Developing a measure of L2 learners' productive knowledge of English collocations

Dale Brown

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

Cardiff University

Thesis submitted for the degree of PhD
Cardiff University
March 2018

Summary

Obtaining accurate measurements of L2 learners' productive knowledge of collocations has proven difficult. The goal of the work reported in this thesis was to develop and test a means of eliciting from learners a reliable and representative sample of their productive knowledge of collocations. The two main methods typically used for this purpose are demonstrated to suffer from a number of drawbacks, yet one instrument is identified as having potential. This instrument, *LexCombi*, originally devised by Barfield (2009a), presents noun cues to learners and asks for three collocates in response to each cue, which are then evaluated as either canonical or not. In this thesis, *LexCombi* is taken forward and, through an iterative series of empirical studies, developed further. Specifically, after trialling *LexCombi* and exploring how learners interact with it, the format is adapted to more clearly guide respondents towards producing collocations; the scoring of learners' responses is reviewed to gain a more complete picture of learners' knowledge; and a new set of cue words is trialled and selected to resolve a number of issues identified with the original cues. After this development process, an empirical evaluation of the final form of the instrument, *LexCombi 2,* is conducted and its capacity to provide useful data on learners' productive knowledge of collocations is evaluated. Following this, the empirical data is used to consider what can be learned about collocation knowledge using *LexCombi 2*. Explorations include the relationship between collocation knowledge and general L2 proficiency, the types of words that are used as responses to *LexCombi 2*, and how *LexCombi 2* scores are affected by different conceptions of collocation. Finally, the thesis considers the overall significance of this work for our understanding of collocation knowledge more generally.

Acknowledgements

A great many people have assisted me along the way in producing this thesis. I should first acknowledge Geoff Hall, who acted initially as my supervisor and got me started on this journey, Tess Fitzpatrick, who supervised the bulk of the empirical research, and Alison Wray for guiding the writing up of the research.

Many friends and colleagues gave practical support in various forms. I wish to thank Ken Kobayashi for help with translations; Trane Devore and Bob Perkins for acting as raters; Mike Linacre for advice on carrying out Rasch analyses; Yasuhiro Imao for developing the scoring program; Kimi Klassen, Mike Green and Ian Munby for collecting data from students on my behalf; Peter Thwaites and Matthew Rooks for recruiting informants for me; and Jon Clenton, Brandon Kramer, Lewis Murray and especially George Higginbotham for providing feedback on draft chapters of the thesis.

I also owe thanks to the many, many students and other informants who kindly gave up their time to provide me with the data on which this thesis is based.

Finally, I would like to thank my family, and in particular my wife, Yoko Brown, who has provided endless support, both practical and moral.

Table of contents

List of tables

List of figures

Abbreviations

ALC	the All Lists Combined scoring approach (Section 7.2.6)
BBI	*BBI Combinatory Dictionary of English*
BNC	British National Corpus
COCA	Corpus of Contemporary American English
EAT	*Edinburgh Associative Thesaurus*
ELF	English as a Lingua Franca
FDCAE	*A Frequency Dictionary of Contemporary American English*
L1	first language
L2	second or foreign language
OCD1	*Oxford Collocations Dictionary* (1st edition)
OCD2	*Oxford Collocations Dictionary* (2nd edition)
OSTI	Office for Scientific and Technical Information

Chapter 1 Introduction and overview

1.1 Introduction

Insofar as it is the aspiration of any second or foreign language (L2) learner to develop fluency and accuracy sufficient for effective communication with L1 users and other L2 users of the target language, one important element that needs to be mastered is collocation, that is, combinations of words that co-occur regularly.

The importance of collocation in a first language (L1) has long been recognised. Sinclair (1991), building on Firth's (1951, 1952/3, 1957) work, assigned collocation a central role in his description of two principles that govern language production: the open-choice principle and the idiom principle. The open-choice principle suggests text is formed by the application of grammatical rules which specify slots to be filled by any lexical item of the correct paradigm. However, Sinclair argues that this does not fully account for constraints on a speaker's choice of words. The idiom principle accounts for these further constraints; Sinclair's argument is that phrases and collocations, while appearing to be analysable, are in fact single choices: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (1991, p. 110).

Similarly, much research has been directed at collocational learning and knowledge in L2 (as reviewed in Chapters 2 and 3), with the prevailing view being that collocation is an area of difficulty for learners. Bahns and Eldaw (1993), for example, reported that "L2 learners often have particular problems with word combinations, even at a relatively advanced level" (p. 101); Paquot and Granger (2012) describe collocations (and other multi-word units) as "notoriously difficult for learners" (p. 130); Bonk (2000) notes that researchers and teachers "have long spoken of learners' inadequate proficiency to produce acceptable collocations in a foreign language" (p. 9-10); and Nesselhauf (2005), discussing collocations as one type of prefabricated unit, states that "knowledge of and the ability to use prefabricated units . . . [are] essential for the language learner; unfortunately, however, they also pose considerable difficulties, even for the advanced learner" (p. 2).

In particular, the production of collocations is seen as challenging for learners (Bonk, 2000; Laufer and Waldman, 2011) because collocations are often unpredictable and because L2 collocations may differ from collocations in the L1. Comprehension of collocations is viewed as less of an issue since many collocations are at least somewhat transparent and can be understood with little difficulty in context. Peters (2016) provides empirical evidence of this in a study that included both receptive and productive measures of collocation knowledge and concludes that "it is productive knowledge where learners' main difficulties with collocations lie" (p. 135).

Yet, while the challenge of collocations is oft-noted, it is also clear that learners have some knowledge of them. Even learners at relatively early stages in their learning make use of collocations such as *make a mistake*, *have breakfast* and *hard work* in their speech and writing.

This contrast between the widely held view of collocation as a difficult area for learners and the fact that learners at all levels of proficiency make use of at least some collocations raises a number of questions: How much productive knowledge of collocations do learners have? How does that knowledge develop? How does learners' knowledge of collocations vary with proficiency? What types of knowledge do they have or do they lack? Why do collocations appear to pose problems for learners while also clearly being acquired to some extent?

This thesis aims to provide a means of answering such questions through the development and testing of an effective instrument for eliciting L2 learners' productive knowledge of collocations. A persistent problem in L2 collocation research has been the accurate measurement of learners' knowledge, particularly their productive knowledge. Specifically, it has been difficult methodologically to elicit a reliable and representative sample of what they know (see Chapter 3).

First, however, there are some preliminary considerations to address.

1.2 Collocation and its relation to other fields of enquiry

Currently, in the field of applied linguistics, collocation is often described as one type of multi-word unit or one form of formulaic language. That is, "formulaic language" and "multi-word units" are used as superordinate terms under which collocation can be found alongside other subordinate types such as idioms and fixed

expressions. This is the case both for scholars primarily focused on formulaic language (e.g. Christiansen & Arnon, 2017; Ellis, Simpson-Vlach, & Maynard, 2008; Schmitt & Carter, 2004; Siyanova-Chanturia & Martinez, 2014; Wood, 2015) and for those with a primary focus on collocations (e.g. Durrant & Schmitt, 2009; Gablasova, Brezina, & McEnery, 2017; Handl & Graf, 2010; Henriksen, 2013; Sonbul, 2015; Wolter & Yamashita, 2015).

Collocation can also, however, be seen as having its roots in a number of separate, though related, fields of enquiry. The first significant strand is discussions of collocation in linguistics. Firth's work is usually described as the starting point, with his ideas being taken on by scholars such as Halliday, McIntosh, Mitchell and Greenbaum, who formalised some aspects of the study of collocation, used collocation as a means of linguistic description and made theoretical suggestions. Sinclair's work in this area was the most sustained and significant, leading in time to his important suggestions for linguistic theory, briefly described above (Barnbrook, Mason, & Krishnamurthy, 2013; Nesselhauf, 2004).

