5126	2k
a handful of	B2	965	5K
a little	B2	20296	1K
add to	B2	1424	4K
as a result	B2	7939	2K
as a whole	B2	3615	3K
as follows	B2	2620	3K
as for	B2	3157	3K
as if	B2	14470	1K
as though	B2	4988	2K
at risk	B2	1419	4K
back up	B2	1042	5K
bear in mind	B2	1398	4K
by far	B2	925	5K
can tell	B2	1011	5K
catch up	B2	1095	5K
come across	B2	1362	4K
even though	B2	5664	2K
fail to	B2	10263	2K
find oneself	B2	3917	3K
focus on	B2	3703	3K
follow up	B2	2128	3K
for the moment	B2	1092	5K
for the sake of	B2	865	5K
from time to time	B2	1627	4K
get away with	B2	824	5K

get away	B2	1165	5K
get to (do sth)	B2	1437	4K
go through	B2	5857	2K
hand over	B2	1798	4K
if so	B2	1830	4K
in any case	B2	2159	3K
in other words	B2	3159	3K
in practice	B2	3609	3K
in return	B2	1720	4K
in the first place	B2	1897	4K
in the meantime	B2	1161	5K
in theory	B2	1065	5K
in this respect	B2	1110	5K
in view of	B2	1477	4K
kind of	B2	3510	3K
lead to	B2	13555	1K
make out	B2	842	5K
make sense	B2	1608	4K
may well	B2	4931	2K
might as well	B2	1348	4K
more or less	B2	2579	3K
no matter	B2	1888	4K
no wonder	B2	820	5K
not only	B2	14110	1K
on average	B2	1014	5K
on behalf of	B2	6734	2K
on the one hand...	B2	1406	4K
point out	B2	6325	2K
rely on	B2	3488	3K
result in	B2	5763	2K
run out (ALL USED)	B2	1430	4K
shake one's head	B2	3250	3K
shut up	B2	1079	5K
sort of	B2	13361	1K
sort out	B2	2696	3K
stand for (REPRESENT)	B2	977	5K
take account of	B2	1217	4K
take for granted	B2	1120	5K
take over	B2	5394	2K
tend to	B2	10504	2K
the following	B2	12963	1K
the former	B2	1842	4K
the latter	B2	5519	2K
the sight of	B2	1024	5K
to date	B2	2600	3K
to some extent	B2	1688	4K
turn down	B2	951	5K
with respect to	B2	1325	4K
work out	B2	4432	3K

Appendix 7 – Phrases missing in English Profile Wordlist

(This is a part of a summary report on the analyses conducted. In the left-hand column of the table below, the full list of all the phrases that were found to be 'missing' [i.e. in the PHRASE List but not in EPW], with further observations in the adjacent columns.)

Re-analysis of 'missing' phrases

The expressions below were multiword items from the PHRASE List originally identified as not present in the English Profile Wordlists. They are re-categorized below and colour-coded into 'not present' (red), 'in list but embedded' (yellow) and 'already in list as headword' (green). (See 'Colour Key' below.) As suspected, many of the truly missing phrases are actually likely candidates for the C Levels. There are many items that were 'hidden' in the current EP lists, however – highlighted in yellow below – that may merit reconsideration as perhaps better serving the list as phrases with their own entries. Moreover, a number of the phrases found to be embedded in senses (again, in yellow) are suspected of actually being outside of the proficiency range of the sense identified in the list.

As before, the phrases are ranked in ascending frequency order, with a minimum level of 787 representing the upper range of the BNC 5000 word family level. Note that a frequency of around 1700 was originally suggested as a possible cut-off point for the C1-C2 levels, but in the light of the many items that are scattered above that frequency range that may be useful for the list, it could be that such a putative limit is actually counterproductive.

COLOUR KEY

RED = NOT PRESENTLY IN EP LIST

YELLOW = IN LIST, BUT EMBEDDED IN OTHER SENSE(S)

GREEN = ALREADY A HEADWORD IN LIST

PHRASE List phrase	EP LOCATION	BNC Frequency (per 100 million)
AT WORK ('operating')		787
COME ABOUT	Meaning 'come to pass'	787
LAY OUT	This in the meaning 'Lay out plans' or 'Lay out a vision', etc.	789
THE LOT	Meaning 'everything'	789
LOOK TO ('want')	As in 'He's looking to get married'.	790

HOLD UP	B1	791
OWING TO		793
TO GO ('REMAINING')	EXAMPLE UNDER 'TIME' – B2	801
WELL BEING		809
TURN BACK		827
WITH A VIEW TO		829
FOND OF	B1	831
COME TO TERMS WITH		839
AT BEST		844
WEALTH OF		844
BACKED BY		850
TO COME (FUTURE)	EXAMPLE UNDER 'HAPPEN' – B2 ('...still to come') THIS PHRASE OFTEN OCCURS AS A TIME ADVERBIAL WITHOUT 'STILL': 'In the weeks to come...'	862
JUST AS ('WHEN')		869
NO SIGN OF	EXAMPLE UNDER 'SHOWING' – B1 ('no sign of') but this meaning does not really hold in sentences like 'there's was no sign of him'.	871
IN WHICH CASE		872
COME UP TO	B2 ('MOVE TOWARDS')	881
MORE SO		887
A CASE OF	As in 'It's simply a case of studying harder'.	888
IN ONE'S OWN RIGHT		891
UP AND DOWN	As in 'He was pacing up and down the hallway'.	898
A DEGREE OF		921
ON THE MARKET		928
IN THE INTEREST OF		938
UNDER WAY	A2 – UNDER 'ROUTE'. The example is 'The coach stopped for us to eat lunch but within half an hour we were **on** our way/**under** way again. I think it's highly unlikely that this is actually A2, and should probably have its own entry anyway, I think.	939
BY VIRTUE OF		954
(AT) THE OUTSET		963
THE BULK OF		968
TO BLAME	B1 – Example: 'Poor housing is to blame' → this particular formulation may actually be of a higher proficiency	973
BETTER OFF		980
TOUCH OF		982
AS IT WERE		985
AND ALL THAT		989
FREE FROM		991

AT ONE TIME	As in, 'It was, at one time, a very nice place.'	1014
HEAD TO	B2 – UNDER 'HEAD', BUT MISSING THE PARTICLE 'TO'	1014
IN A SENSE		1014
THE WHOLE THING	In the sense of 'everything' (e.g. 'He ate the whole thing by himself'), I believe the phrase merits its own entry.	1015
GO ROUND	A2 – UNDER 'VISIT' ('I'll go round and see her later.')	1022
BACK UP		1042
AS GOOD AS ('LIKE')		1043
BOTHER TO	B2 – UNDER 'MAKE AN EFFORT' ('He hasn't even bothered to write'). Would be interesting to see if 'bother to' actually is used a lot in B2.	1046
COMMON SENSE	B1 – UNDER 'GOOD JUDGMENT'. I really do think this deserves a separate entry.	1049
OPPOSED TO		1059
FOR GOOD		1061
THOUGHT OF (AS)		1065
WOULD APPEAR	B1 – UNDER 'SEEM': 'It would appear that nobody…'. →My guess is that 'WOULD APPEAR' actually pertains to a register that is not as common as B1.	1068
AS OF		1069
OVER TIME	B1 – UNDER 'DURING'. The example is 'It's fascinating to watch how a baby changes and develops over time.' I believe that 'over lunch' and 'over the summer' do mean 'during lunch' and 'during summer', but '…develops over time' does not mean '…develops during time'.	1069
OF LITTLE		1079
PUT TOGETHER		1082
THINGS LIKE THAT		1082
AT THE EXPENSE OF		1086
A GO ('ATTEMPT')	B2 – UNDER 'ATTEMPT'	1093
ALLOW FOR ('CALCULATE IN')		1105
PROVIDED THAT	B2	1110
IN NEED		1122
LITTLE MORE THAN		1133
SOMETHING OF A		1154
THE ODD		1154
FOR LIFE		1172
A QUESTION OF		1174

FOR ALL ('CONSIDERING')	As in, 'For all the hoopla, I was expecting much more.'	1182
WHEN IT COMES TO		1188
MAKE USE OF		1191
SOMETHING LIKE THAT	A2	1192
CARE TO		1223
IN LINE WITH		1240
GIVEN THAT		1247
SO FAR AS		1251
TO THE EXTENT		1253
LET AL.ONE		1261
IN CONJUNCTION WITH		1270
LONG AGO		1273
LONG BEFORE		1282
DAY TO DAY		1289
ACT ON		1296
EXCEPT THAT	A2 → This actually deceptively simple. Consider the example 'I want to go, except that I'm tired.' This does not really fit the definition offered in the dictionary.	1296
A GOOD ('AT LEAST')		1298
CONTRARY TO		1303
THIRD PARTY		1307
WAY OUT	As in 'He lives way out in Nottingham.'	1309
CONSISTENT WITH		1315
AT A TIME ('SIMULTANEOUSLY')	B2 – Listed as 'ONE AT A TIME' (meaning 'separately'), but, curiously, examples of 'two at a time', 'three at a time' and so on are not provided. Perhaps a separate entry?	1340
MIND YOU		1342
TO ME ('IN MY OPINION')		1345
LIMITED TO	As in 'Eligibility is limited to those who have not previously applied.' (i.e. 'restricted to')	1346
FIRST OF ALL	A2 – UNDER 'BEFORE EVERYTHING': 'First of all, check you have...'. The LEARNER EXAMPLE, however, is only 'first', so I wonder about the representativeness here.	1352
COME ACROSS	B2	1362
EVEN SO		1363
PROVIDE FOR		1370
MOST LIKELY		1376
OR ANYTHING		1379
OLD FASHIONED	B1	1393

ON THE ROAD		1400
BY WAY OF		1404
SHOWN TO	It's not clear that the learner example under B2 'PROVE', 'Experiments have even shown some astonishing results' is as complex as the slightly more opaque passive construction, e.g. 'The medicine is shown to work'. I have a feeling that it shows up later (i.e. C level).	1409
A MERE	B2 : 'It cost a mere twenty dollars'. To me, this item is synonymous to 'ONLY'. In the corpus, the use of 'THE MERE' and 'A MERE' are actually slightly different.	1410
AS YET		1423
IN PRINCIPLE		1429
BY CONTRAST	B2 – I think this deserves its own entry.	1430
PARTY TO		1439
IN HAND		1443
MAKE ITS/ONE'S WAY	A2 – UNDER 'ROUTE'. It seems this sense may be quite over-conflated, as another item – 'under way' – is also found here (see above). I wouldn't surprised is this is actually a C-level candidate.	1446
TO THE POINT		1447
NEXT DOOR		1449
OUT OF ('DUE TO')	B2	1453
SUCH THAT		1454
IN THE FACE OF		1457
NOTHING BUT	A2 – UNDER 'NOT ANYTHING': 'She did nothing but criticize'. I can't really imagine an A2-level candidate writing or saying this, to be honest. The construction is not that transparent, with 'but' not holding its usual meaning here. The phrase is really a synonym for 'only'.	1466
REFLECTED IN		1471
THAT WHICH		1491
KEY TO		1499
YET ANOTHER		1500
NO GOOD	B2	1528
IN THE ABSENCE OF		1538
SOME MORE	A1 – UNDER 'UNKNOWN AMOUNT': 'I've got some more work before I can go out'. Grammatically speaking,	1539

	this is actually quite complex, and I'm pretty sure not A1.	
(BE) RUN BY	(Although this could possibly fit under 'ORGANIZE')	1540
FOR LONG	A2 – UNDER 'TIME': 'Have you been waiting **for long**?' This is more difficult than the learner example, 'It doesn't last longer than 45 minutes'. Another example from the dictionary that probably doesn't fit here: 'We've been walking **all day long'**.	1553
A LONG WAY	B1 – UNDER 'DISTANCE'. But the figurative use, as in 'A little bit can go a long way' is not listed.	1557
GOOD AT	A1 – UNDER 'SUCCESSFUL': 'She's very good at geography'. The learner example, 'He's [a] good football player' is clearly of a less complex construction.	1562
ALL TOO		1571
BUT THEN (AGAIN)		1589
A BIT OF A		1599
BY MEANS OF	B2 - 'She tried to explain by means of sign language'. The sense offered in the learner example is a much more common and concrete one: 'This is the cheapest means of transport'. This phrase is much more likely a C-level candidate.	1602
IN SHORT		1602
LARGE SCALE	B2 – UNDER 'SIZE'	1610
GIVE RISE TO		1613
THE MEANS (e.g. money)		1626
PUT FORWARD		1640
DO(ING) SO		1647
FACED WITH	B2 – UNDER 'PROBLEM'. Where this phrase gets especially tricky is when it starts a sentence: 'Faced with the truth, he decided to change his mind'.	1658
HELD THAT ('BELIEVED')		1661
HAPPEN TO (BE)		1664
GREATER THAN		1666
NO IDEA	B1	1670
SOME KIND OF		1678

------------------- PREVIOUS C1/C2 LEVEL CUT-OFF POINT -----------------

ON THE GROUNDS	B2 – This might be useful on its own.	1713
TO DEATH	B2	1714
FOR SOME TIME	B2 – UNDER 'LARGE AMOUNT'	1725
THE EXTENT TO WHICH		1751
THAT MUCH		1781
IN THE SAME WAY		1784
IN THE LIGHT OF		1789
OR WHATEVER		1806
UP TO (DECISIONS)	B1	1807
YET TO		1818
IN ITSELF		1855
AT THIS POINT ('now')		1884
IN COMMON	B1 – 'HAVE STH IN COMMON'	1890
PROVE TO BE	B2 – UNDER 'RESULT'.	1893
IN QUESTION ('under consideration')		1898
COULD HARDLY	B2 – UNDER 'CERTAINLY NOT'	1911
BY NO MEANS		1925
OVER THE YEARS		1942
SHORT OF	As in 'Short of selling the house, we have no money left.'	1944
SO AS TO		1954
ON THE PART OF		1965
SIGHT OF	B2	1988
AFFORD TO	As in 'We can't afford to be late.'	1989
IN EFFECT	As in, 'It is, in effect, a completely new system.'	1995
ALL THE WAY		2007
AT TIMES		2014
CALL ON		2016
IN ACCORDANCE WITH		2017
BRING ABOUT		2022
HAD BETTER	A2	2022
THINK SO		2033
BY NOW		2044
SOMETHING ABOUT	As in 'there's something about Mary'.	2071
TURN INTO	B1	2072
MEANT TO	As in 'It's mean to be sunny next week.'	2098
FOUND TO		2104
WOULD SAY		2128
THANKS TO		2159
ALL BUT		2214