A second important strand has been lexicography. One of the first significant discussions of collocation was in Palmer's *Second interim report on English collocations* (1933), the ideas within which were taken on by Hornby in his lexicographic work (see Barnbrook, et al., 2013; Cowie, 1998b). Lexicographic description, in particular the compilation of learners' dictionaries, has played an important part in the development of the concept of collocation. In parallel, collocation has also become a key concern of lexicography. It is now seen as central to the identification of the senses of a word and the exemplification of the collocational behaviour of words is a primary concern in learners' dictionaries (Rundell & Kilgarriff, 2011).

The third strand that has influenced the developing notion of collocation has been corpus linguistics. The study of collocation is a central aspect of corpus linguistics, providing insights into language which, it is argued, are not available to intuition. The advancement of new ways of identifying collocations has also been a central concern in the field. At the same time, corpus linguistics has had a crucial role in making collocations visible and thereby raising the awareness of linguists and language professionals about collocation and other forms of recurrence (Xiao, 2015).

These three fields are of course linked, with Sinclair in particular having a key

role as an important figure in all three. Sinclair took on theoretical ideas about collocation and began exploring them through computer corpora, thereby playing a major role in establishing the practices of corpus linguistics. He then went on to lead the COBUILD project (a project which itself involved considerable development and use of the first large corpus), which changed the face of lexicography and established several current conventions in the field.

Collocation can then be seen as a concept with roots in theoretical ideas in linguistics, lexicography and corpus linguistics, as depicted in Figure 1.1, but at the same time the development of the concept of collocation has influenced these fields. These three areas are also themselves of course closely inter-related, as the figure indicates.

Figure 1.1: Collocation and related fields of enquiry.

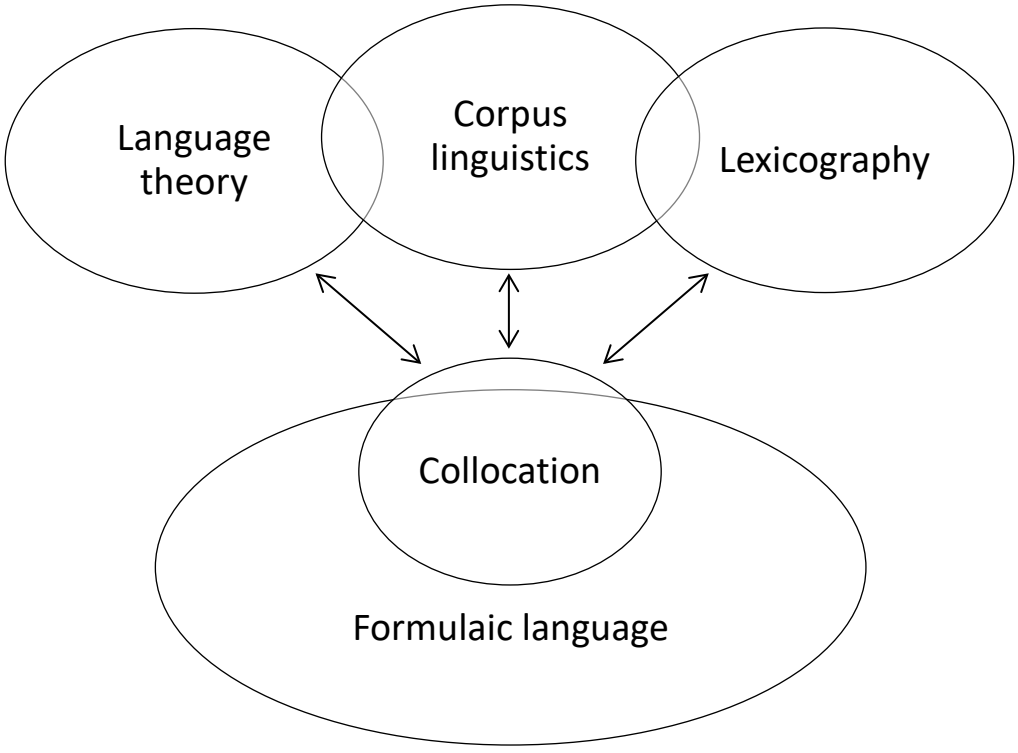


Figure 1.1 also depicts the fact that, as described above, collocation is often, though not always, described as one form of formulaic language. Is this the case? For Wray (2002), collocations "hover at the edge of definitions of formulaicity" (p. 47) since they are less fixed and more fluid than other types of formulaic sequence, being about tendencies and preferences. Bonk (2000) contrasts the terms

"collocation" and "formulaic speech" and suggests that whereas the former is about links between words in the mental lexicon based on lexical and semantic characteristics, the latter is about chunked storage and psycholinguistic reality. Whether collocation is a form of formulaic language ultimately depends of course on how each term is defined, since neither concept has a settled definition. However, if formulaic language is understood as a broader phenomenon to do with lexical patterning (Schmitt, 2004), collocation (under most definitions) would seem to fit under its umbrella, and this may be how the scholars cited above see the situation. Alternatively, it may be that though collocation does not fit entirely within formulaic language, the overlap is such that it is useful to see collocation as one form of formulaic language.

In this thesis, collocation will be treated as one type of formulaic language. Accordingly, literature on both collocation itself and from the wider field of formulaic language will be referred to.

1.3 The significance of collocation

Collocations are widely seen as significant and this is for two primary reasons. The first reason is their sheer ubiquity. Hoey (2005) suggests that collocations are pervasive in discourse and much of the language we produce is characterised by the stringing together of chains of collocations. Indeed, it was the ubiquity of collocations which in part led Sinclair (1991) to postulate the idiom principle.

There have been a number of studies that quantify this ubiquity to an extent. With respect to formulaic language, Conklin and Schmitt (2012) review a range of studies that have attempted to estimate the proportion of text that is formulaic and suggest that between one third and one half of discourse is made up of formulaic language. As for collocations, Howarth (1998b) found, looking at frequently used verbs in a corpus of L1 academic writing, that 38% of the verb + noun combinations were restricted collocations or idioms, as reported in Section 3.2.1. Also reported in that section, Laufer and Waldman (2011) found that 10.2% of the occurrences of frequent nouns in a corpus of writing by L1 school and university students were in verb + noun collocations. Siyanova and Schmitt (2008) report that 51.9% of the adjective + noun combinations identified in a corpus of 22 essays by L1 English university students were collocations as judged by their frequency of occurrence in

the British National Corpus (BNC). These figures are quite disparate, but even if collocations account for just 10 percent of discourse, they would nonetheless merit serious attention as a major linguistic phenomenon.

The second reason collocations are significant is that they are seen as having a number of important functions. According to Henriksen (2013), collocational competence enables language users to (1) sound like L1 users, (2) process language fluently, (3) establish islands of reliability which free up cognitive energy for other tasks, (4) disambiguate polysemous words, and (5) understand connotational meaning. Handl and Graf (2010) add that collocations have a pragmatic role in that they form part of communicative competence, and a developmental role in aiding acquisition, while Nesselhauf (2005) suggests collocations additionally support comprehension and indicate group membership. Barfield (2009b), furthermore, suggests that for learners studying in an L2 academic context the learning of collocations can constitute the acquisition of content knowledge.

More generally, collocations, as one form of formulaic language, may be part of the answer to Pawley and Syder's (1983) puzzles of "nativelike[1] selection" and "nativelike fluency". Nativelike selection refers to the fact that there are standard, preferred ways of expressing ideas, despite the fact that following the rules of syntax and using the mental dictionary one might come up with numerous ways of expressing a particular idea. That is, certain selections, which are somewhat arbitrary and unpredictable, from among the possible linguistic realizations of an idea, are nativelike. Nativelike fluency refers to the fact that although producing speech is an intense mental activity and the human capacity for encoding is limited, in routine conversation people seem able to go beyond these limits. Pawley and Syder argue that what they call "memorised sequences" and "lexicalised sentence stems" explain both these phenomena.