THIS STAGE ('now')		2223
IN CONTRAST (TO)	B2 (see 'BY CONTRAST' earlier in this list)	2229
MOVE ON		2245
SET TO		2258
AS SUCH		2290
PART TIME	B1	2343
IN SO FAR AS		2344
HEARD OF	B2	2403
IN CHARGE		2432
PUT IT ('SAY')		2449
HAVE A LOOK	B1 – UNDER 'SEE': 'Can I have a look at your dictionary?' I think this could really be its own entry.	2464
LED BY	As in 'She was led by greed.'	2511
OUT THERE		2513
AIMED AT	B2	2573
SHORT TERM	B2	2574
IN THE COURSE OF		2585
THE ABOVE	B1 (but 'THE ABOVE' in this sense is more of an anaphor)	2608
IN MIND	B2	2638
IN PART		2652
THAT'S IT (i.e. 'that's all)		2674
OVER THERE	B1	2678
IN A WAY		2684
FULL TIME	B1	2761
NO DOUBT ('SURELY')		2791
IN PLACE		2805
WHETHER OR NOT	B1 – UNDER 'IF': 'It all depends on whether or not she's got the time'.	2824
POINT OF VIEW	B2	2864
OUT OF ('USING')	B1	2907
IN RESPECT OF		2909
SO CALLED	B2	2944
THEY SAY		2962
IN THE EVENT (OF)		2998
IN THE WAY	B2 – UNDER 'FREE SPACE': 'I couldn't see the stage because there was a pillar in the way'. I think this deserves to feature more prominently in the list, really.	3013
IN TOUCH (WITH)	B1	3060
KNOWN TO ('notorious for')		3091
SOMETHING LIKE ('AROUND')		3092
CHOOSE TO	A1 – 'Katie chose to stay	3099

	away from work that day'. I really think that this 'choose' is somewhat different from the sense in the list. 'Choose to' is more like 'to resolve to' do something. (And hence not likely that common in A1, I would imagine.)	
PRIOR TO		3110
NOT EVEN		3128
NO MORE THAN ('ONLY')		3226
APPEAL TO	B2 – UNDER 'ATTRACT': 'Cycling has never really appealed to me'.	3299
ON THE BASIS (OF)	B2 – UNDER 'REASON': 'Marks are awarded on the basis of progress and performance'. I think this is common enough as a form that it should be pulled out as especially useful.	3515
IN TURN		3558
SAID TO BE		3586
LOTS OF	A1	3605
BY THE TIME		3607
IT TAKES	There is something similar under 'TIME', but 'IT TAKES' is about what is 'necessary' in general (e.g. 'It takes patience.')	3670
GET UP	A1	3857
IN FAVOUR	B2	4073
AT LAST	B1	4306
OUT OF ('IN'/'FROM')	B1 (in 'FROM AMONG')	4361
OTHER THAN		4380
SUPPOSED TO	B1	4586
CONCERNED WITH		4619
THE CASE ('TRUE')	B1	4794
IN (THE SENSE) THAT		4805
FOLLOWED BY	B1 – UNDER 'HAPPEN AFTER': 'There was a bang, followed by a cloud of smoke'.	4816
ALONG WITH		4948
COME TO ('EVOLVE TO')		4970
ON THE WAY	A2 – UNDER 'ROUTE' (AGAIN – SEE OTHER ITEMS UNDER THIS HEADING): 'We'll have to stop for fuel on the way to the airport'. This, I believe, is a useful item to single out.	5085

TO DO WITH	B2	5184
SUBJECT TO		5218
ON ONE'S OWN	B1	5240
AT THE TIME ('WHEN THIS HAPPENED') - anaphor	A2 – UNDER 'OCCASION': 'We were very young at the time'. This is quite different from the learner example, 'We had a good time and all the guests were happy'.	5282
A RANGE OF	B1 – UNDER 'OF THINGS'	5651
THAT IS (REPHRASING)		5737
LONG TERM	B2	5831
SEEK TO		5937
BE EXPECTED TO	B2 – UNDER 'BEHAVE'	5964
A FURTHER ('ANOTHER')		6121
OR TWO	B1 – the phrase 'ONE OR TWO' appears in the list, but this phrase is more like 'I need to say word or two about dress code'.	6192
MANAGE TO	B1 – UNDER 'SUCCEED'	6234
CALL FOR STH		6243
INSTEAD OF	A2	6907
A COUPLE OF	B1 – UNDER 'SOME'	7007
YOU SEE	The sense 'UNDERSTAND' (B1) comes close, but you see the difference in utterances like 'You see, the thing I don't like about Sao Paulo is the traffic'.	7102
THINK ABOUT	A2 – UNDER 'CONSIDER DOING': 'I'm thinking about buying a new car'.	7243
WORK ON		7600
LAST NIGHT	'LAST WEEK/YEAR' etc. is in the list, but not 'LAST NIGHT' (which probably should at least be one of that bunch, if not a separate item).	7992
LOOK FOR	A1 – UNDER 'SEARCH': 'I'm looking for my keys'.	8377
A SINGLE ('ANY')	B2 – UNDER 'ONE': 'Not a single person offered to help her'. Since this form does not change (i.e. it is always 'A SINGLE', and as it is not likely to be salient and/or obvious in the input, it might be worth singling it out as a phrase.	8710
NO ONE	A2	9597

BASED ON	B1	11440
AS TO		11535
HAVE GOT TO	A2	12270
HAVE GOT (+NP)	B1	12734
THOSE WHO		13951
A NUMBER OF	B1 UNDER 'AMOUNT': 'There were a number of soldiers present at the rally'. (Well, this one we know about!)	15090
BE LIKELY TO	B1	15854
SO THAT	B1	20966
A FEW	A2	26451
GOING TO (FUTURE)	A2	28259
HAVE TO	A2	83092

Appendix 8 – The Phrasal Vocabulary Size Test

A vocabulary size test of multiword expressions

Choose the right phrase to go with each meaning and tick the correct answer. If you have no idea about the meaning of a phrase, do not guess. (Begins on following page.)

First 1000

1.	go on: It will **go on**.	
	a.	sleep
	b.	repeat
	c.	be fast
	d.	continue

2.	lead to: No one knows what it will **lead to.**	
	a.	want
	b.	have inside
	c.	cause in the future
	d.	find

3.	so that: He sat **so that** they could do it.	
	a.	to make it possible that
	b.	because
	c.	very slowly and then
	d.	before

4.	at all: I don't like it **at all.**	
	a.	all the time
	b.	in any way
	c.	at first
	d.	sometimes

5.	I mean: Two, **I mean**, three.	
	a.	I am guessing
	b.	maybe
	c.	then later
	d.	I correct myself

6.	at least: **At least** it is warm.	
	a.	other things may be bad, but
	b.	many days have passed and now
	c.	I cannot believe that
	d.	the least important thing is

7.	is likely to: He **is likely to** go.	
	a.	likes to
	b.	can
	c.	wants to
	d.	probably will

8.	is to: He **is to** speak this afternoon.	
	a.	will
	b.	can
	c.	wants to
	d.	may

9.	deal with: I can **deal with it.**	
	a.	fix
	b.	try
	c.	find
	d.	see

10.	used to: I **used to** go.	
	a.	want to
	b.	did before
	c.	usually
	d.	always

Second 1000

1.	so: It's good **so far.**	
	a.	until now
	b.	but not really
	c.	sometimes
	d.	from a distance

2.	to do with: It is **to do with** money.	
	a.	making
	b.	for
	c.	about
	d.	our

3.	take over: They will **take over.**	
	a.	be finished
	b.	have control
	c.	come later
	d.	think about it

4.	in particular: I want that **in particular.**	
	a.	especially
	b.	in private
	c.	because it is different
	d.	maybe

5.	for instance: **For instance**, it is cheaper.
a.	maybe
b.	for a short time
c.	In my opinion
d.	as an example

6.	as a result: **As a result** it was done.
a.	no person knows if
b.	after a long time
c.	before that
d.	because of that

7.	as soon as: I'll go **as soon as** I can.
a.	from the moment
b.	only if
c.	after
d.	before

8.	carry out: It was **carried out** yesterday.
a.	started
b.	found
c.	read
d.	done

9.	be about to: I **am about to** read the newspaper.
a.	cannot wait to
b.	am soon going to
c.	really like to
d.	am trying to

10.	be expected to: We **are expected to** do it.
a.	are waiting
b.	hoping to
c.	must
d.	are able to

Third 1000

1.	give up: I **give up**.
a.	try very hard
b.	am starting
c.	will now stop
d.	exercise

2.	feel like: I just did not **feel like** it.
a.	love
b.	want to do
c.	think about
d.	try to do

3.	turn out: It **turned out** different.
a.	started
b.	seemed
c.	became
d.	did not look

4.	other than: **Other than** that, it's good.
a.	not including
b.	if you include
c.	because of
d.	after

5.	get to: She **got to** the car.
a.	arrived at
b.	drove
c.	received
d.	entered

6.	all over: It is **all over** the bed.
a.	covering
b.	inside
c.	on top of
d.	beside

7.	in touch: Keep **in touch**.
a.	feeling it
b.	communicating
c.	pushing it
d.	thinking

8.	get rid of: They **got rid of** it.
a.	decided to not have
b.	received
c.	became bored with
d.	chose

9.	at once: I did it **at once**.
a.	one time
b.	many times
c.	early
d.	immediately

10	in time: **In time** they bought a house.
a.	quickly
b.	earlier
c.	eventually
d.	recently

Fourth 1000

1.	prove to be: It has **proved to be** important.
a.	will possibly become
b.	shown itself to be
c.	continued to be
d.	never been

2.	next door: It's just **next door**.
a.	coming soon
b.	common
c.	perfect
d.	very close

3.	run out: I think we **ran out** of it.
a.	had no more
b.	were bored
c.	thought
d.	moved outside

4.	take advantage: You must **take advantage**.
a.	go slowly
b.	use the opportunity
c.	pay attention
d.	relax

5.	in effect: It is **in effect** the reason.
a.	possibly
b.	not
c.	now
d.	actually

6.	in the light of: It was accepted **in the light of** the money.
a.	despite
b.	because of
c.	in addition to
d.	instead of

7.	by no means: He is **by no means** rich.
a.	very
b.	not at all
c.	more or less
d.	considered

8.	come across: They **came across** a hotel.
a.	stayed in
b.	opened
c.	were near
d.	found

9.	happen to: She **happened to** call.
a.	pretended
b.	tried hard to
c.	did not want to
d.	by chance did

10.	even so: **Even so** it's better.
a.	despite that
b.	that way
c.	it is the same and
d.	maybe

Fifth 1000

1.	by far: She is **by far** the most intelligent.
a.	trying to be
b.	not at all
c.	really
d.	sometimes

2.	come up to: He just **came up to** me.
a.	approached
b.	rejected
c.	did not like
d.	copied

3.	straight away: They did it **straight away**.
a.	immediately
b.	the correct way
c.	slowly
d.	because they wanted to

4.	would appear: It **would appear** it's true.
a.	cannot be that
b.	is certain that
c.	seems that
d.	is assumed that

5.	turn down: She **turned down** the money.	
	a.	hid
	b.	lost
	c.	made
	d.	refused

6.	to blame: We are not **to blame**.	
	a.	in total agreement
	b.	interested
	c.	accusing anyone
	d.	the cause of the problem

7.	take for granted: She **took it for granted**.	
	a.	kept it
	b.	did not give it importance
	c.	wanted it a lot
	d.	thought about it carefully

8.	as of: It changes **as of** today.	
	a.	starting
	b.	sometime
	c.	perhaps
	d.	because of

9.	can tell: You **can tell**.	
	a.	may speak
	b.	are smart
	c.	can see
	d.	might

10	over time: **Over time** it was cheaper.	
	a.	long ago
	b.	eventually
	c.	when it was too late
	d.	at the perfect moment

- END OF TEST-

Appendix 9 – Phrasal VLT pilot test (three versions)

Vocabulary Levels Test
– Phrasal Expressions -

This is a vocabulary test. Choose the right phrase to go with each meaning. Write the number of that phrase next to its meaning. Here is an example.

1 a little		
2 in fact	______	not much
3 too much	______	yesterday at night
4 instead of	______	really
5 apart from		
6 last night		

1 a little		
2 in fact	1	not much
3 too much	6	yesterday at night
4 instead of	2	really
5 apart from		
6 last night		

Some phrases in the test are extra. In the example above, these phrases are *too much*, *instead of*, and *apart from*. If you have no idea about the meaning of a phrase, do not guess. But if you think you know the meaning, then try to find the answer.

(CONTINUED ON NEXT PAGE)

Version A

Phrasal Expressions (first 5,000 words)

First 1,000

1 go on
2 lead to
3 deal with
4 have to
5 I mean
6 be going to

_____ must
_____ will
_____ continue

1 such as
2 at least
3 rather than
4 a number of
5 at all
6 sort of

_____ in any way
_____ many
_____ for example

From 1,000 to 2,000

1 as soon as
2 no longer
3 a range of
4 on your own
5 all right
6 and so on

_____ alone
_____ when something begins to happen
_____ many

1 set up
2 carry out
3 take place
4 look like
5 find out
6 manage to

_____ learn or discover
_____ happen
_____ be able

From 2,000-3,000

1 look after
2 find oneself
3 it takes
4 take into account
5 sort out
6 feel like

_____ want to
_____ solve
_____ is necessary

1 a variety of
2 or so
3 in time
4 whether or not
5 in a way
6 above all

_____ if
_____ many
_____ maybe a little more or less

From 3,000-4,000

1 by no means
2 in the light of
3 by the way
4 as yet
5 let al.one
6 on the whole

_____ taking something into account
_____ in general
_____ still has not happened

1 call on
2 take advantage
3 prove to be
4 keep up
5 give rise to
6 run out of

_____ cause something to happen
_____ eventually become
_____ have no more of something

From 4,000-5,000

1 take for granted
2 provided that
3 stand for
4 to blame
5 get away with
6 keep on

_____ as long as
_____ not value someone or something
_____ continue

1 something of a
2 little more than
3 for the moment
4 as of
5 for good
6 by far

_____ starting at a certain point in time
_____ permanently
_____ only

Version B

First 1,000

1 is likely to
2 have got ______ did in the past
3 deal with ______ probably will
4 used to ______ try to solve a problem
5 there is
6 is to

1 of course
2 a lot ______ much or many
3 so that ______ not much
4 a bit ______ also
5 as well as
6 as if

From 1,000 to 2,000

1 as to
2 a couple of ______ until now
3 as a result ______ because of that
4 for instance ______ more than one
5 long term
6 so far

1 have got to
2 result in ______ cause
3 point out ______ related to
4 take over ______ show
5 to do with
6 get out

From 2,000-3,000

1 turn out
2 work out ______ indicate your answer is no
3 carry on ______ try to see or inspect something
4 shake your head ______ continue
5 have a look
6 look forward to

1 all over
2 at last ______ immediately
3 at once ______ again
4 in touch ______ everywhere
5 all but
6 once more

From 3,000-4,000

1 all the way
2 in effect ______ regardless of
3 in advance ______ modern or recent
4 up to date ______ actually
5 no matter
6 in the first place

1 happen to
2 make sense ______ by chance
3 amount to ______ find
4 get on with ______ become
5 come across
6 bear in mind

From 4,000-5,000

1 make use of
2 would appear ______ seem
3 bother to ______ reject
4 turn down ______ approach a person or place
5 come up to
6 come to terms with

1 at one time
2 as it were ______ at fault
3 under way ______ in the past, but no more
4 at best ______ already started
5 to blame
6 straight away

Version C

First 1,000

1 lead to
2 I mean _____ try to solve a problem
3 have got _____ to say it a different way
4 there is _____ cause
5 is to
6 deal with

1 at least
2 rather than _____ as would be expected
3 sort of _____ the minimum
4 of course _____ similar to, but not exactly
5 so that
6 as if

From 1,000 to 2,000

1 no longer
2 all right _____ OK
3 and so on _____ as an example
4 as to _____ et cetera (etc.)
5 for instance
6 long term

1 set up
2 carry out _____ control
3 look like _____ must
4 have got to _____ do
5 take over
6 get out

From 2,000-3,000

1 look after
2 find oneself _____ becomes the end result
3 turn out _____ try to understand
4 take into account _____ to consider as part of an analysis
5 work out
6 look forward to

1 in time
2 in a way _____ eventually
3 above all _____ communicating
4 at last _____ most important
5 in touch
6 all but

From 3,000-4,000

1 by no means
2 by the way _____ before
3 let al.one _____ not at all
4 all the way _____ completely
5 in advance
6 most likely

1 call on
2 take advantage _____ to consider
3 keep up _____ logical
4 make sense _____ use an opportunity
5 get on with
6 bear in mind

From 4,000-5,000

1 stand for
2 catch up _____ make an effort to do something
3 get away with _____ accept a situation
4 make use of _____ represent
5 bother to
6 come to terms with

1 something of a
2 for the moment _____ temporarily
3 by far _____ immediately
4 as it were _____ much more
5 at best
6 straight awa

Appendix 10 – Prototype version of Phrasal Vocabulary Size Test

Name: ______________ Group: ____________ Sex: ____________ Date: ____________

First language: ____________ Age: ________ email (optional): ____________________

A vocabulary size test of multiword expressions

This is a vocabulary test. Choose the right phrase to go with each meaning and tick the correct answer. Here is an example.

1.	have got: He **has got** it.	
	a.	bought
	b.	has
	c.	received
	d.	wants

1.	have got: He **has got** it.	
	a.	bought
✓	b.	has
	c.	received
	d.	wants

If you have no idea about the meaning of a phrase, do not guess. But if you think you know the meaning, then try to find the answer.

(THIS PAGE HAS BEEN LEFT INTENTIONALLY BLANK)

- TEST BEGINS ON FOLLOWING PAGE -

First 1000

1.	lead to: No one knows what it will **lead to.**	
	a.	want
	b.	have inside
	c.	cause in the future
	d.	find

2.	have to: You **have to** go.	
	a.	must
	b.	want to
	c.	can
	d.	will

3.	a number of: **A number of** people came.	
	a.	almost no
	b.	several
	c.	no
	d.	two

4.	go on: It will **go on.**	
	a.	sleep
	b.	repeat
	c.	be fast
	d.	continue

5.	a bit: It's **a bit** different.	
	a.	very
	b.	a little
	c.	sometimes
	d.	always

6.	is likely to: He **is likely to** go.	
	a.	likes to
	b.	can
	c.	wants to
	d.	probably will

7.	deal with: I can **deal with** it.	
	a.	fix
	b.	try
	c.	find
	d.	see

8.	at all: I don't like it **at all.**	
	a.	all the time
	b.	in any way
	c.	at first
	d.	sometimes

9.	is to: He **is to** speak this afternoon.	
	a.	will
	b.	can
	c.	wants to
	d.	may

10.	a lot: I go there **a lot.**	
	a.	sometimes
	b.	often
	c.	never
	d.	alone

Second 1000

1.	as soon as: I'll go **as soon as** I can.	
	a.	from the moment
	b.	only if
	c.	after
	d.	before

2.	find out: How did you **find out?**	
	a.	arrive
	b.	like it
	c.	get the information
	d.	leave

3.	so: It's good **so far.**	
	a.	until now
	b.	but not really
	c.	sometimes
	d.	during the day

4.	to do with: It is **to do with** money.	
	a.	making
	b.	for
	c.	about
	d.	our

5.	for instance: **For instance**, it is cheaper.	
	a.	maybe
	b.	for a short time
	c.	In my opinion
	d.	as an example

6.	take over: They will **take over.**	
	a.	be finished
	b.	have control
	c.	come later
	d.	think about it

7.	a range of: They have **a range of** them.	
	a.	many of
	b.	very few of
	c.	two of
	d.	one of

8.	as a result: **As a result** it was done.	
	a.	no person knows if
	b.	after a long time
	c.	before that
	d.	because of that

9.	take place: It will **take place** tonight.	
	a.	be fun
	b.	travel
	c.	arrive
	d.	happen

10.	and so on: There are children **and so on.**	
	a.	around
	b.	over there
	c.	and similar things
	d.	because of it

Third 1000

1.	it takes: He has what **it takes** to learn languages.	
	a.	makes it difficult
	b.	is necessary
	c.	is common
	d.	are bad skills

2.	other than: **Other than** that, it's good.	
	a.	not including
	b.	if you include
	c.	because of
	d.	after

3.	carry on: You can **carry on.**	
	a.	continue
	b.	stop
	c.	lift it
	d.	go faster

4.	all over: It is **all over** the bed.	
	a.	covering
	b.	not on
	c.	on top of
	d.	near

5.	turn out: It **turned out** different.	
	a.	started
	b.	seemed
	c.	became
	d.	did not look

6.	in time: **In time** they bought a house.	
	a.	quickly
	b.	early
	c.	eventually
	d.	instead

7.	feel like: I just did not **feel like** it.	
	a.	love
	b.	want to do
	c.	think about
	d.	try to do

8.	or so: It was a day **or so.**	
	a.	that was not good
	b.	and maybe more
	c.	exactly
	d.	but probably less

9.	shake your head: **shook her head.**	
	a.	said hello
	b.	was very surprised
	c.	hurt herself
	d.	said no

10.	whether or not: I don't know **whether or not** it's expensive.	
	a.	why
	b.	if
	c.	how much
	d.	who says

Fourth 1000

1.	as yet: The have not travelled **as yet.**	
	a.	still
	b.	like now
	c.	unfortunately
	d.	very fast

2.	prove to be: It has **proved to be** important.	
	a.	not been
	b.	become
	c.	no chance to be
	d.	not usually been

3.	in effect: It is **in effect** the reason.	
	a.	possibly
	b.	not
	c.	now
	d.	actually

4.	happen to: She **happened to** call.	
	a.	pretended
	b.	tried hard to
	c.	did not want to
	d.	by chance did

5.	by no means: He is **by no means** tall.	
	a.	very
	b.	not at all
	c.	more or less
	d.	a little

6.	take advantage: You must **take advantage.**	
	a.	go slowly
	b.	use the opportunity
	c.	pay attention
	d.	relax

7.	in the light of: It was accepted **in the light of** the money.	
	a.	despite
	b.	because of
	c.	in addition to
	d.	instead of

8.	give rise to: That **gave rise to** questions.	
	a.	created an increase in
	b.	stopped the
	c.	caused
	d.	was the best of the

9.	no matter: I goes **no matter.**	
	a.	anyway
	b.	often
	c.	only sometimes
	d.	because he must

10.	come across: They **came across** a hotel.	
	a.	stayed in
	b.	drove past
	c.	were near
	d.	found

Fifth 1000

1.	take for granted: She **took it for granted.**	
	a.	kept it
	b.	did not give it importance
	c.	wanted it a lot
	d.	thought about it carefully

2.	as of: It changes **as of** today.	
	a.	starting
	b.	sometime
	c.	perhaps
	d.	because of

3.	would appear: It **would appear** it's true.	
	a.	cannot be that
	b.	is certain that
	c.	seems that
	d.	is assumed that

4.	to blame: We are not **to blame.**	
	a.	in total agreement
	b.	interested
	c.	accusing anyone
	d.	the cause of the problem

5.	stand for: It **stands for** wealth.	
	a.	represents
	b.	is the cause of
	c.	the opposite of
	d.	is related to

6.	by far: She is **by far** the most intelligent.	
	a.	trying to be
	b.	not at all
	c.	really
	d.	sometimes

7.	keep on: We had to **keep on.**	
	a.	not give anything away
	b.	not tell anyone
	c.	rest
	d.	continue

8.	over time: **Over time** it was cheaper.	
	a.	long ago
	b.	eventually
	c.	when it was too late
	d.	at the perfect moment

9.	come up to: He just **came up to** me.	
	a.	approached
	b.	rejected
	c.	did not like
	d.	copied

10.	straight away: They did it **straight away.**	
	a.	immediately
	b.	the correct way
	c.	slowly
	d.	because they wanted to

(THIS PAGE HAS BEEN LEFT INTENTIONALLY BLANK)

Appendix 11 – Research information sheets

The University of Nottingham

Special vocabulary test of multi-word expressions at Regent Oxford

Do you know the difference between the word ***might*** and the expression ***might as well***? What about ***good*** and ***for good***? All are very common in English, but they have completely different meanings. In fact, there are thousands of such expressions in English, and not knowing them can be a problem for comprehension.

Starting on December 15, a special test will be offered for free to students who want to know how good their knowledge of these expressions is. The test is being used for research purposes at the University of Nottingham, but students at Regent who take this test will be able to receive individual feedback on what their strengths and weaknesses are in this important area, and how they can improve.

Also, a free movie voucher for the Odeon Cinema in Oxford will be given to each student who takes the test and also agrees to a short interview.

In total the test will take no more than 1.5 hours. Should you have any further questions, you can ask your teacher or ask the researcher directly at aexrm3@nottingham.ac.uk.

Please sign up below. (Limit 10 per class.)

The University of Nottingham

Research Title: Testing L2 Phrasal Vocabulary

Researcher responsible:
Ron Martinez Email: aexrm3@nottingham.ac.uk Tel: 07786 [redacted]

What is the proposed research and what is its motivation?

Vocabulary has traditionally been seen as being made up of individual words (e.g. *might, large, place*). However, it is now known that vocabulary is also composed of multiword expressions (e.g. *might as well, by and large, take place*), which often cannot be easily decoded by only understanding the individual words in the expression, and which also have been shown to negatively affect reading comprehension.

Likewise, vocabulary testing has so far primarily focused on individual words, and so little is known about how L2 English learners' phrasal vocabulary size compares to their single-word vocabulary. The present research aims to address this gap.

What will the procedure be?

The researcher and/or teachers in the school will inform the students that a special vocabulary test is being offered for those who would like to know their proficiency in phrasal expressions. They should also be briefly informed as to the benefits of taking the test (see below). Those interested will be asked to go to a designated room at a time to be determined, and they will take a test that should take no more than an hour and a half.

In addition, students will also be offered the opportunity to receive their scores and feedback right away if they stay for an interview immediately following the test. This interview is meant to determine how closely students' answers on the test reflect their actual knowledge. These students will be offered a special incentive for this extra time (see below).

All participants who wish to do so can choose to provide their email addresses so that the researcher can provide individualized feedback, including their scores and how to interpret them.

What do the students get out of it?

The incentive to take the test is two-fold, including a pedagogical incentive, and a small financial incentive.

On a purely pedagogical level, students who take the test will become more familiar with an important part of vocabulary, and with how extensive their knowledge is of that important part. They will also be offered advice regarding how to improve their vocabulary in general, and phrasal lexicons in particular.

Those participants who choose to give a little extra of their time to respond to a post-test interview will receive a gift voucher for the Odeon cinema in Oxford, to attend the film of their choice.