Wray (2002, 2017) develops these ideas further and argues that formulaic language is a linguistic solution to a non-linguistic problem. That problem is how to get what you want, and it is argued that formulaic language makes this possible in a

[1] The term *native* (and its counter-part *non-native*) has attracted considerable attention in recent years (e.g. Cook, 1999; Dewaele, 2018). In this thesis, *L1 user* and *L2 user* are generally used in preference to the terms *native* and *non-native*. However, *native* and *non-native* are used when they appear in direct quotes from scholars or when discussing work that is concerned with these terms or their implications.

number of ways: it reduces processing effort for the speaker, thereby contributing to fluency; it minimises the chances of a message being misunderstood; it signals that speaker and hearer belong to a shared community; it helps differentiate old from new information; and, perhaps most importantly, makes processing easier for the listener. As noted earlier, even L2 learners of relatively low proficiency make use of some collocations in the early stages of L2 production, with those collocations seeming to facilitate communicative expression for these learners, while Millar (2011) provides empirical evidence indicating that the use of standard collocations makes the processing of language easier for listeners. Collocations, then, as one type of formulaic language, may be part of this linguistic solution.

In addition to the above arguments, there are also a number of empirical studies showing the importance of collocations and formulaic language to successful L2 language use. It has been found, for example, that the number of formulaic word combinations used by learners during oral tasks correlates well with oral proficiency ratings (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Stengers, Boers, Housen, & Eyckmans, 2011); that among a large number of lexical variables, the strongest predictor of L2 learners' speaking proficiency scores was trigram frequency (i.e. the frequency of three-word phrases) (Kyle & Crossley, 2015); that gains in L2 fluency (in terms of temporal measures) over a period of time are associated with gains in the use of formulaic sequences (Wood, 2010); that there is a strong positive correlation between knowledge of collocations and speaking proficiency (Hsu & Chiu, 2008); and that the average degree to which the word pairs in a text are collocations has a strong positive correlation with evaluations of text quality (Bestgen, 2017).

Having clarified the links between collocation and other related fields of enquiry, and having discussed why collocation is significant in language, it is possible to proceed to a more detailed consideration of the phenomenon and deal with the issue of how collocation can be defined (Chapter 2). This will be followed in Chapter 3 by a consideration of both what is currently known about L2 learners' productive knowledge of collocations and the methods used to investigate this. While a number of issues are identified with previous investigations, one instrument is considered to show potential, and thus the subsequent chapters (Chapters 4-9) set out to trial and make improvements to that instrument. Chapter 10 then begins to

explore what we can discover from data collected with the final form of the instrument, broader issues raised across the thesis are discussed in Chapter 11, before Chapter 12 concludes the thesis.

Chapter 2 Defining collocation

2.1 Introduction

If we are to seek a greater understanding of L2 learners' productive knowledge of collocations, then the first, fundamental, question that must be answered is: What are "collocations"? There is no settled definition of "collocation" and a wide array of definitions can be found in the literature. This chapter sets out to review this issue and settle on a definition of collocation to guide this thesis. It should be noted that in this thesis, for practical rather than theoretical reasons, only two-word combinations will be explored, even though collocations can be of three or more words. In this respect, the thesis follows the majority of studies of collocation.

The first part of the chapter surveys discussions of collocation and evaluates elements that are important in considering its definition. Section 2.2 outlines two broad approaches to collocation, the phraseological approach and the frequency-based approach. Section 2.3 then considers a number of more particular areas of debate in work on collocation. The second part of the chapter engages with collocation from a different direction, by exploring what empirical studies on the processing of formulaic sequences and collocations have revealed about its psychological nature (Section 2.4). Section 2.5 draws the strands together, to offer a definition of "collocation" that is compatible with the evidence reviewed.

2.2 Two approaches to collocation

A number of scholars (e.g. Barfield & Gyllstad, 2009; Granger & Paquot, 2008; Nesselhauf, 2004) discuss two traditions in research on collocation: the phraseological approach and the frequency-based approach. An understanding of these two approaches is useful for placing in context many of the studies on collocation to be reviewed in this chapter and later in this thesis.

2.2.1 The phraseological approach

The phraseological approach can be traced back to Palmer's research on vocabulary and Hornby's work on lexicography in the 1930s (see Barnbrook, et al., 2013; Cowie, 1998b) and has been heavily influenced by Russian scholarship. Key recent figures have been Benson, Mel'čuk, Hausmann, Cowie, Howarth and Nesselhauf

(see Nesselhauf (2005), Gyllstad (2007) and Barnbrook, Mason & Krishnamurthy (2013) on the history of scholarship on collocation). This approach seeks to define multi-word units linguistically: that is, it describes linguistic criteria by which phraseological units can be identified and by which one type of phraseological unit can be distinguished from another. The two principal linguistic criteria used are semantic opacity and restrictedness, each of which is viewed as a scale. Collocations are seen as occupying a certain region along these scales, being less restricted and less opaque than idioms at one extreme, but more restricted and more opaque than free combinations (which "lie outside the limits of phraseology altogether" (Cowie, 1998a, p. 6)) at the other.

One example of this approach is Howarth's (1998a) phraseological continuum, which is a refinement of Cowie's (1988) earlier work. This continuum is "derived from the application of such criteria as restricted collocability, semantic specialization, and idiomaticity, each of which is gradable" (Howarth, 1998a, p. 28). Howarth divides the continuum into four categories: free combinations, restricted collocations, figurative idioms and pure idioms. Free combinations (e.g. *blow a trumpet*) "consist of elements [which are] used in their literal senses and [are] freely substitutable" (p. 28); restricted collocations (e.g. *blow a fuse*) "have one component . . . that is used in a specialized, often figurative sense only found in the context of a limited number of collocates" (p. 28); figurative idioms (e.g. *blow your own trumpet*) "have metaphorical meanings in terms of the whole and have a current literal interpretation" (p. 28); and pure idioms (e.g. *blow the gaff*) "have a unitary meaning that cannot be derived from the meanings of the components" (p. 28).

The phraseological approach has faced a number of criticisms. First, as Nesselhauf (2005) observes, phraseological approaches often use a mix of criteria (a problem Wray (2002) has noted with taxonomies of formulaic language also). There seems to be an assumption that the various criteria coincide with each other, and thus there is little clarity about whether, for example, a combination should be considered a collocation if it qualifies under one criterion but not under another. For instance, the combination *run a business* is relatively unrestricted (you can *run a factory*/*shop*/*organisation*/*project*/*campaign*/*country*/*programme* and more, while you can *operate*/*manage*/*control*/*head*/*lead*/*be in charge of a business*), but may be considered somewhat opaque depending on how *run* is defined.

A second issue is that phraseological approaches do not allow for the possibility that a collocating pair of words is selected as a single choice. Phraseological approaches suggest that one word is first selected, which then restricts the selection of the other (Howarth, 1998b). For example, Hausmann's (1991, 1999) work describes one part of a collocation as the "basis", an autonomous word selected first on the basis of its meaning, and the other as a "collocator", a word added to the basis with a meaning sense dependent on the basis. Indeed, one way that Nesselhauf (2005) distinguishes collocations from idioms is that in idioms "the whole combination is selected at once" (p. 32).

Third, the reliability of researchers' identification of collocations may be doubted. The phraseological approach requires researchers to make a number of decisions, each of which may be challenging. Depending on the details of the particular scheme, the researcher may need to:

- Distinguish between senses of words (e.g. Are *front* in *front door* and in *front row* used in the same sense?).
- Determine whether combinations are possible (e.g. Is *run a situation* a possible combination?).
- Determine a combination's degree of semantic opacity (e.g. To what extent is *double vision* opaque?).
- Identify other words that may co-occur with a word (e.g. What words similar to *avid* co-occur with *reader*?).
- Decide how many co-occurring synonymous words constitute restrictedness (e.g. *Voracious*, *devoted*, *passionate*, *ardent*, *insatiable* and *fanatical* all co-occur with *reader*; does this constitute restrictedness?).