Appendix 12 – Phrasal VLT and Phrasal VST knowledge discrepancy analysis sheet (sample)

PARTICIPANT: C.L. (VLT first)

PHRASE	VST	VLT	Questionnaire	Observations
lead to	1	1	1	
have to	1	1	1	
a number of	1	1	1	
go on	1	1	1	
a bit	1	1	1	
is likely to	1	1	1	
deal with	1	1	1	
at all	1	1	1	
is to	1	--	--	
a lot	1	1	1	
as soon as	1	0	1	
find out	1	1	1	
so far	1	0	1	
to do with	1	1	1	
for instance	1	1	1	
take over	1	1	1	
a range of	1	1	1	
as a result	1	1	1	
take place	1	1	1	
and so on	1	1	1	Correct on VLT-C, but missed as distractor on VLT-A
it takes	1	0	1	
other than	1	0	1	
carry on	1	1	1	

all over	0	1	1	Candidate first chose 'covering' (correct answer)
turn out	1	0	1	
in time	0	0	0	
feel like	0	0	0	
or so	1	1	1	
shake your head	1	1	1	
whether or not	1	--	--	
as yet	1	1	1	
prove to be	1	0	1	
in effect	1	1	1	
happen to	0	0	0	
by no means	0	0	0	
take advantage	1	1	1	
in the light of	1	0	1	
give rise to	0	1	1	Candidate distracted by 'increase', also indicated by NS
no matter	1	1	1	
come across	1	0	1	
take for granted	1	0	1	
as of	0	0	0	
would appear	1	1	1	
to blame	0	0	0	
stand for	1	1	1	
by far	0	1	1	Distractor 'really' perhaps faulty
keep on	1	1	1	
over time	0	0	0	
come up to	1	1	1	
straight away	1	1	1	
Discrepancies	3 (6%)	9 (18%)		

Appendix 13 – Breakdown by phrase of PVLT x PVST knowledge discrepancies

lead_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

have_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

a_number_of

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

go_on

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

a_bit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

is_likely_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item unknown, PVLT discrepant	1	10.0	10.0	80.0
	Item known, PVST discrepant	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

deal_with

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	5	50.0	50.0	50.0
	Item known, PVST discrepant	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

at_all

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item known, PVLT discrepant	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

a_lot

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

as_soon_as

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	5	50.0	50.0	50.0
	Item known, PVLT discrepant	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

find_out

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	9	90.0	90.0	90.0
	Item known, PVLT discrepant	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

so_far

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	6	60.0	60.0	60.0
	Item known, PVLT discrepant	4	40.0	40.0	100.0
	Total	10	100.0	100.0	

to_do_with

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	5	50.0	50.0	50.0
	Item known, PVLT discrepant	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

for_instance

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

take_over

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	4	40.0	40.0	40.0
	Item unknown, PVLT discrepant	3	30.0	30.0	70.0
	All incorrect and match	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

a_range_of

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

as_a_result

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item known, PVLT discrepant	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

take_place

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

and_so_on

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

it_takes

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	5	50.0	50.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

other_than

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	4	40.0	40.0	70.0
	All incorrect and match	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

carry_on

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

all_over

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	8	80.0	80.0	80.0
	Item known, PVST discrepant	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

turn_out

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	6	60.0	60.0	60.0
	Item known, PVLT discrepant	4	40.0	40.0	100.0
	Total	10	100.0	100.0	

In_time

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	2	20.0	20.0	20.0
	Item known, PVLT discrepant	3	30.0	30.0	50.0
	All incorrect and match	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

feel_like

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	8	80.0	80.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

or_so

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item unknown, PVLT discrepant	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

shake_your_head

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

as_yet

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item known, PVLT discrepant	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

prove_to_be

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	6	60.0	60.0	60.0
	Item known, PVLT discrepant	4	40.0	40.0	100.0
	Total	10	100.0	100.0	

In_effect

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	5	50.0	50.0	50.0
	Item known, PVLT discrepant	3	30.0	30.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

happen_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	1	10.0	10.0	40.0
	Item unknown, PVLT discrepant	1	10.0	10.0	50.0
	All incorrect and match	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

by_no_means

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	6	60.0	60.0	60.0
	Item unknown, PVLT discrepant	2	20.0	20.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

take_advantage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	9	90.0	90.0	90.0
	Item known, PVLT discrepant	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

in_the_light_of

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	2	20.0	20.0	50.0
	All incorrect and match	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

give_rise_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	6	60.0	60.0	60.0
	Item known, PVST discrepant	2	20.0	20.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

no_matter

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	8	80.0	80.0	80.0
	Item known, PVLT discrepant	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

come_across

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	5	50.0	50.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

take_for_granted

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	3	30.0	30.0	30.0
	Item known, PVLT discrepant	2	20.0	20.0	50.0
	Item unknown, PVLT discrepant	1	10.0	10.0	60.0
	All incorrect and match	4	40.0	40.0	100.0
	Total	10	100.0	100.0	

as_of

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Item known, PVLT discrepant	2	20.0	20.0	20.0
	Item unknown, PVLT discrepant	1	10.0	10.0	30.0
	All incorrect and match	7	70.0	70.0	100.0
	Total	10	100.0	100.0	

would_appear

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	8	80.0	80.0	80.0
	Item unknown, PVLT discrepant	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

to_blame

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Item known, PVLT discrepant	3	30.0	30.0	30.0
	Item unknown, PVLT discrepant	2	20.0	20.0	50.0
	All incorrect and match	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

stand_for

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	8	80.0	80.0	80.0
	All incorrect and match	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

by_far

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	2	20.0	20.0	20.0
	Item known, PVLT discrepant	3	30.0	30.0	50.0
	Item known, PVST discrepant	2	20.0	20.0	70.0
	All incorrect and match	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

keep_on

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	10	100.0	100.0	100.0

over_time

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Item known, PVLT discrepant	2	20.0	20.0	20.0
	All incorrect and match	8	80.0	80.0	100.0
	Total	10	100.0	100.0	

come_up_to

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	7	70.0	70.0	70.0
	Item known, PVLT discrepant	1	10.0	10.0	80.0
	Item unknown, PVLT discrepant	1	10.0	10.0	90.0
	All incorrect and match	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

straight_way

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All correct and match	5	50.0	50.0	50.0
	Item unknown, PVST discrepant	2	20.0	20.0	70.0
	All incorrect and match	3	30.0	30.0	100.0
	Total	10	100.0	100.0	

Appendix 14 – PVST with extra items for Innsbruck field test

Name: ________________ Group: ______________ Sex: ______________ Date: ____________

First language: ______________ Age: ________ email (optional): ________________________

A vocabulary size test of multiword expressions

This is a vocabulary test. Choose the right phrase to go with each meaning and tick the correct answer. Here is an example.

1.	have got: He **has got** it.	
	a.	bought
	b.	has
	c.	received
	d.	wants

1.	have got: He **has got** it.	
	a.	bought
✓	b.	has
	c.	received
	d.	wants

If you have no idea about the meaning of a phrase, do not guess. But if you think you know the meaning, then try to find the answer.

First 1000

1.	lead to: No one knows what it will **lead to.**	
	a.	want
	b.	have inside
	c.	cause in the future
	d.	find

2.	have to: You **have to** go.	
	a.	must
	b.	want to
	c.	can
	d.	will

3.	a number of: **A number of** people came.	
	a.	almost no
	b.	several
	c.	no
	d.	two

4.	go on: It will **go on.**	
	a.	sleep
	b.	repeat
	c.	be fast
	d.	continue

5.	a bit: It's **a bit** different.	
	a.	very
	b.	a little
	c.	sometimes
	d.	always

6.	is likely to: He **is likely to** go.	
	a.	likes to
	b.	can
	c.	wants to
	d.	probably will

7.	deal with: I can **deal with** it.	
	a.	fix
	b.	try
	c.	find
	d.	see

8.	at all: I don't like it **at all.**	
	a.	all the time
	b.	in any way
	c.	at first
	d.	sometimes

9.	is to: He **is to** speak this afternoon.	
	a.	will
	b.	can
	c.	wants to
	d.	may

10.	a lot: I go there **a lot.**	
	a.	always
	b.	often
	c.	never
	d.	sometimes

11.	I mean: Two, **I mean**, three.	
	a.	I am guessing
	b.	maybe
	c.	then later
	d.	I correct myself

12.	at least: **At least** it is warm.	
	a.	other things may be bad, but
	b.	many days have passed and now
	c.	I cannot believe that
	d.	the least important thing is

13.	so that: He sat **so that** they could do it.	
	a.	to make it possible that
	b.	because
	c.	very slowly and then
	d.	before

14.	used to: I **used to** go.	
	a.	want to
	b.	did before
	c.	usually
	d.	always

15.	rather than: I'll cook **rather than** eat.	
	a.	or maybe
	b.	but I prefer to
	c.	before I
	d.	and not

Second 1000

1.	as soon as: I'll go **as soon as** I can.	
	a.	from the moment
	b.	only if
	c.	after
	d.	before

2.	find out: How did you **find out**?	
	a.	arrive there
	b.	like it
	c.	get the information
	d.	leave

3.	so: It's good **so far**.	
	a.	until now
	b.	but not really
	c.	sometimes
	d.	from a distance

4.	to do with: It is **to do with** money.	
	a.	making
	b.	for
	c.	about
	d.	our

5.	for instance: **For instance**, it is cheaper.	
	a.	maybe
	b.	for a short time
	c.	In my opinion
	d.	as an example

6.	take over: They will **take over**.	
	a.	be finished
	b.	have control
	c.	come later
	d.	think about it

7.	a range of: They have **a range of** colours.	
	a.	many different
	b.	very few of
	c.	all of the
	d.	one or two

8.	as a result: **As a result** it was done.	
	a.	no person knows if
	b.	after a long time
	c.	before that
	d.	because of that

9.	take place: It will **take place** tonight.	
	a.	be fun
	b.	travel
	c.	arrive
	d.	happen

10.	and so on: There are children **and so on**.	
	a.	around
	b.	over there
	c.	and similar things
	d.	only

11.	carry out: It was **carried out** yesterday.	
	a.	started
	b.	found
	c.	read
	d.	done

12.	each other: They have **each other**.	
	a.	one yes, one no
	b.	themselves
	c.	all of them
	d.	some of them

13.	in particular: I want that **in particular**.	
	a.	especially
	b.	in private
	c.	because it is different
	d.	maybe

14.	be expected to: We **are expected to** do it.	
	a.	are waiting
	b.	hoping to
	c.	must
	d.	are able to

15.	be about to: I **am about to** read the newspaper.	
	a.	cannot wait to
	b.	am soon going to
	c.	really like to
	d.	am trying to

Third 1000

1.	it takes: He has what **it takes** to learn languages.	
	a.	makes it difficult
	b.	is necessary
	c.	is common
	d.	are bad skills

2.	other than: **Other than** that, it's good.	
	a.	not including
	b.	if you include
	c.	because of
	d.	after

3.	carry on: You can **carry on**.	
	a.	continue
	b.	stop
	c.	lift it
	d.	go faster

4.	all over: It is **all over** the bed.	
	a.	covering
	b.	inside
	c.	on top of
	d.	beside

5.	turn out: It **turned out** different.	
	a.	started
	b.	seemed
	c.	became
	d.	did not look

6.	in time: **In time** they bought a house.	
	a.	quickly
	b.	earlier
	c.	eventually
	d.	recently

7.	feel like: I just did not **feel like it.**	
	a.	love
	b.	want to do
	c.	think about
	d.	try to do

8.	or so: It will take a week **or so.**	
	a.	exactly
	b.	and maybe much more
	c.	and maybe less
	d.	maybe more, maybe less

9.	shake your head: She **shook her head.**	
	a.	said hello
	b.	was very surprised
	c.	hurt herself
	d.	said no

10.	whether or not: We don't know **whether or not** it's expensive.	
	a.	why
	b.	if
	c.	how much
	d.	who says

11.	get to: She **got to** the car.	
	a.	arrived at
	b.	drove
	c.	received
	d.	entered

12.	at once: I did it **at once.**	
	a.	one time
	b.	many times
	c.	early
	d.	immediately

13.	give up: I **give up**.	
	a.	try very hard
	b.	am starting
	c.	will now stop
	d.	exercise

14.	in touch: Keep **in touch**.	
	a.	feeling it
	b.	communicating
	c.	pushing it
	d.	thinking

15.	get rid of: They **got rid of** it.	
	a.	decided to not have
	b.	received
	c.	became bored with
	d.	chose

Fourth 1000

1.	as yet: They have not travelled **as yet**.	
	a.	not now but maybe later
	b.	because they don't want to
	c.	because they have no time
	d.	not now and not ever

2.	prove to be: It has **proved to be** important.	
	a.	will possibly become
	b.	shown itself to be
	c.	continued to be
	d.	never been

3.	in effect: It is **in effect** the reason.	
	a.	possibly
	b.	not
	c.	now
	d.	actually

4.	happen to: She **happened to** call.	
	a.	pretended
	b.	tried hard to
	c.	did not want to
	d.	by chance did

5.	by no means: He is **by no means** rich.	
	a.	very
	b.	not at all
	c.	more or less
	d.	considered

6.	take advantage: You must **take advantage**.	
	a.	go slowly
	b.	use the opportunity
	c.	pay attention
	d.	relax

7.	in the light of: It was accepted **in the light of** the money.	
	a.	despite
	b.	because of
	c.	in addition to
	d.	instead of

8.	give rise to: That **gave rise to** questions.	
	a.	increased the number of
	b.	stopped the
	c.	caused
	d.	was the best of the

9.	no matter: He will do it **no matter how**.	
	a.	any way possible
	b.	the way he planned
	c.	when he can
	d.	as needed

10.	come across: They **came across** a hotel.	
	a.	stayed in
	b.	opened
	c.	were near
	d.	found

11.	even so: **Even so** it's better.	
	a.	despite that
	b.	that way
	c.	it is the same and
	d.	maybe

12.	run out: I think we **ran out** of it.	
	a.	had no more
	b.	were bored
	c.	thought
	d.	moved outside

13.	might as well: You **might as well** go.	
	a.	possibly will
	b.	ought to
	c.	have to
	d.	can

14.	next door: It's just **next door.**	
	a.	coming soon
	b.	common
	c.	perfect
	d.	very close

15.	on the one hand: **On the one hand** he's fine.	
	a.	before everything happens
	b.	considering one aspect
	c.	usually
	d.	when alone

Fifth 1000

1.	take for granted: She **took it for granted.**	
	a.	kept it
	b.	did not give it importance
	c.	wanted it a lot
	d.	thought about it carefully

2.	as of: It changes **as of** today.	
	a.	starting
	b.	sometime
	c.	perhaps
	d.	because of

3.	would appear: It **would appear** it's true.	
	a.	cannot be that
	b.	is certain that
	c.	seems that
	d.	is assumed that

4.	to blame: We are not **to blame.**	
	a.	in total agreement
	b.	interested
	c.	accusing anyone
	d.	the cause of the problem

5.	stand for: It **stands for** wealth.	
	a.	represents
	b.	is the cause of
	c.	the opposite of
	d.	is related to

6.	by far: She is **by far** the most intelligent.	
	a.	trying to be
	b.	not at all
	c.	really
	d.	sometimes

7.	keep on: We had to **keep on.**	
	a.	not give anything away
	b.	not tell anyone
	c.	rest
	d.	continue

8.	over time: **Over time** it was cheaper.	
	a.	long ago
	b.	eventually
	c.	when it was too late
	d.	at the perfect moment

9.	come up to: He just **came up to** me.	
	a.	approached
	b.	rejected
	c.	did not like
	d.	copied

10.	straight away: They did it **straight away.**	
	a.	immediately
	b.	the correct way
	c.	slowly
	d.	because they wanted to

11.	shut up: He won't **shut up**.	
	a.	be shy
	b.	stop trying
	c.	lock the door
	d.	stop talking

12.	a handful of: **A handful of** us went.	
	a.	a few
	b.	none
	c.	most
	d.	all

13.	can tell: You **can tell.**	
	a.	may speak
	b.	are smart
	c.	can see
	d.	might

14.	under way: It is **under way**.	
	a.	coming
	b.	happening now
	c.	stopped
	d.	an obstacle

15.	turn down: She **turned down** the money.	
	a.	hid
	b.	lost
	c.	made
	d.	refused

(THIS PAGE HAS BEEN LEFT INTENTIONALLY BLANK)

Appendix 15 – Full item analysis of Phrasal Vocabulary Size Test (Innsbruck field test)

Frequency Band K1

K1, Item A1 (item-total correlation .454)

1.	lead to: No one knows what it will **lead to.**		Facility	Upper	Lower	D
	a.	want	.01	.00	.04	-.04
	b.	have inside	. 02	.00	.06	-.06
	c.	cause in the future	.94	1.00	.86	.14
	d.	find	.01	.00	.04	-.04
		No attempt		0 (%)	9 (4.5%)	

K1, Item A2 (item-total correlation .347)

2.	have to: You **have to** go.		Facility	Upper	Lower	D
	a.	must	.97	.99	.94	.05
	b.	want to	.00	.00	.00	.00
	c.	can	.01	.01	.04	-.03
	d.	will	.01	.00	.02	-.02
		No attempt		0 (%)	2 (1%)	

K1, Item A3 (item-total correlation .329)

3.	a number of: **A number of** people came.		Facility	Upper	Lower	D
	a.	almost no	.01	.01	.01	.00
	b.	several	.98	.99	.96	.03
	c.	no	.00	.00	.00	.00
	d.	two	.01	.01	.03	-.02
		No attempt		0 (%)	2 (1%)	

K1, Item A4 (item-total correlation .453)

4.	go on: It will **go on.**		Facility	Upper	Lower	D
	a.	sleep	.00	.00	.01	-.01
	b.	repeat	. 01	.01	.02	-.01
	c.	be fast	.01	.00	.04	-.04
	d.	continue	.97	.99	.93	.06
		No attempt		0 (%)	2 (1%)	

K1, Item A5 (item-total correlation .359)

5.	a bit: It's **a bit** different.		Facility	Upper	Lower	D
	a.	very	.01	.00	.01	-.01
	b.	a little	. 97	1.00	.94	.06
	c.	sometimes	.02	.00	.04	-.04
	d.	always	.00	.00	.01	-.01
		No attempt		1 (0.5%)	1 (0.5%)	

K1, Item A6 (item-total correlation .444)

6.	is likely to: He **is likely to** go.		Facility	Upper	Lower	D
	a.	likes to	.02	.01	.03	-.02
	b.	can	. 02	.01	.05	-.04
	c.	wants to	.27	.02	.49	-.47
	d.	probably will	.68	.97	.42	.55
		No attempt		0 (%)	5 (2.5%)	

K1, Item B7 (item-total correlation .395)

7.	deal with: I can **deal with** it.		Facility	Upper	Lower	D
	a.	fix	.66	.89	.43	.46
	b.	try	.23	.07	.36	-.29
	c.	find	.02	.02	.06	-.04
	d.	see	.03	.01	.05	-.04
		No attempt		4 (2.1%)	21 (10.8%)	

K1, Item B8 (item-total correlation (.407)

8.	at all: I don't like it **at all.**		Facility	Upper	Lower	D
	a.	all the time	.07	.01	.13	-.12
	b.	in any way	. 91	.99	.82	.07
	c.	at first	.01	.00	.02	-.02
	d.	sometimes	.00	.00	.02	-.02
		No attempt		0 (0%)	4 (2.1%)	

K1, Item B9 (item-total correlation .350)

9.	is to: He **is to** speak this afternoon.		Facility	Upper	Lower	D
	a.	will	.61	.85	.44	.41
	b.	can	. 11	.05	.16	-.11
	c.	wants to	.16	.05	.32	-.27
	d.	may	.06	.05	.07	-.02
		No attempt		2 (1%)	16 (8.2%)	

K1, Item B10 (item-total correlation .427)

10.	a lot: I go there **a lot**.		Facility	Upper	Lower	D
	a.	always	.01	.00	.01	-.01
	b.	often	. 97	1.00	.94	.06
	c.	never	.00	.00	.01	-.01
	d.	sometimes	.01	.00	.03	-.03
		No attempt		0 (0%)	3 (1.5%)	

K1, Item B11 (item-total correlation .398)

11.	I mean: Two, **I mean**, three.		Facility	Upper	Lower	D
	a.	I am guessing	.04	.00	.12	-.12
	b.	maybe	.04	.02	.07	-.05
	c.	then later	.00	.00	.00	.00
	d.	I correct myself	.90	.98	.78	.20
		No attempt		0 (0%)	6 (3.1%)	

K1, Item B12 (item-total correlation .538)

12.	at least: **At least** it is warm.		Facility	Upper	Lower	D
	a.	other things may be bad, but	.72	.99	.42	.57
	b.	many days have passed and now	. 13	.01	.29	-.28
	c.	I cannot believe that	.02	.00	.05	-.05
	d.	the least important thing is	.09	.00	.24	-.24
		No attempt		2 (1%)	15 (7.7%)	

K1, Item C1 (item-total correlation .467)

1.	lead to: No one knows what it will **lead to.**		Facility	Upper	Lower	D
	a.	want	.02	.00	.07	-.07
	b.	have inside	. 02	.00	.06	-.06
	c.	cause in the future	.92	1.00	.79	.21
	d.	find	.01	.00	.04	-.04
		No attempt		0 (0%)	10 (5.1%)	

K1, Item C2 (item-total correlation .325)

2.	have to: You **have to** go.		Facility	Upper	Lower	D
	a.	must	.96	.99	.97	.02
	b.	want to	.01	.00	.02	-.02
	c.	can	.01	.01	.01	.00
	d.	will	.01	.00	.00	.00
		No attempt		0 (0%)	4 (2.1%)	

K1 ,Item C3 (item-total correlation .477)

3.	a number of: **A number of** people came.		Facility	Upper	Lower	D
	a.	almost no	.01	.00	.03	-.03
	b.	several	. 98	1.00	.95	.05
	c.	no	.00	.00	.02	-.02
	d.	two	.00	.00	.00	.00
		No attempt		0 (0%)	3 (1.5%)	

K1, Item C13 (item-total correlation .425)

13.	so that: He sat **so that** they could do it.		Facility	Upper	Lower	D
	a.	to make it possible that	.93	.99	.86	.13
	b.	because	. 02	.01	.03	-.02
	c.	very slowly and then	.01	.00	.02	-.02
	d.	before	.01	.00	.04	-.04
		No attempt		0 (0%)	10 (5.1%)	

K1, Item C14 (item-total correlation .389)

14.	used to: I **used to** go.		Facility	Upper	Lower	D
	a.	want to	.12	.01	.29	-.28
	b.	did before	.26	.55	.07	.48
	c.	usually	.56	.40	.54	-.14
	d.	always	.07	.05	.09	-.04
		No attempt		1 (0.5%)	14 (7.2%)	

K1, Item C15 (item-total correlation .240)

15.	rather than: I'll cook **rather than** eat.		Facility	Upper	Lower	D
	a.	or maybe	.01	.01	.03	-.02
	b.	but I prefer to	.55	.41	.63	-.22
	c.	before I	.22	.20	.23	-.03
	d.	and not	.22	.38	.12	.26
		No attempt		5 (2.6%)	14 (7.2%)	

Frequency Band K2

K2, Item A1 (item-total correlation .380)

1.	as soon as: I'll go **as soon as** I can.		Facility	Upper	Lower	D
	a.	from the moment	.74	.93	.56	.37
	b.	only if	.14	.00	.25	-.25
	c.	after	.11	.04	.16	-.12
	d.	before	.01	.03	.03	.00
		No attempt		0 (0%)	17(8%)	

K2, Item A2 (item-total correlation .373)

2.	find out: How did you **find out**?		Facility	Upper	Lower	D
	a.	arrive there	.00	.00	.01	-.01
	b.	like it	.01	.00	.00	.00
	c.	get the information	.99	1.00	.99	.01
	d.	leave	.00	.00	.00	.00
		No attempt		1 (0.5%)	2 (1%)	

K2, Item A3 (item-total correlation .488)

3.	so: It's good **so far**.		Facility	Upper	Lower	D
	a.	until now	.90	1.00	.75	.25
	b.	but not really	.04	.00	.08	-.08
	c.	sometimes	.01	.00	.02	-.02
	d.	from a distance	.05	.00	.15	-.15
		No attempt		0 (0%)	12(5%)	

K2, Item A4 (item-total correlation .415)

4.	to do with: It is **to do with** money.		Facility	Upper	Lower	D
	a.	making	.08	.01	.19	-.18
	b.	for	.03	.02	.05	-.03
	c.	about	.89	.97	.75	.22
	d.	our	.00	.00	.01	-.01
		No attempt		2 (1%)	14 (6.5%)	

K2, Item A5 (item-total correlation .403)

5.	for instance: **For instance**, it is cheaper.		Facility	Upper	Lower	D
	a.	maybe	.05	.03	.09	-.06
	b.	for a short time	.09	.00	.14	-.14
	c.	In my opinion	.04	.00	.11	-.11
	d.	as an example	.82	.97	.66	.31
		No attempt		0 (0%)	13 (6%)	

K2, Item A6 (item-total correlation .537)

6.	take over: They will **take over**.		Facility	Upper	Lower	D
	a.	be finished	.10	.01	.23	-.23
	b.	have control	.83	.98	.63	.35
	c.	come later	.02	.01	.04	-.03
	d.	think about it	.05	.01	.10	-.09
		No attempt		1 (0.5%)	18(9%)	

K2, Item B7 (item-total correlation .428)

7.	a range of: They have **a range of** colours.		Facility	Upper	Lower	D
	a.	many different	.97	1.00	.90	.10
	b.	very few of	.03	.00	.08	-.02
	c.	all of the	.00	.00	.01	-.01
	d.	one or two	.00	.00	.01	-.01
		No attempt		0 (0%)	8 (4%)	

K2, B8 (item-total correlation .445)

8.	as a result: **As a result** it was done.		Facility	Upper	Lower	D
	a.	no person knows if	.01	.00	.01	-.01
	b.	after a long time	.18	.03	.33	-.30
	c.	before that	.01	.00	.03	-.03
	d.	because of that	.78	.97	.63	.34
		No attempt		0 (0%)	17 (8%)	

K2, Item B9 (item-total correlation .529)

9.	take place: It will **take place** tonight.		Facility	Upper	Lower	D
	a.	be fun	.01	.00	.02	-.02
	b.	travel	.02	.00	.03	-.03
	c.	arrive	.03	.01	.10	-.09
	d.	happen	.94	.99	.85	.14
		No attempt		0 (0%)	7(3%)	

K2, B10 (item-total correlation .472)

10.	and so on: There are children **and so on**.		Facility	Upper	Lower	D
	a.	around	.02	.00	.05	-.05
	b.	over there	.02	.00	.06	-.06
	c.	and similar things	.94	1.00	.85	.15
	d.	only	.01	.00	.04	-.04
		No attempt		1 (0.5%)	9 (4.5%)	

K2, Item B11 (item-total correlation .403)

11.	carry out: It was **carried out** yesterday.		Facility	Upper	Lower	D
	a.	started	.14	.06	.24	-.18
	b.	found	.11	.01	.24	.23
	c.	read	.01	.00	.01	-.01
	d.	done	.74	.93	.50	.43
		No attempt		2 (1%)	23 (12%)	

K2, B12 (item-total correlation .464)

12.	each other: They have **each other**.		Facility	Upper	Lower	D
	a.	one yes, one no	.01	.00	.01	-.01
	b.	themselves	.91	.99	.80	.19
	c.	all of them	.05	.01	.13	-.12
	d.	some of them	.02	.00	.07	-.07
		No attempt		0 (0%)	11(5%)	

K2, Item C1 (item-total correlation .378)

1.	as soon as: I'll go **as soon as** I can.		Facility	Upper	Lower	D
	a.	from the moment	.84	.92	.73	.19
	b.	only if	.10	.03	.15	-.12
	c.	after	.05	.04	.08	-.04
	d.	before	.01	.01	.04	-.03
		No attempt		1 (0.5%)	26 (10%)	

K2, Item C2 (item-total correlation .522)

2.	find out: How did you **find out**?		Facility	Upper	Lower	D
	a.	arrive there	.00	.00	.01	-.01
	b.	like it	.01	.00	.02	-.02
	c.	get the information	.98	1.00	.95	.05
	d.	leave	.01	.00	.02	-.02
		No attempt		0 (0%)	0 (0%)	