In a study of just one of the above considerations, Boers and Webb (2015) investigated ratings of semantic opacity for idioms and found poor inter-rater agreement among L1-user language teachers. If this applies to the other considerations above, there could be significant problems with the reliability of claims in published studies. These concerns would be best addressed with evidence that the researcher had used multiple raters and checked for the similarity of their judgements. However, as will be seen later, phraseological studies seldom involve multiple raters and few provide statistics on inter-rater or intra-rater reliability.

A particular difficulty for phraseological approaches is distinguishing

collocations from free combinations (see Nesselhauf, 2005). This may be because the notion of free combinations is more problematic than many phraseologists recognise. Sinclair argued against the idea that words are selected freely, since "when writers and speakers co-select words, they create a new meaning which makes other instances of the same individual words and other co-selections involving these same words irrelevant" (Cheng, Greaves, Sinclair, & Warren, 2009, p. 237). Hanks (2013) has argued that only phrases have meanings: a word in isolation has only meaning potential, with one element of this potential being realised depending on the phrase in which it is used. Wray's (2015) discussion of the concept of "word" suggests that we may never truly encounter the same word twice since each instance of what may appear the same word is in fact subtly different: thus comparing one use of a word with other uses of that word is misguided. These ideas suggest that there are no genuinely free combinations, and this may be why distinguishing them from collocations is so challenging.

2.2.2 The frequency-based approach

With its roots in the work of Firth (1951, 1952/3, 1957), the frequency-based approach was largely developed by Sinclair and has since been taken forward by a number of scholars. This approach sees collocations as bonded combinations of words, with those bonds arising from their frequency of co-occurrence. In research practice, a collocation is a pair of words co-occurring in texts within a certain distance of each other, specified by the researcher.

Firth was primarily concerned with collocation as an aspect of meaning. He argued that "words must not be treated as if they had isolate meaning[,] and occurred[,] and could be used[,] in free distribution" (1952/3, p. 18). He pointed out that the meaning of a word is partly defined by the words it frequently co-occurs with: "One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*" (1951, p. 196). This position is summarised in his oft-quoted statement: "You shall know a word by the company it keeps" (1957, p. 179). Firth also seems to have regarded collocations as single units in some sense: "The collocation of a word . . . is not to be regarded as mere juxtaposition, it is an order of *mutual expectancy*. The words are mutually expectant and mutually prehended" (1957, p. 181).

Sinclair led a project for the UK government's Office for Scientific and Technical Information (OSTI), and the report of this project (Sinclair, Jones, & Daley, 1970) may be considered the foundational text of the frequency-based approach. After initially defining collocation as "the co-occurrence of two items in a text within a specified environment" (p. 10), the authors distinguished "significant collocation" from "casual collocation". The former was defined as word combinations that "co-occur more often than their respective frequencies and the length of the text in which they appear would predict" (p. 10), and the report set out how the significance of word combinations could be calculated statistically. Following the OSTI report, "collocation" came to be used by frequency-oriented scholars to mean "significant collocations". Sinclair et al. also investigated what the "specified environment" should be and found that 95% of the influence a word has on others is found within a span of +/-4 words, a span which became the norm.

In the frequency-based approach today, researchers use corpora and specialised software to find collocations, using two principal methods. One technique, pursued in particular by Biber and colleagues (e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004) identifies recurrent sequences of words (often referred to as "lexical bundles" or "n-grams"). The other follows Sinclair and, using one word as a node, counts the recurrence of items appearing within a certain span of the node. Collocations are then defined as items reaching a certain threshold on a variety of statistical measures (see Manning & Schütze (1999) and Evert (2008) for overviews of such measures).

Two commonly used statistical measures are *t*-scores and MI scores. *T*-scores are a measure of certainty of collocation, indicating the likelihood that the two words occur independently of each other. This is calculated as observed co-occurrence minus expected co-occurrence (calculated on the basis of the frequency of the component words), divided by the square root of observed co-occurrence. MI scores measure the strength of association, showing how much observed co-occurrence surpasses expected co-occurrence. MI scores are the log-transformed ratio between observed co-occurrence and expected co-occurrence (Evert, 2008). For example, in the Corpus of Contemporary American English (COCA) (Davies, 2008-), two of the most frequently occurring adjectives with the word *test* are *standardized* and *positive*, co-occurring 811 and 755 times respectively. In the corpus as a whole

standardized is considerably less frequent than *positive* (5,686 vs. 54,325 occurrences) and has an expected frequency with *test* of 7, as opposed to an expected frequency for *positive* with *test* of 66. Calculated as explained above, *standardized* has a *t*-score with *test* of 28 and *positive* a *t*-score of 25, while the MI scores are 6.8 and 3.5 respectively. Thus, while *standardized* and *positive* have a similar number of occurrences with *test* and similar *t*-scores, the MI scores are rather different (see also Sections 4.5.1 and 7.2.2).

There are two main issues with the frequency-based approach. First, Howarth (1998a) states that many statistically significant combinations identified in corpora are in fact free combinations: that is, they occur simply because of the subject matter in hand, and there is no restriction of form or meaning and therefore no issues for processing. Similarly, a number of scholars (e.g. Barnbrook, 1996; Henriksen, 2013) have observed that relying purely on frequency highlights combinations that are frequent simply because their component words are frequent but between which there is no relationship. This latter problem has been well noted since the time of the OSTI report, not least by frequency-oriented scholars themselves, and is why statistical measures which take account of the frequency of the component words are often used, rather than pure frequency alone (see Evert, 2008). Howarth's criticism, however, is more deep-seated in that it simply does not accept frequency as relevant to identifying word pairs that have a relationship which can be described as collocation.

A second issue is that, as with the phraseological approach, in its practical application, the frequency-based approach requires that researchers make a number of decisions. These include the span in which to conduct the search (while +/-4 is conventional, other choices are possible), the choice of statistic (a wide array of statistics have been developed), and the establishment of cut-off values (here too conventions have arisen, though without strong foundations). The frequency-based approach then, despite sometimes being described as "objective" and "without subjective elements" (Nguyen & Webb, 2017, p. 300), does involve an element of subjectivity: as Sinclair, Jones and Daley (1970) said: "although the significance tests in themselves are objective, they rest on essentially arbitrary choices made by their users" (p. 72). Wray (2008) has observed, regarding the identification of formulaic sequences, that in making methodological choices intuition often has a

role, and this is the case, though not always explicitly, even in frequency-based studies as researchers evaluate the list of collocations produced by the initial set of choices they have made and may adjust those choices as a result.

In current research on collocation, the frequency-based approach is perhaps more central, but certain aspects of the phraseological approach remain of interest to scholars. Most current studies of collocation involving L2 learners combine the two approaches in some way: for example, frequency-based techniques may be used to identify collocations before the phraseological approach is applied in refining and classifying the collocations identified. In this thesis also, as will be seen, both frequency-based and phraseological concerns are included.

2.3 Issues in research on collocation

In addition to debate over the two broad approaches outlined above, there are a number of more specific issues that have been discussed in research on collocation. These issues come to the fore when collocation is operationalised, since decisions, implicit or explicit, are required with regard to each issue. The issues, to be discussed below, are:

1. What should count as the base unit that one is observing?
2. Are both lexical and grammatical words of interest?
3. Is collocation limited to orthographically separate words?
4. Are collocations necessarily semantically opaque?
5. Do collocations necessarily have structural integrity?
6. Which variety of English should be the basis for identifying collocations?
7. Is collocation a textual or mental phenomenon?