K2, Item C3 (item-total correlation .561)

3.	so: It's good **so far**.		Facility	Upper	Lower	D
	a.	until now	.91	.99	.72	.27
	b.	but not really	.03	.01	.09	-.08
	c.	sometimes	.01	.00	.02	-.02
	d.	from a distance	.05	.00	.16	-.16
		No attempt		1 (0.5%)	23 (8%)	

K2, Item C13 (item-total correlation .512)

13.	in particular: I want that **in particular**.		Facility	Upper	Lower	D
	a.	especially	.82	.99	.70	.29
	b.	in private	.04	.00	.10	-.10
	c.	because it is different	.04	.01	.11	-.10
	d.	maybe	.03	.01	.04	-.03
		No attempt		0 (0%)	20 (12%)	

K2, C14 (item-total correlation .441)

14.	be expected to: We **are expected to** do it.		Facility	Upper	Lower	D
	a.	are waiting	.18	.04	.27	-.13
	b.	hoping to	.16	.05	.21	-.16
	c.	must	.56	.88	.31	.57
	d.	are able to	.10	.03	.21	-.18
		No attempt		0 (0%)	20 (12%)	

K2, Item C15 (item-total correlation .526)

15.	be about to: I **am about to** read the newspaper.		Facility	Upper	Lower	D
	a.	cannot wait to	.04	.01	.06	-.05
	b.	am soon going to	.73	.97	.47	.50
	c.	really like to	.07	.00	.17	-.17
	d.	am trying to	.17	.02	.30	-.28
		No attempt		0 (0%)	40 (22%)	

Frequency Band K3

K3, Item A1 (item-total correlation .525)

1.	it takes: He has what **it takes** to learn languages.		Facility	Upper	Lower	D
	a.	makes it difficult	.01	.00	.04	-.04
	b.	is necessary	.95	1.00	.85	.15
	c.	is common	.04	.00	.11	-.11
	d.	are bad skills	.00	.00	.00	-.00
		No attempt		0 (0%)	19 (11%)	

K3, Item A2 (item-total correlation .522)

2.	other than: **Other than** that, it's good.		Facility	Upper	Lower	D
	a.	not including	.85	.99	.58	.41
	b.	if you include	.06	.00	.13	-.13
	c.	because of	.07	.00	.21	-.21
	d.	after	.03	.01	.08	-.07
		No attempt		2 (1%)	18 (10%)	

K3, Item A3 (item-total correlation .547)

3.	carry on: You can **carry on**.		Facility	Upper	Lower	D
	a.	continue	.92	1.00	.76	.24
	b.	stop	.00	.00	.01	-.01
	c.	lift it	.01	.00	.04	-.04
	d.	go faster	.07	.00	.19	-.19
		No attempt		0 (0%)	4 (2%)	

K3, A4 (item-total correlation, .564)

4.	all over: It is **all over** the bed.		Facility	Upper	Lower	D
	a.	covering	.78	.97	.47	.50
	b.	inside	.03	.01	.07	-.06
	c.	on top of	.18	.03	.42	-.39
	d.	beside	.02	.00	.04	-.04
		No attempt		0 (0%)	9 (5%)	

K3, Item A5 (item-total correlation .506)

5.	turn out: It **turned out** different.		Facility	Upper	Lower	D
	a.	started	.02	.00	.06	-.06
	b.	seemed	.12	.02	.24	-.22
	c.	became	.84	.98	.66	.32
	d.	did not look	.01	.00	.04	-.04
		No attempt		0 (0%)	9 (5%)	

K3, Item A6 (item-total correlation .247)

6.	in time: **In time** they bought a house.		Facility	Upper	Lower	D
	a.	quickly	.14	.20	.13	.17
	b.	earlier	.11	.06	.18	-.12
	c.	eventually	.25	.41	.18	.23
	d.	recently	.50	.33	.51	-.18
		No attempt		0 (0%)	30 (16%)	

K3, Item B7 (item-total correlation .512)

7.	feel like: I just did not **feel like** it.		Facility	Upper	Lower	D
	a.	love	.05	.01	.11	-.10
	b.	want to do	.85	.98	.61	.37
	c.	think about	.07	.01	.19	-.18
	d.	try to do	.04	.00	.09	-.09
		No attempt		0 (0%)	15 (8%)	

K3, B8 (item-total correlation .549)

8.	or so: It will take a week **or so**.		Facility	Upper	Lower	D
	a.	exactly	.01	.00	.03	-.03
	b.	and maybe much more	.02	.00	.06	-.06
	c.	and maybe less	.01	.00	.02	-.02
	d.	maybe more, maybe less	.96	1.00	.83	.17
		No attempt		0 (0%)	0 (0%)	

K3, B9 (item-total correlation .560)

9.	shake your head: She **shook her head**.		Facility	Upper	Lower	D
	a.	said hello	.05	.01	.13	-.12
	b.	was very surprised	.04	.00	.07	-.07
	c.	hurt herself	.01	.00	.02	-.02
	d.	said no	.90	.99	.77	.22
		No attempt		0 (0%)	14 (8%)	

K3, B10 (item-total correlation .586)

10.	whether or not: We don't know **whether or not** it's expensive.		Facility	Upper	Lower	D
	a.	why	.04	.00	.12	-.12
	b.	if	.90	1.00	.75	.25
	c.	how much	.05	.00	.10	-.10
	d.	who says	.01	.00	.03	-.03
		No attempt		0 (0%)	25 (14%)	

K3, Item B11 (item-total correlation .544)

11.	get to: She **got to** the car.		Facility	Upper	Lower	D
	a.	arrived at	.82	.96	.62	.34
	b.	drove	.01	.00	.03	-.03
	c.	received	.08	.03	.13	-.10
	d.	entered	.09	.02	.21	-.19
		No attempt		0 (0%)	22 (12%)	

K3, Item B12 (item-total correlation .503)

12.	at once: I did it **at once**.		Facility	Upper	Lower	D
	a.	one time	.47	.16	.78	-.62
	b.	many times	.00	.00	.00	.00
	c.	early	.02	.00	.06	-.06
	d.	immediately	.43	.81	.16	.65
		No attempt		4 (2%)	29 (16%)	

K3, C1 (item-total correlation .570)

1.	it takes: He has what **it takes** to learn languages.		Facility	Upper	Lower	D
	a.	makes it difficult	.02	.00	.07	-.07
	b.	is necessary	.93	.99	.80	.19
	c.	is common	.05	.01	.12	-.11
	d.	are bad skills	.00	.00	.01	-.01
		No attempt		0 (0%)	0 (0%)	

K3, C2 (item-total correlation .563)

2.	other than: **Other than** that, it's good.		Facility	Upper	Lower	D
	a.	not including	.84	.97	.57	.56
	b.	if you include	.06	.01	.14	-.09
	c.	because of	.07	.01	.20	-.13
	d.	after	.03	.01	.02	-.06
		No attempt		2 (1%)	22 (33%)	

K3, C3 (item-total correlation .658)

3.	carry on: You can **carry on**.		Facility	Upper	Lower	D
	a.	continue	.88	1.00	.82	.18
	b.	stop	.01	.00	.04	-.04
	c.	lift it	.00	.00	.02	-.02
	d.	go faster	.04	.00	.12	-.12
		No attempt		0 (0%)	8 (5%)	

K3, C13 (item-total correlation .633)

13.	give up: I **give up**.		Facility	Upper	Lower	D
	a.	try very hard	.03	.00	.06	-.06
	b.	am starting	.01	.00	.03	-.03
	c.	will now stop	.87	1.00	.65	.35
	d.	exercise	.02	.00	.04	-.04
		No attempt		1 (0.5%)	13 (8%)	

K3, C14 (item-total correlation .598)

14.	in touch: Keep **in touch**.		Facility	Upper	Lower	D
	a.	feeling it	.06	.00	.13	-.13
	b.	communicating	.78	1.00	.47	.53
	c.	pushing it	.04	.00	.07	-.07
	d.	thinking	.04	.00	.05	-.05
		No attempt		0 (0%)	21 (13%)	

K3, C15 (item-total correlation .547)

15.	get rid of: They **got rid of** it.		Facility	Upper	Lower	D
	a.	decided to not have	.58	.91	.22	.69
	b.	received	.06	.00	.12	-.12
	c.	became bored with	.25	.09	.32	-.23
	d.	chose	.00	.00	.03	-.03
		No attempt		0 (0%)	28 (17%)	

Frequency Band K4

K4, Item A1 (item-total correlation .456)

1.	as yet: They have not travelled **as yet.**		Facility	Upper	Lower	D
	a.	not now but maybe later	.90	.99	.79	.20
	b.	because they don't want to	.01	.00	.03	-.03
	c.	because they have no time	.04	.00	.08	-.08
	d.	not now and not ever	.05	.01	.10	-.09
		No attempt		4 (2%)	2 (1%)	

K4, Item A2 (item-total correlation .622)

2.	prove to be: It has **proved to be** important.		Facility	Upper	Lower	D
	a.	will possibly become	.06	.00	.16	-.16
	b.	shown itself to be	.89	1.00	.67	.33
	c.	continued to be	.04	.00	.15	-.15
	d.	never been	.01	.00	.03	-.03
		No attempt		0 (0%)	0 (0%)	

K4, Item A3 (item-total correlation .520)

3.	in effect: It is **in effect** the reason.		Facility	Upper	Lower	D
	a.	possibly	.31	.04	.31	.27
	b.	not	.03	.00	.03	-.03
	c.	now	.13	.02	.13	-.11
	d.	actually	.53	.94	.53	.41
		No attempt		3 (1.5%)	8 (5%)	

K4, A4 (item-total correlation .596)

4.	happen to: She **happened to** call.		Facility	Upper	Lower	D
	a.	pretended	.21	.04	.53	-.49
	b.	tried hard to	.09	.00	.23	-.23
	c.	did not want to	.06	.04	.12	-.08
	d.	by chance did	.47	.92	.11	.81
		No attempt		1 (0.5%)	34 (21%)	

K4, Item A5 (item-total correlation .410)

5.	by no means: He is **by no means** rich.		Facility	Upper	Lower	D
	a.	very	.12	.06	.13	-.07
	b.	not at all	.73	.92	.62	.30
	c.	more or less	.11	.02	.21	-.19
	d.	considered	.05	.00	.04	-.04
		No attempt		2 (1%)	23 (14%)	

K4, A6 (item-total correlation .612)

			Facility	Upper	Lower	D
6.	take advantage: You must **take advantage**.		Facility	Upper	Lower	D
	a.	go slowly	.01	.00	.05	-.05
	b.	use the opportunity	.84	.99	.55	.44
	c.	pay attention	.14	.01	.37	-.36
	d.	relax	.01	.00	.04	-.04
		No attempt		1 (0.5%)	17 (10%)	

K4, Item B7(item-total correlation .497)

7.	in the light of: It was accepted **in the light of** the money.		Facility	Upper	Lower	D
	a.	despite	.03	.02	.07	-.05
	b.	because of	.75	.94	.49	.46
	c.	in addition to	.15	.02	.31	-.29
	d.	instead of	.07	.02	.14	-.12
		No attempt		8 (4%)	0 (0%)	

K4, Item B8 (item-total correlation .380)

8.	give rise to: That **gave rise to** questions.		Facility	Upper	Lower	D
	a.	increased the number of	.52	.35	.60	-.35
	b.	stopped the	.03	.00	.11	-.11
	c.	caused	.42	.64	.23	.41
	d.	was the best of the	.02	.01	.06	-.05
		No attempt		2 (1%)	16 (11%)	

K4, B9 (item-total correlation .650)

9.	no matter: He will do it **no matter how**.		Facility	Upper	Lower	D
	a.	any way possible	.94	.98	.82	.16
	b.	the way he planned	.02	.00	.05	-.05
	c.	when he can	.01	.00	.06	-.06
	d.	as needed	.03	.02	.07	-.07
		No attempt		0 (0%)	0 (0%)	

K4, B10 (item-total correlation .505)

10.	come across: They **came across** a hotel.		Facility	Upper	Lower	D
	a.	stayed in	.06	.01	.15	-.14
	b.	opened	.01	.00	.03	-.03
	c.	were near	.31	.10	.48	-.38
	d.	found	.63	.89	.34	.55
		No attempt		0 (0%)	1 (0.5%)	

K4, Item B11 (item-total correlation .417)

11.	even so: **Even so** it's better.		Facility	Upper	Lower	D
	a.	despite that	.38	.96	.15	.81
	b.	that way	.52	.00	.67	-.67
	c.	it is the same and	.05	.03	.06	-.03
	d.	maybe	.05	.01	.11	-.10
		No attempt		0 (0%)	7 (5%)	

K4, B12 (item-total correlation .618)

12.	run out: I think we **ran out** of it.		Facility	Upper	Lower	D
	a.	had no more	.86	.99	.61	.38
	b.	were bored	.06	.01	.14	-.13
	c.	thought	.01	.00	.03	-.03
	d.	moved outside	.07	.02	.21	-.19
		No attempt		0 (0%)	3 (2.2%)	

K4, C1 (item-total correlation .534)

1.	as yet: They have not travelled **as yet**.		Facility	Upper	Lower	D
	a.	not now but maybe later	.88	.97	.73	.24
	b.	because they don't want to	.00	.00	.02	-.02
	c.	because they have no time	.05	.01	.10	-.09
	d.	not now and not ever	.07	..02	.15	-.13
		No attempt		7 (3.6%)	1 (0.5%)	

K4, C2 (item-total correlation .665)

2.	prove to be: It has **proved to be** important.		Facility	Upper	Lower	D
	a.	will possibly become	.06	.00	.19	-.19
	b.	shown itself to be	.89	1.00	.64	.36
	c.	continued to be	.04	.00	.14	-.14
	d.	never been	.01	.00	.03	-.03
		No attempt		0 (0%)	12 (9%)	

K4, C3 (item-total correlation .572)