In what follows, the intention is not to provide definitive answers to these questions, but to identify the main considerations that can contribute to operational decisions in empirical research.

2.3.1 Word forms and lemmas

Collocation is generally said to be about co-occurrence, but there are different views as to what it is that is co-occurring. Some would argue (e.g. Revier, 2009) that it is a mistake to see collocations as involving units: the collocation itself is a unit; one that appears to be divisible, but in fact is not. This is supported by Bybee's (2002) linear fusion hypothesis, which suggests very frequently used combinations fuse and

become forms themselves in the lexicon. Similarly, both Sinclair and Hanks have argued that the phrase is a more important unit than the word. Sinclair (2008) states that "the normal primary carrier of meaning is the phrase and not the word; the word is the limiting case of the phrase, and has no other status in the description of meaning" (p. 409), and Hanks (2013), as noted earlier, argues that only phrases have meaning, with isolated words having merely meaning potential.

Words are, however, salient and, crucially, appear readily identifiable, and perhaps as a result the majority of work on collocation essentially treats it as a property of words. The issue thus becomes what is counted as a word, and one particular area of debate has been whether collocational relationships exist between word forms or between lemmas.

Among many researchers with a frequency orientation (e.g. Hoey, 2005; Kjellmer, 1994), individual word forms have their own collocates, a position which may stem from the OSTI report (Sinclair, et al., 1970). This demonstrated that conflating different forms of nouns and different forms of verbs did not provide additional information on collocational behaviour (i.e. there were more significant collocations when looking at forms separately than when forms were conflated). Durrant (2009) notes a further concern that "lemmatisation has the potential to disguise important differences in collocational preferences between different forms of a lemma" (p. 162). Taylor (2012) shows, for example, that different inflections of the verb *decide* have different sets of collocates, leading him to suggest that "each of the derived and inflected forms of a word may have its own, unique set of collocational preferences, distinct from those of the base word" (p. 160). One recent corpus-based study (Michelbacher, Evert, & Schütze, 2011), however, reports employing lemmatisation after preliminary work showed collocation to often exist between lemmas (though unfortunately no details on this preliminary work are provided).

In phraseology, this issue has not always received similar attention. This may be because there is not the same need for explicitness that is required by frequency-oriented scholars entering search terms in corpus query software. Some scholars do address the issue, however: Nesselhauf (2005), for example, is explicit in stating that in her study "the elements involved in collocations are assumed to be lexemes" (p. 25) and thus "it is assumed that combinations such as *pay attention*, *pays attention*,

paid attention and *attention was paid* are instantiations of the same collocation" (p. 25). No justification is given for this assumption, but it may stem from her belief that the term "collocation" refers to "an abstract unit of language and its instantiations in text" (p. 25). That is, collocations are seen as a mental phenomenon abstracted from textual instantiations, with those textual instantiations possibly involving various grammatical forms, as in the examples above.

Meanwhile, more pedagogically oriented work tends to focus on base forms, and, while not stated explicitly, these base forms appear to stand for lemmas. This can be seen, for example, in work on the teaching of collocations, such as Lewis (1993, 1997, 2000) and Boers and Lindstromberg (2009), as well as in materials designed for learners (e.g. McCarthy & O'Dell, 2005).

Finally, it should be recognised that the question of whether word forms or lemmas should be used applies to both parts of a collocation: that is, we can ask what type of lexical unit the entry/node is and we can ask the same question about the collocates of the entry/node.

2.3.2 Lexical and grammatical words

A second issue regarding the units involved is whether collocation is limited to lexical words or whether grammatical words are also included. The combination of two lexical words is termed "lexical collocation", while the combination of one lexical and one grammatical word is called "grammatical collocation". Many recent studies concentrate on lexical collocations, and in both phraseological and frequency-based research there is a particular focus on two types: verb + noun and adjective + noun collocations. A further distinction, occasionally confused with the above, is between collocation and colligation, the latter referring to the combination of a word with a particular class of items. Thus, the word *lap* tends to be preceded by a possessive determiner (*my*, *his*, etc.), with the determiner itself typically preceded by a preposition (*on*, *into*, etc.) (Taylor, 2012).

2.3.3 Orthographic words and word parts

A third, less frequently discussed, issue regarding the units is whether collocation can involve word parts as well as separate orthographic words. Work on collocation has in general failed to engage with questions of wordhood. As Granger and Paquot (2008) point out, while one of the foundations of the study of collocations is that the

units are made up of at least two words, “in view of the ambiguity surrounding the definition of the concept of word, this definition is not as helpful as may seem at first sight” (p. 32). They observe that “linguists often make quite arbitrary decisions as to what they include or exclude” (p. 32) and that often no explicit statement is made on this issue, such that “one regularly has to scan through the examples given by the authors” (p. 32). This issue was in fact recognised in early work on collocation: Sinclair and Jones (1974), in a paper summarising the OSTI report, note that they “present this account with a heavy reliance on the vagueness of [the] conventional notion of ‘word’” (p. 253). The problem of wordhood has long troubled linguists (Wray, 2015), while for lay people the problem seems equally vexing when it is probed (Davis, 2001). Indeed, it may be the case that the distinction made between collocation on the one hand and multi-morphemic words and compounds on the other is somewhat artificial, an issue discussed in Section 11.5.

2.3.4 Semantic opacity

As noted in Section 2.2.1, a central tenet of the phraseological tradition is semantic opacity: the idea that the meaning of a collocation is not fully compositional (i.e. a collocation’s meaning does not follow straightforwardly from combining the regular meanings of its component words). Phraseologists often describe a continuum of opacity and distinguish between several points along this continuum, with collocations occupying a region at a mid-point along this scale.

Most frequency-based work, in contrast, does not take account of semantic opacity. Kjellmer (1994) argues that since there is a continuum of semantic opacity, there are not really any categories at all beyond those created arbitrarily and since, in his view, collocations may span the continuum, it is better to disregard opacity in defining collocation.

Across research on collocations in general, there has been something of a move away from the requirement for semantic opacity; Cowie (1998a) observing that “studies of collocations have pushed the boundary that roughly demarcates the ‘phraseological’ more and more into the zone formerly thought of as free” (p. 20).

2.3.5 Structural integrity

Another notion emphasised by phraseological approaches is that collocations are units with structural integrity: that is, that there must be a grammatical relationship

between the component words. Indeed, Gyllstad (2007) describes this as an inherent feature of the methodology. Consequently, collocations can be described as belonging to structurally based categories: for example, the phraseologically based *BBI Combinatory Dictionary of English* (Benson, Benson, & Ilson, 2009) classifies collocations according to eight categories of grammatical collocation (e.g. noun + preposition, noun + *to* + infinitive) and seven categories of lexical collocation (e.g. adjective + noun, adverb + adjective).

Nesselhauf (2005) states that views on structural integrity are one way in which the phraseological and frequency-based approaches differ, and certainly the frequency-based approach can be applied with no account taken of structural integrity. However, there are some similar concerns in frequency-based research. Kjellmer (1994) argues that, as he terms it, grammatical well-formedness is an essential criterion for collocation, since frequency of occurrence alone may designate combinations such as *but too* and *night he* as collocations; Evert (2008) distinguishes between syntactic collocations (elements syntactically related), textual collocations (elements within a single sentence/utterance) and surface collocations (elements within a certain span, regardless of syntactic relationship or sentence-utterance boundaries) and argues for the superiority of the first of these. Moreover, as noted in Section 2.3.2, many recent frequency-based studies focus on particular types of collocation, such as verb + noun or adjective + noun collocations, and so show a concern for structural integrity.

2.3.6 Language variety

An important issue which has received less attention is which variety of English should serve as the basis for identifying collocations: in other words, whose collocations count?

There has been much criticism in applied linguistics of the notion that L1 users are or should be the model for second language learners. Widdowson (2000, 2003) complains that corpus linguistics in particular has promoted the idea of the "native-speaker model" and the implication that learners are somehow rehearsing to be L1 users: thus "it is no wonder that a common feeling among English learners is a sense of inadequacy in failing to measure up to native-speaker norms" (2003, p. 114). V. Cook (1999) similarly criticises the deficit view that sees their vocabulary usage and

other features of their language "treated as signs of L2 users' failure to become native speakers, not of their accomplishments in learning to use the L2" (p. 194-195). G. Cook (1998) mentions collocations in particular in asking "Why should the attested language use of a native-speaker community be a model for learners of English as an international language? If a certain collocation occurs frequently among British or American English speakers, must it also be used by the Japanese or the Mexicans?" (p. 60). Also relevant is Grosjean's (1989) argument that "the bilingual is NOT the sum of two complete or incomplete monolinguals; rather, he or she has a unique and specific linguistic configuration" (p. 6). This is not to suggest that bilinguals' L2 knowledge is inadequate, lacking or impaired by L1 interference; the point is simply that their L2 knowledge exists alongside knowledge of their L1. This is not the case for monolingual L1 users, and consequently the "native-speaker model" is inappropriate with respect to L2 learning as, by definition, an L2 learner can never become a monolingual in the target language.

Despite these critiques and much interest in English as an International Language and English as a Lingua Franca (ELF), research on L2 collocation, for the most part, remains untroubled. L1-user corpora and L1-user performance are routinely used as the standard against which L2 learners/users are compared, with little comment or debate (though see Nesselhauf (2005) and Henriksen (2013) for some discussion). Research on L2 collocation needs, as a minimum, to give careful consideration to the selection of a reference point for judging learners' production of collocations.

A related issue is the influence of culture on collocation. The wider field of phraseology has given some attention, of a descriptive nature, to cultural influences on phraseology (see, for example, Ooi (2000) and various papers in Skandera (2007)) and this has provided examples of such influences (e.g. the collocation *weekend car* in Singaporean English, which refers to a car that can only be legally driven outside of peak commuter times (Ooi, 2000)). Yet it is not often recognised that this is of importance when learners' production of collocations is evaluated, since there may be a cultural element to learners' collocational production and a cultural element to judgements of collocational acceptability. Any mismatch between the cultural basis of the production and of the judgements is then problematic and will not provide a full picture of learners' knowledge of collocations.

2.3.7 Textual and mental views of collocation

The final area of debate is whether collocation should be viewed as a property of language or as a property of our linguistic knowledge. That is, language can be seen as an artefact, an object in the world, with collocation as one phenomenon observed in that object; or language can be seen as a form of knowledge in the brain, with collocation as part of that knowledge, which is glimpsed through external behaviour (e.g. responses in tests) and external texts (e.g. as gathered in corpora). Partington (1998), for example, contrasts textual approaches to collocation, in which "one item collocates with another if it appears somewhere near it in a given text" (p. 15), and associative or psychological approaches, which consider it "part of a native speaker's communicative competence . . . to know what are normal and what are unusual collocations in given circumstances" (p. 16). Mollin (2009) terms essentially the same contrast the "corpus linguistic view" of collocation and the "psycholinguistic view", while Gyllstad (2007) refers to "textual instantiations" and "abstractions", but argues that collocations are both at the same time since "it seems reasonable to assume that any textual instantiations stem originally from associative connections between words present in these language users' minds" (p. 20). Gyllstad points to Hoey's (2005) view that the statistical approach is a method for revealing the truly interesting thing, the abstractions: "collocation is a psycholinguistic phenomenon, the evidence for which can be found statistically in computer corpora" (Hoey, 2005, p. 5).

Ellis, Frey and Jalkanen (2009), however, caution against such views, and question the transfer of findings in terms of textual collocations to the psycholinguistic realm:

> "observations of textual corpora naturally provoked linguists to make inferences about language *users* and about cognitive processes of meaning, speech production and comprehension . . . [but] however appealing these statements, they go beyond the data. While there is no denying that texts have been produced by language users, and thus must somehow reflect their thinking, corpus analyses say nothing about the cognitive loci of sensitivity of language learners and fluent users to these patterns of co-occurrence" (p. 91).

Their point then is not that there is no link between textual instantiations and

mental abstractions, but that this link cannot be assumed; empirical evidence must be sought for it. This was in fact recognised in the early days of frequency-based research: Sinclair, Jones and Daley (1970) describe one of their concerns as "the relationship between the physical evidence of collocation and the psychological sensation of meaning" (p. 3). Recent research (see Section 2.4.1) has begun to explore the textual-mental link.

Partington (1998), additionally, makes a slightly different point about the relationship between the textual/statistical notion of collocation and the mental/psychological view, stating that while "their interrelation makes it possible for a text receiver to judge whether a particular collocation in a text is usual or unusual . . . it must be stressed that there is a distinction between the statistical information about 'the language as a whole' and the individual's psychological knowledge of what constitutes normal collocation" (p. 17). Indeed, language exists both in individuals (as an idiolect), as a result of their own unique experience of the language, and in a community of users (as a communal language). This has significant consequences for how we evaluate learners' production of collocations: while we can never replicate an individual's unique exposure to the language, the reference point against which they are judged should as closely as possible reflect the language they have been exposed to (see Section 7.2.4).

2.4 Towards defining collocation

As Sections 2.2 and 2.3 have shown, there are many unresolved issues around collocation, but a way forward can perhaps be identified by considering what motivates both frequency-oriented scholars and phraseologists. As Section 2.3.7 noted, collocation can be considered a textual phenomenon or a mental phenomenon, and among many in the frequency tradition, there is a belief that observable collocations and other formulaic sequences have psychological reality. Conklin and Schmitt (2012) note that this is a general assumption, and Ellis, Frey and Jalkanen (2009) highlight a number of comments from scholars that explicitly connect sequences found in texts with mental processes or states.

Among researchers following the phraseological approach, there is often a similar belief in the psycholinguistic reality of collocation, especially among those who are oriented towards second language learning. This is seen, for example, in

Nesselhauf's (2005) description, quoted earlier, of a collocation as "an abstract unit of language and its instantiations in text" (p. 25); in Benson, Benson and Ilson's (1986) assertion that collocations are "psychologically salient" (p. 253); in Gyllstad's (2007) statement, quoted previously, that "it seems reasonable to assume that any textual instantiations stem originally from associative connections between words present in these language users' minds" (p. 20); and in Howarth's (1998a) suggestion that "the significance of composites is regarded as psychological, their degree of restrictedness related to mental storage and processing" (p. 28). Indeed, Howarth (1998a) is critical of the frequency-based approach precisely because, in his view, it is not valid psycholinguistically: "The mental lexicon clearly holds more abstract entities than are identified by computational searches, and neither native speakers nor learners produce word combinations on the basis of their frequency and probability of co-occurrence" (p. 26).

Part of the argument about collocation is therefore about what is most important in psychological terms. Frequency-oriented scholars see the sheer frequency of recurrence of collocations as reflective of and influential on psychological reality: there is storage of collocations in some form in the mental lexicon. Phraseologists see opacity and restrictedness as important since they believe these phenomena have psychological consequences: they necessitate storage of collocations in some form in the mental lexicon. The key question thus is what studies that probe the mental lexicon show. If such studies provide evidence that frequency of occurrence has psychological consequences, the frequency-based approach to collocation is strengthened. Alternatively, or in addition, if there is evidence of psychological effects for opacity or restrictedness, the phraseological approach is bolstered. The following section reviews such studies.

2.4.1 Psycholinguistic studies of formulaic sequences and collocations

This review seeks to reveal what psycholinguistic studies suggest about the nature of collocation. Primary attention is given to studies of collocation, and especially those involving L2 learners/users, but related studies of formulaic language are also referred to. The studies included in this review that involve L2 learners/users focus on those at higher proficiency levels. This is not by choice: it simply reflects the fact that, perhaps for methodological reasons or due to practicality, there is a lack of

collocation processing studies involving learners at lower proficiency levels. The research does not, therefore, give insights into development, but does provide information on the advanced L2 lexicon.