3.	in effect: It is **in effect** the reason.		Facility	Upper	Lower	D
	a.	possibly	.21	.02	.33	-.31
	b.	not	.00	.03	.01	.02
	c.	now	.08	.00	.14	-.14
	d.	actually	.71	.95	.52	.43
		No attempt		3 (1.5%)	18 (13.5%)	

K4, Item C13 (item-total correlation .079)

13.	might as well: You **might as well** go.		Facility	Upper	Lower	D
	a.	possibly will	.30	.15	.47	-.32
	b.	ought to	.16	.14	.14	-.00
	c.	have to	.04	.02	.09	-.07
	d.	can	.50	.69	.30	.39
		No attempt		2 (1%)	12 (9%)	

K4, C14 (item-total correlation .552)

14.	next door: It's just **next door**.		Facility	Upper	Lower	D
	a.	coming soon	.11	.05	.24	-.31
	b.	common	.02	.01	.02	-.01
	c.	perfect	.01	.00	.05	-.11
	d.	very close	.86	.94	.70	.24
		No attempt		1 (0.5%)	0 (0%)	

K4, C15 (item-total correlation .662)

15.	on the one hand: **On the one hand** he's fine.		Facility	Upper	Lower	D
	a.	before everything happens	.01	.00	.05	-.05
	b.	considering one aspect	.87	1.00	.89	.11
	c.	usually	.02	.00	.05	-.05
	d.	when alone	.00	.00	.02	-.02
		No attempt		0 (0%)	7 (5.3%)	

Frequency Band K5

K5, Item A1 (item-total correlation .458)

1.	take for granted: She **took** it **for granted**.		Facility	Upper	Lower	D
	a.	kept it	.14	.07	.21	-.14
	b.	did not give it importance	.54	.82	.27	.65
	c.	wanted it a lot	.16	.06	.29	-.23
	d.	thought about it carefully	.16	.05	.23	-.18
		No attempt		7 (3%)	13 (9%)	

K5, Item A2 (item-total correlation .468)

2.	as of: It changes **as of** today.		Facility	Upper	Lower	D
	a.	starting	.57	.85	.27	.58
	b.	sometime	.05	.00	.11	-.11
	c.	perhaps	.12	.05	.28	-.23
	d.	because of	.26	.10	.34	-.24
		No attempt		7 (3%)	25 (18%)	

K5, Item A3 (item-total correlation .394)

3.	would appear: It **would appear** it's true.		Facility	Upper	Lower	D
	a.	cannot be that	.02	.02	.05	-.03
	b.	is certain that	.09	.02	.17	-.15
	c.	seems that	.69	.80	.58	.22
	d.	is assumed that	.20	.16	.21	-.05
		No attempt		1 (0.5%)	11 (8%)	

K5, Item A4 (item-total correlation .491)

4.	to blame: We are not **to blame**.		Facility	Upper	Lower	D
	a.	in total agreement	.03	.00	.09	-.09
	b.	interested	.03	.00/	.14	-.14
	c.	accusing anyone	.39	.18	.53	-.35
	d.	the cause of the problem	.55	.82	.25	.57
		No attempt		0 (0%)	13 (9%)	

K5, A5 (item-total correlation .585)

5.	stand for: It **stands for** wealth.		Facility	Upper	Lower	D
	a.	represents	.96	.99	.89	.10
	b.	is the cause of	.01	.00	.04	-.04
	c.	the opposite of	.01	.00	.04	-.04
	d.	is related to	.02	.01	.04	-.03
		No attempt		0 (0%)	1 (.7%)	

K5, A6 (item-total correlation .560)

6.	by far: She is **by far** the most intelligent.		Facility	Upper	Lower	D
	a.	trying to be	.03	.00	.07	-.07
	b.	not at all	.07	.02	.17	-.15
	c.	really	.88	.98	.69	.29
	d.	sometimes	.02	.01	.07	-.06
		No attempt		1 (0.5%)	0 (0%)	

K5, B7 (item-total correlation .661)

7.	keep on: We had to **keep on**.		Facility	Upper	Lower	D
	a.	not give anything away	.01	.01	.04	-.03
	b.	not tell anyone	.01	.00	.06	-.06
	c.	rest	.02	.00	.08	-.08
	d.	continue	.95	.99	.82	.07
		No attempt		0 (0%)	0 (0%)	

K5, Item B8 (item-total correlation .378)

8.	over time: **Over time** it was cheaper.		Facility	Upper	Lower	D
	a.	long ago	.61	.38	.73	-.35
	b.	eventually	.27	.57	.06	.51
	c.	when it was too late	.08	.03	.13	-.10
	d.	at the perfect moment	.04	.03	.08	-.05
		No attempt		12 (6%)	10 (7%)	

K5, B9 (item-total correlation .626)

9.	come up to: He just **came up to** me.		Facility	Upper	Lower	D
	a.	approached	.87	.99	.59	.40
	b.	rejected	.09	.00	.27	-.27
	c.	did not like	.02	.00	.08	-.08
	d.	copied	.02	.01	.06	-.05
		No attempt		1 (0.5%)	24 (17%)	

K5, B10 (item-total correlation .671)

10.	straight away: They did it **straight away**.		Facility	Upper	Lower	D
	a.	immediately	.80	.99	.33	.66
	b.	the correct way	.13	.01	.41	-.40
	c.	slowly	.02	.00	.08	-.08
	d.	because they wanted to	.05	.00	.18	-.18
		No attempt		0 (0%)	17 (12.5%)	

K5, B11 (item-total correlation .625)

11.	shut up: He won't **shut up**.		Facility	Upper	Lower	D
	a.	be shy	.02	.00	.03	-.03
	b.	stop trying	.01	.00	.05	-.05
	c.	lock the door	.00	.00	.08	-.08
	d.	stop talking	.97	1.00	.91	.09
		No attempt		0 (0%)	4 (3%)	

K5, B12 (item-total correlation .620)

12.	a handful of: **A handful of** us went.		Facility	Upper	Lower	D
	a.	a few	.93	.97	.80	.17
	b.	none	.00	.00	.02	-.02
	c.	most	.05	.02	.14	-.12
	d.	all	.01	.00	.05	-.05
		No attempt		1 (0.5%)	5 (4%)	

K5, C1 (item-total correlation .497)

1.	take for granted: She **took** it **for granted**.		Facility	Upper	Lower	D
	a.	kept it	.14	.08	.21	-.13
	b.	did not give it importance	.55	.83	.26	.57
	c.	wanted it a lot	.12	.05	.17	-.12
	d.	thought about it carefully	.19	.04	.36	-.12
		No attempt		3 (1.6%)	23 (19%)	

K5, C2 (item-total correlation .501)

2.	as of: It changes **as of** today.		Facility	Upper	Lower	D
	a.	starting	.51	.85	.20	.65
	b.	sometime	.06	.03	.13	-.10
	c.	perhaps	.17	.04	.30	-.26
	d.	because of	.26	.09	.38	-.29
		No attempt		11 (5%)	20 (17%)	

K5, Item C3 (item-total correlation .455)

3.	would appear: It **would appear** it's true.		Facility	Upper	Lower	D
	a.	cannot be that	.02	.01	.07	-.06
	b.	is certain that	.11	.02	.18	-.16
	c.	seems that	.73	.90	.61	.29
	d.	is assumed that	.14	.07	.14	-.07
		No attempt		7 (4%)	3 (2%)	

K5, Item C13 (item-total correlation .456)

13.	can tell: You **can tell.**		Facility	Upper	Lower	D
	a.	may speak	.58	.26	.79	-.53
	b.	are smart	.07	.10	.04	.06
	c.	can see	.28	.62	.05	.57
	d.	might	.07	.02	.12	-.10
		No attempt		2 (1%)	0 (0%)	

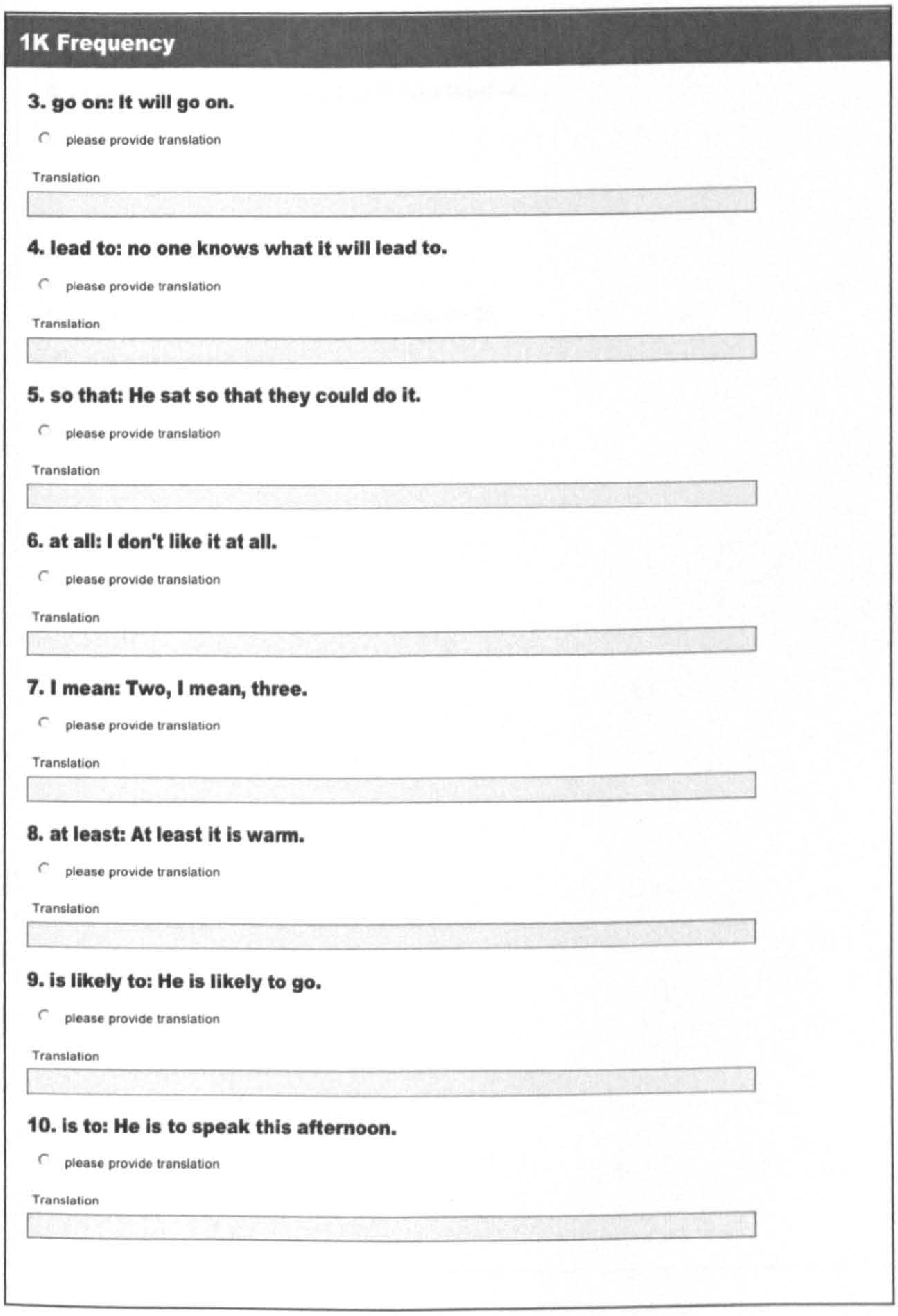
K5, Item C14 (item-total correlation .162)

14.	under way: It is **under way**.		Facility	Upper	Lower	D
	a.	coming	.56	.57	.43	.14
	b.	happening now	.25	.31	.28	.03
	c.	stopped	.08	.05	.11	-.06
	d.	an obstacle	.11	.08	.17	-.09
		No attempt		24 (12%)	18 (15%)	

K5, C15 (item-total correlation .497)

15.	turn down: She **turned down** the money.		Facility	Upper	Lower	D
	a.	hid	.14	.03	.22	-.19
	b.	lost	.27	.11	.48	-.37
	c.	made	.02	.01	.03	-.02
	d.	refused	.57	.86	.27	.59
		No attempt		3 (2%)	11 (9%)	

Appendix 16 – Online version of PVST for Brazil validation study



1K Frequency

3. go on: It will go on.

please provide translation

Translation

4. lead to: no one knows what it will lead to.

please provide translation

Translation

5. so that: He sat so that they could do it.

please provide translation

Translation

6. at all: I don't like it at all.

please provide translation

Translation

7. I mean: Two, I mean, three.

please provide translation

Translation

8. at least: At least it is warm.

please provide translation

Translation

9. is likely to: He is likely to go.

please provide translation

Translation

10. is to: He is to speak this afternoon.

please provide translation

Translation

1K Frequency

3. go on: It will go on.

- a sleep
- b. repeat
- c. be fast
- d. continue

4. lead to: no one knows what it will lead to.

- a. want
- b. have inside
- c. cause in the future
- d. find

5. so that: He sat so that they could do it.

- a. to make it possible that
- b. because
- c. very slowly and then
- d. before

6. at all: I don't like it at all.

- a. all the time
- b. in any way
- c. at first
- d. sometimes

7. I mean: Two, I mean, three.

- a. I am guessing
- b maybe
- c. then later
- d. I correct myself

8. at least: At least it is warm.

- a. other things may be bad, but
- b. many days have passed and now
- c. I cannot believe that
- d. the least important thing is

Appendix 17 – The PHRASE List A to Z

Phrase	Frequency
A BIT	19618
A BIT OF A	1599
A CASE OF	888
A COUPLE OF	7007
A DEGREE OF	921
A FEW	26451
A FURTHER ('ANOTHER')	6121
A GO ('ATTEMPT')	1093
A GOOD ('AT LEAST')	1298
A GOOD/GREAT DEAL ('MUCH')	5126
A HANDFUL OF	965
A LITTLE	20296
A LONG WAY	1557
A LOT	22332
A MERE	1410
A NUMBER OF	15090
A QUESTION OF	1174
A RANGE OF	5651
A SINGLE ('ANY')	8710
A VARIETY OF	4283
ABOUT TO	4600
ABOVE ALL	2212
ACCOUNT FOR	4642
ACT ON	1296
ADD TO	1424
AFFORD TO	1989
AFTER ALL (adv.)	5197
AIM TO	3415
AIMED AT	2573
ALL BUT	2214
ALL OVER ('EVERYWHERE')	4420
ALL RIGHT	5230
ALL SORTS OF	1535
ALL THE TIME	3527
ALL THE WAY	2007
ALL TOO	1571
ALLOW FOR ('CALCULATE IN')	1105
ALONG WITH	4948

AMOUNT TO	1556
AND ALL THAT	989
AND SO ON	4584
APART FROM	6287
APPEAL TO	3299
AS A RESULT	7939
AS A WHOLE	3615
AS FAR AS	4619
AS FOLLOWS	2620
AS FOR	3157
AS GOOD AS ('LIKE')	1043
AS IF	14470
AS IT WERE	985
AS LONG AS	5084
AS OF	1069
AS OPPOSED TO	1615
AS SOON AS	5323
AS SUCH	2290
AS THOUGH	4988
AS TO	11535
AS USUAL	1287
AS WELL AS	18041
AS WELL	11519
AS YET	1423
AT A TIME ('SIMULTANEOUSLY')	1340
AT ALL	14650
AT BEST	844
AT FIRST	4275
AT LAST	4306
AT LEAST	25034
AT ONCE	3684
AT ONE TIME	1014
AT PRESENT	2847
AT RISK	1419
AT THE EXPENSE OF	1086
AT THE MOMENT	5001
(AT) THE OUTSET	963
AT THE SAME TIME ('CONVERSELY')	2892
AT THE TIME ('WHEN THIS HAPPENED')	5282
AT THIS POINT	1884
AT TIMES	2014
AT WORK	787
BACK UP	1042
BACKED BY	850
BASED ON	11440
BE EXPECTED TO	5964

BE LIKELY TO	15854
BEAR IN MIND	1398
BELIEVE IN	2450
BETTER OFF	980
BOTHER TO	1046
BOUND TO	2278
BREAK UP	1595
BRING ABOUT	2022
BRING UP	1958
BUT THEN (AGAIN)	1589
BY CONTRAST	1430
BY FAR	925
BY MEANS OF	1602
BY NO MEANS	1925
BY NOW	2044
BY THE TIME	3607
BY THE WAY	1433
BY VIRTUE OF	954
BY WAY OF	1404
CALL FOR	6243
CALL ON	2016
CAN TELL	1011
CARE FOR	3004
CARE TO	1223
CARRY ON	3759
CARRY OUT	10753
CATCH UP	1095
CHOOSE TO	3099
COME ABOUT	787
COME ACROSS	1362
COME BACK	6772
COME ON	4519
COME OUT	6031
COME TO ('EVOLVE TO')	4970
COME TO TERMS WITH	839
COME UP TO	881
COME UP WITH	1898
COMMON SENSE	1049
CONCERNED WITH	4619
CONSIST OF	5362
CONSISTENT WITH	1315
CONTRARY TO	1303
COULD HARDLY	1911
DAY TO DAY	1289
DEAL WITH	13634
DO(ING) SO	1647

DUE TO	10454
EACH OTHER	10160
END UP	3285
ENTITLED TO	3113
EVEN SO	1363
EVEN THOUGH	5664
EVER SINCE	1622
EXCEPT THAT	1296
FACED WITH	1658
FAIL TO	10263
FEEL LIKE	2431
FILL IN	1187
FIND ONESELF	3917
FIND OUT	6499
FIRST OF ALL	1352
FOCUS ON	3703
FOLLOW UP	2128
FOLLOWED BY	4816
FOND OF	831
FOR ALL ('CONSIDERING')	1182
FOR GOOD	1061
FOR INSTANCE	7138
FOR LIFE	1172
FOR LONG	1553
FOR SALE	1379
FOR SOME TIME	1725
FOR THE MOMENT	1092
FOR THE SAKE OF	865
FOUND TO	2104
FREE FROM	991
FROM TIME TO TIME	1627
FULL TIME	2761
GET AWAY WITH	824
GET AWAY	1165
GET BACK	3178
GET INTO	3508
GET ON ('RELATE')	944
GET ON WITH	1535
GET ON/OFF (TRANs)	3656
GET OUT	4858
GET TO ('ARRIVE AT')	3979
GET TO ('OPPORTUNITY')	1437
GET UP	3857
GIVE RISE TO	1613
GIVE UP	3997
GIVEN THAT	1247

GO AHEAD	2023
GO AWAY	2150
GO BACK	3722
GO FOR	3421
GO INTO	2595
GO OFF	2551
GO ON	15610
GO OUT	5173
GO ROUND	1022
GO THROUGH	5857
GOING TO (FUTURE)	28259
GOOD AT	1562
GREATER THAN	1666
HAD BETTER	2022
HALF PAST	1325
HAND OVER	1798
HANG ON	2074
HAPPEN TO (BE)	1664
HAVE A LOOK	2464
HAVE GOT (+NP)	12734
HAVE GOT TO	12270
HAVE TO	83092
HEAD TO	1014
HEARD OF	2403
HELD THAT ('BELIEVED')	1661
HOLD UP	791
HOW ABOUT	805
I MEAN	23616
IF ONLY	1830
IF SO	1488
IF YOU LIKE	1256
I'M AFRAID	1495
IN (THE SENSE) THAT	4805
IN A POSITION TO	864
IN A SENSE	1014
IN A WAY	2684
IN ACCORDANCE WITH	2017
IN ADDITION (TO)	7822
IN ADVANCE	1983
IN ANY CASE	2159
IN CASE	2536
IN CHARGE	2432
IN COMMON	1890
IN CONJUNCTION WITH	1270
IN CONTRAST (TO)	2229
IN DETAIL	1473

IN EFFECT	1995
IN FACT	15983
IN FAVOUR	4073
IN FRONT OF ('BEFORE')	5190
IN FULL	1698
IN HAND	1443
IN ITSELF	1855
IN LINE WITH	1240
IN MIND	2638
IN NEED	1122
IN ONE'S OWN RIGHT	891
IN ORDER TO	12762
IN OTHER WORDS	3159
IN PART	2652
IN PARTICULAR	7092
IN PLACE	2805
IN PRACTICE	3609
IN PRINCIPLE	1429
IN QUESTION	1898
IN RESPECT OF	2909
IN RETURN (FOR)	1720
IN SHORT	1602
IN SO FAR AS	2344
IN SPITE OF	2676
IN TERMS OF	9881
IN THE ABSENCE OF	1538
IN THE COURSE OF	2585
IN THE END	3050
IN THE EVENT (OF)	2998
IN THE FACE OF	1457
IN THE FIRST PLACE	1897
IN THE INTEREST OF	938
IN THE LIGHT OF	1789
IN THE MEANTIME	1161
IN THE SAME WAY	1784
IN THE WAY	3013
IN THEORY	1065
IN THIS RESPECT	1110
IN TIME	3566
IN TOUCH (WITH)	3060
IN TURN	3558
IN VIEW OF	1477
IN WHICH CASE	872
INSTEAD OF	6907
IS TO	15232
IT TAKES	3670

JUST ABOUT	1622
JUST AS ('WHEN')	869
KEEP ON	788
KEEP UP	1674
KEY TO	1499
KIND OF	3510
KNOWN TO	3091
LARGE SCALE	1610
LAST NIGHT	7992
LAY OUT	789
LEAD TO ('CAUSE')	13555
LED BY	2511
LET AL.ONE	1261
LIMITED TO	1346
LITTLE MORE THAN	1133
LONG AGO	1273
LONG BEFORE	1282
LONG TERM	5831
LOOK AFTER	4332
LOOK FOR	8377
LOOK FORWARD TO	2331
LOOK LIKE	6595
LOOK TO	790
LOTS OF	3605
MAKE ITS/ONE'S WAY	1446
MAKE OUT	842
MAKE SENSE	1608
MAKE SURE	5510
MAKE UP ('COMPRISE')	3394
MAKE UP ONE'S MIND	788
MAKE USE OF	1191
MANAGE TO	6234
MAY WELL ('COULD')	4931
MEANT TO	2098
MIGHT AS WELL	1348
MIND YOU	1342
MORE AND MORE	2468
MORE OR LESS	2579
MORE SO	887
MOST LIKELY	1376
MOVE ON	2245
NEVER MIND	915
NEXT DOOR	1449
NEXT TO	2882
NO DOUBT ('SURELY')	2791
NO GOOD	1528

NO IDEA	1670
NO LONGER	8556
NO MATTER	1888
NO MORE THAN ('ONLY')	3226
NO ONE	9597
NO SIGN OF	871
NO SUCH	1470
NO WONDER	820
NOT EVEN	3128
NOT ONLY	14110
NOTHING BUT	1466
OF COURSE	26966
OF LITTLE	1079
OH DEAR	1700
OH NO	2642
OH WELL	2129
OLD FASHIONED	1393
ON AVERAGE	1014
ON BEHALF OF	6734
ON BOARD	1698
ON ONE'S OWN	5240
ON THE BASIS (OF)	3515
ON THE GROUNDS	1713
ON THE MARKET	928
ON THE ONE HAND	1406
ON THE OTHER HAND	5267
ON THE PART OF	1965
ON THE ROAD	1400
ON THE WAY	5085
ON THE WHOLE	1238
ONCE AGAIN	3532
ONCE MORE	2146
ONE ANOTHER	2623
OPPOSED TO	1059
OR ANYTHING	1379
OR SO	4164
OR SOMETHING ('PERHAPS')	2683
OR TWO	6192
OR WHATEVER	1806
OTHER THAN	4380
OUGHT TO	5002
OUT OF ('DUE TO')	1453
OUT OF ('IN'/'FROM')	4361
OUT OF ('USING')	2907
OUT THERE	2513
OVER THE YEARS	1942

OVER THERE	2678
OVER TIME	1069
OWING TO	793
PART TIME	2343
PICK UP	9252
POINT OF VIEW	2864
POINT OUT	6325
PRIOR TO	3110
PROVE TO BE	1893
PROVIDE FOR	1370
PROVIDED THAT	1110
PUT FORWARD	1640
PUT IT ('SAY')	2449
PUT TOGETHER	1082
PUT UP	1571
QUITE A LOT	1260
RATHER THAN	21085
REFLECTED IN	1471
(WITH) REGARD TO	2630
RELY ON	3488
RESULT IN	5763
RID OF	2212
RIGHT NOW	1482
(BE) RUN BY	1540
RUN OUT (OF)	1430
SAID TO BE	3586
SEEK TO	5937
SET OFF	1644
SET OUT	4624
SET TO	2258
SET UP	11560
SHAKE ONE'S HEAD	3250
SHORT OF	1944
SHORT TERM	2574
SHOWN TO	1409
SHUT UP	1079
SIGHT OF	1988
SO AS TO	1954
SO CALLED	2944
SO FAR ('UNTIL NOW')	5018
SO FAR AS	1251
SO THAT	20966
SOME KIND OF	1678
SOME MORE	1539
SOMETHING ABOUT	2071
SOMETHING LIKE ('AROUND')	3092

SOMETHING LIKE THAT	1192
SOMETHING OF A	1154
SORT OF	13361
SORT OUT	2696
STAND FOR	977
STRAIGHT AWAY	794
SUBJECT TO	5218
SUCH A(N)	23894
SUCH AS	30857
SUCH THAT	1454
SUPPOSED TO	4586
SWITCH ON	1935
TAKE ACCOUNT OF	1217
TAKE ADVANTAGE	1949
TAKE CARE OF	1034
TAKE FOR GRANTED	1120
TAKE INTO ACCOUNT	2921
TAKE OFF	1957
TAKE ON	4451
TAKE OVER	5394
TAKE PART IN	2374
TAKE PLACE	10556
TAKE UP	4717
TEND TO	10504
THANKS TO	2159
THAT IS (REPHRASING)	5737
THAT MUCH	1781
THAT SORT OF THING	843
THAT WHICH	1491
THAT'S IT	2674
THE ABOVE	2608
THE BULK OF	968
THE CASE ('TRUE')	4794
THE EXTENT TO WHICH	1751
THE FOLLOWING	12963
THE FORMER	1842
THE LATTER (anaphor)	5519
THE LOT	789
THE MEANS	1626
THE ODD	1154
THE OTHER DAY	1066
THE SIGHT OF	1024
THE WHOLE THING	1015
THERE IS/ARE	59833
THESE DAYS	2440
THEY SAY	2962

THINGS LIKE THAT	1082
THINK ABOUT	7243
THINK SO	2033
THIS STAGE	2223
THOSE WHO	13951
THOUGHT OF (AS)	1065
TO BLAME	973
TO COME (FUTURE)	862
TO DATE	2600
TO DEATH	1714
TO DO WITH	5184
TO GO ('REMAINING')	801
TO ME ('IN MY OPINION')	1345
TO SOME EXTENT	1688
TO THE EXTENT	1253
TO THE POINT	1447
TOO MANY	2587
TOO MUCH	7123
TOUCH OF	982
TURN BACK	827
TURN DOWN	951
TURN INTO	2072
TURN ON	2266
TURN OUT	4354
TURN UP	2865
UNDER WAY	939
UP AND DOWN	898
UP TO (DECISIONS)	1807
UP TO (MAXIMUM)	8733
UP TO (UNTIL)	2409
UP TO DATE	1268
USED TO (ACCUSTOMED)	2831
USED TO (PAST)	14411
WAS TO	14366
WAY OUT	1309
WAY ROUND	1013
WEALTH OF	844
WELL BEING	809
WHAT ABOUT	3160
WHEN IT COMES TO	1188
WHETHER OR NOT	2824
WITH A VIEW TO	829
WITH RESPECT TO	1325
WORK ON	7600
WORK OUT	4432
WORTH OF	1299

WOULD APPEAR	1068
WOULD SAY	2128
WOULD YOU LIKE	1133
YET ANOTHER	1500
YET TO	1818
YOU SEE	7102