Ellis, Frey and Jalkanen (2009) looked at the processing by L1 users of adverb + adjective collocations taken from Kennedy (2003) and verb + noun collocations from Kennedy (2005). The frequency of each collocation was found in the British National Corpus (BNC) and a set of non-collocations created by re-combining the collocations' component words. Using a lexical decision task, the authors found that higher frequency correlated with faster lexical decisions for both collocation types. This does not necessarily indicate that textual collocations are mental collocations (see Section 2.3.7), but does suggest that these two aspects of collocation overlap.

Wolter and Gyllstad (2013) made similar findings with advanced L2 learners. Defining collocations as adjective + noun combinations with a frequency of at least 10 in the COCA, collocations with a wide range of frequencies were identified. A set of non-collocations was created by randomly combining frequent adjectives and nouns and checking that each resulting combination did not occur in the COCA. L1 users and advanced Swedish-L1 learners performed an acceptability judgement task in which they were asked to determine "whether the word combinations are commonly used in English or not" (p. 460). While the study was also concerned with the effects of congruence (i.e. comparing collocations that have a direct L1 equivalent with collocations that do not), with respect to frequency, it was found that both groups of participants had faster response times for more frequent collocations.

The above two studies therefore show that collocations, defined by frequency, are processed more quickly by both L1 users and learners. Interestingly, Wolter and Gyllstad found that there was no effect on reaction times for the frequency of the component words in the collocations: that is, it was the frequency of the collocation itself which affected processing.

These findings are supported by studies of the processing of compositional phrases (examples from several studies are given below). These studies, themselves a reaction to a longer tradition of L1 idiom processing studies, have in recent years begun exploring whether compositional phrases may show signs of holistic storage and processing. Bannard and Matthews (2008) found that young children are sensitive to the frequency of compositional phrases. Using a sentence-repetition task,

it was found that two- and three-year-old children were significantly more likely to repeat four-word compositional sequences identified in a corpus of child-directed speech (e.g. *sit in your chair*) versus matched controls (e.g. *sit in your truck*). Tremblay, Derwing, Libben and Westbury (2011) found significant advantages for lexical bundles over matched non-bundles in adult L1 users. The study featured syntactically and semantically regular four- and five-word lexical bundles (e.g. *in the middle of the*), and, using five separate self-paced reading and sentence-recall tasks, significant advantages for the bundles were found in each task. Arnon and Snider (2010) found continuous sensitivity to the frequency of compositional sequences (e.g. *don't have to worry*) in adult L1 users. Their study used lexical decision tasks to explore the processing of four-word strings, like Tremblay et al., but the strings were selected from across the frequency spectrum, and both word- and substring-frequency were controlled for. Two experiments and a meta-analysis using all the data showed sensitivity to frequency extending across the frequency range (i.e. there was not a threshold at which frequency began to have an effect; rather, there were larger and smaller effects in line with greater or lesser frequency).

Hernández, Costa and Arnon (2016) replicated Arnon and Snider (2010) with two groups of L2 learners/users of high-intermediate proficiency or above and a group of L1 users. The study found that, after controlling for word- and substring-frequency, all three groups were sensitive to the frequency of multi-word strings, with this sensitivity extending across the frequency spectrum. Further, the frequency effects were of similar magnitude for the three participant groups.

These studies confirm the findings above that it is the frequency of the collocation itself, not of the component words, which is key. They also show that frequency has a continuous effect, rather than there being a threshold for frequency effects. These studies thus challenge the established idea of a divide between stored idiomatic language and computed compositional language in that they imply there must be some kind of storage of compositional sequences, with Hernández, Costa and Arnon providing evidence that L2 learners/users also can develop sensitivity to such sequences.

Ellis, Simpson-Vlach and Maynard (2008), in contrast, showed that there may be differences between L1 users and L2 learners/users in which aspect of frequency is most important. They looked at the processing of three-, four- and five-word

sequences. All the sequences passed minimum frequency and MI-score thresholds (MI being a measure of strength of association between words; see Section 2.2.2), and were then further assigned to three frequency bands and three MI-score bands (thus nine sub-sets of sequences). Three experiments explored the processing of the sequences. A grammaticality judgement task found that L1 users responded faster to shorter sequences and sequences with higher MI scores, while for the L2 users there were effects likewise for shorter sequences and sequences of higher frequency, but not for MI score. In experiment two, participants read aloud sequences as fast as possible, with voice onset time as the key measurement. The number of words, number of phonemes, MI scores, but not frequency had effects for L1 users. For L2 users, the number of phonemes and frequency, but not MI score had an effect. Experiment three looked at the priming of the final word in a sequence by the initial words in the sequence. For L1 users, the number of phonemes and MI scores, but not frequency had effects; for L2 users, none of the variables examined had a significant effect.

It appears, therefore, that L1 users are sensitive to how often formulae occur in comparison with the occurrence of the component words in other contexts, while L2 users are sensitive simply to how often formulae occur. The authors argue these findings suggest that for the L2 users acquisition of the formulae was ongoing and so the effects of frequency were evident, but for the L1-user participants, massive amounts of exposure to English meant that all of the formulae had been experienced sufficiently for learning to reach asymptote. The L1 users were sensitive to MI, however, because this reflects co-occurrence greater than chance, which is a marker of cohesiveness. With regard to the above findings, it should be noted that all the formulae in the study were reasonably frequent and had reasonably high MI scores. It is therefore possible that, had low-frequency and low-MI sequences also been included, L1-user sensitivity to frequency and L2-user sensitivity to MI scores may have been found. The study's core finding is, nonetheless, interesting given other indications of differences between L1 users and learners with respect to MI-based and frequency-based collocations (see Sections 3.2.1 and 3.2.2).

Sonbul (2015) provides a more detailed picture of the effects of collocation frequency on processing. This study was innovative in that it looked solely at the processing of semantically related adjective + noun combinations. That is, while

most studies use incongruous combinations as controls, in this study all the combinations were semantically plausible: e.g. *official recognition* (a high-frequency collocation) was compared with *proper recognition* (low frequency) and *standard recognition* (a non-collocation, though congruous semantically). L1 users and advanced learners of English from a variety of L1 backgrounds completed a sentence-reading task with eye-tracking. For both groups, there were effects for collocation frequency on early processing (as seen in first pass reading time), but not for later processing (as seen in total reading time and fixation count). There was also no effect for L2 proficiency (estimated using the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001)), though this may be due to a lack of variance among the participants given that all were advanced learners.

This study confirms the finding described above that collocation frequency has a continuous effect on processing. However, it also gives more detail on this effect in the finding of effects on early, but not later processing. Sonbul explains this finding by suggesting that since collocations are not fixed expressions, language users are able to deal quite easily with an unusual combination. Thus, while collocations have an advantage in initial processing, there is no difference at later stages in which the meaning of the combination is integrated with the text as a whole up to that point. This implies, nonetheless, that collocations have some status in the lexicon.

Durrant and Doherty (2010) explored another factor that may have psycholinguistic effects: word association. They conducted two experiments with several types of word combination: low-frequency combinations (e.g. *direct danger*); mid-frequency collocations (e.g. *greater concern*); high-frequency collocations (e.g. *foreign debt*); and high-frequency collocations which were also word associations as judged by consultation of the *Edinburgh Associative Thesaurus* (EAT) (Kiss, Armstrong, Milroy, & Piper, 1973) and confirmed by eliciting word associations from respondents from the same pool as the participants in the main experiments (e.g. *estate agent*). In addition, incongruous combinations, consisting of the nouns from the above conditions paired with adjectives they did not co-occur with in the BNC (e.g. *armed concept*), were used as controls. In the first experiment, which used a lexical decision priming task, L1-user participants responded significantly more quickly to both the high-frequency collocations and the associated high-frequency collocations over the controls. However, due to concern that participants may have

used strategic processes (i.e. having noticed relationships between some of the primes and targets, some participants may have attempted to guess possible targets on viewing each prime), the second experiment used masked priming, in which the prime is shown so briefly that participants are not consciously aware of having viewed it. This experiment found significantly faster processing only for associated high-frequency collocations.

This study, then, appears to show that high-frequency collocations which are not regular word associations do not effect automatic processing. Durrant and Doherty are, however, cautious about this finding since they report signs of facilitation in the other conditions, and the effect sizes are similar across all four conditions. It may therefore be that while psychological association has an additional effect, frequency does affect processing.

Gyllstad and Wolter (2016) looked at the processing of collocations versus free combinations, as described by Howarth's (1998a) phraseological continuum (described in Section 2.2.1). The collocations (e.g. *run a risk*) included a noun in its literal sense and a verb in a specialised sense; in the free combinations (e.g. *write a letter*) both elements were in their literal senses. The two types of items were matched for verb frequency, noun frequency, phrasal frequency, lemmatised phrasal frequency and length. The expectation was that the collocation items would be processed more slowly by learners of English than matched free combinations due to their lower degree of transparency. Using a semantic judgement task, in which participants had to judge as quickly as possible whether an item was "meaningful and natural" in English, with L1 users and highly proficient Swedish learners of English as participants, a significant advantage was found for free combinations over collocations in terms of both reaction times and error rates, with no differences between the two participant groups. In addition, effects were found for frequency independent of the item type (i.e. in the case of both free combinations and collocations, greater frequency led to faster reaction times).

Support for Gyllstad and Wolter's main finding is seen in Siyanova-Chanturia, Conklin and Schmitt (2011), in which the authors used eye-tracking as participants read short stories containing idioms used figuratively (e.g. *at the end of the day* = eventually), idioms used literally (e.g. *at the end of the day* = in the evening) or a matched novel phrase (e.g. *at the end of the war*). L1-user participants showed a

processing advantage for the idioms (in both figurative and literal uses) over novel strings, but there was no difference between figurative and literal readings. L2 users, however, processed novel strings and idioms used literally at a similar speed, but had significantly slower processing for idioms used figuratively versus idioms used literally. This was despite the idioms being selected on the basis of high familiarity with a separate but similar group of L2 users, and their familiarity being confirmed with the experimental group after the eye-tracking task.

Gyllstad and Wolter's study supports the idea that collocations are distinguished from free combinations psycholinguistically. It should be noted, however, that the authors do not appear to have controlled (either in item selection or statistically) for MI score, which Ellis, Simpson-Vlach and Maynard (2008) found to affect processing, nor for psychological association, which Durrant and Doherty (2010) showed to be important. It is not therefore clear how, or whether, these three factors interact.

2.4.2 Lessons from psycholinguistic studies

The above review of psycholinguistic studies of the processing of formulaic sequences and collocations has explored the status of collocations in the mental lexicon and has pinpointed factors that seem to give collocations this status. The review shows first that there seems to be a strong effect of frequency on processing (Ellis, et al., 2009; Gyllstad & Wolter, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013). Frequency effects have been found across a large number of linguistic variables (see Ellis (2002) who discusses frequency effects in phonology and phonotactics, reading and spelling, lexis, morphosyntax, formulaic language, language comprehension, judgements of grammaticality and syntax), and the above studies demonstrate that this includes various types of formulaic language and collocations also.

There are four notable aspects of the findings with regard to frequency. First, there are frequency effects on both opaque language (Gyllstad & Wolter, 2016) and on compositional language (Arnon & Snider, 2010; Gyllstad & Wolter, 2016; Hernández, et al., 2016). Second, it is the frequency of the collocation/sequence itself, rather than the frequency of the component words, that is key (Hernández, et al., 2016; Wolter & Gyllstad, 2013). Third, different measures of frequency (i.e. MI

scores vs. raw frequency) correlate with L1-user and L2-learner/user processing respectively (Ellis, et al., 2008). Fourth, frequency has a continuous effect on processing, rather than there being a certain threshold of frequency over which effects are seen (Ellis, et al., 2008; Gyllstad & Wolter, 2016; Hernández, et al., 2016; Sonbul, 2015; Wolter & Gyllstad, 2013).

In addition, the studies give indications of other factors affecting processing. Specifically, whether combinations have been found to occur as word associations or not appears to affect processing over and above the effects of frequency (Durrant & Doherty, 2010), and there appear to be differences in the processing of phraseologically defined collocations versus free combinations (Gyllstad & Wolter, 2016).

As noted previously, one feature of the research is that the L2 learners/users involved are almost exclusively at higher proficiency levels. In some cases differences were found between L2 learners/users and L1 users: for example, Ellis, Simpson-Vlach and Maynard (2008) found that L2 users differed in response times to collocations distinguished by frequency while L1-user response times were affected by MI scores. However, it is notable that across the studies, apart from somewhat slower reaction times, there are in general few differences between L2 learners/users and L1 users. Similar patterns of results for both L1 and L2 participants were found by Hernández, Costa and Arnon (2016), Wolter and Gyllstad (2013), Sonbul (2015) and Gyllstad and Wolter (2016). It seems then that the psychological reality of collocation, whatever that may be, can ultimately be developed by L2 learners/users in largely the same way as for L1 users. This in turn means that a single definition of collocation can be applied for work with both L1 and L2 participants.

2.5 Defining collocation

Based on the above, two conclusions may be reached regarding how collocation should be defined. First, it seems that definitions of collocation that allow either for frequency or for opacity/restrictedness to be the defining feature may be most appropriate because there is evidence that these are the characteristics that impact processing, which in turn is assumed to reflect psychological reality. Second, it seems reasonable to make use of the same definition of collocation for both L1 and

L2 focused research, which is a helpful conclusion to be able to draw, given that research involving L1 and L2 learner comparisons should be based on the same design parameters.

Examining definitions in the literature, a number of scholars have attempted to combine phraseological and frequency-oriented criteria. Laufer and Waldman (2011) consider collocations to be "habitually occurring lexical combinations that are characterized by restricted co-occurrence of elements and relative transparency of meaning" (p. 658). Handl and Graf (2010) characterise word combinations on three levels. The first is predictability: "if a word has many potential partners it has a tendency towards a free word combination, if the collocational range is restricted . . . it is similar to an idiom" (p. 129). The second is frequency: "very rare combinations are not considered good candidates for a collocation, nor are very frequent ones" (p. 129). The third is idiomaticity: "both completely opaque combinations and transparent combinations are excluded from the collocational area" (p. 129). For Schmid (2003) collocations, prototypically, are "combinations of lexemes exhibiting a medium degree of observable recurrence, mutual expectancy and idiomaticity" (p. 249).

Various features of these definitions may be discussed, but for current purposes the key point is that each of the definitions includes an idea of frequency ("habitually occurring", "frequency", "observable recurrence") as well as phraseological notions of restriction ("restricted co-occurrence", "predictability", "mutual expectancy") and opacity ("relative transparency", "idiomaticity"). All three definitions exhibit, however, one above-noted problem (Section 2.2.1) in that they include multiple criteria but do not clarify how the criteria interact or which has primacy.

A more successful combination of phraseological and frequency-based criteria is Durrant's (2014) definition. This is based on Palmer (1933), and describes collocation as:

> "combinations of two words that are best learned as integral wholes or independent entities, rather than by the process of placing together their component parts, either because (i) they may not be understood or appropriately produced without specific knowledge, or (ii) they occur with sufficient frequency that their independent learning will facilitate fluency" (Durrant, 2014, p. 448).

This definition then not only combines elements of the two approaches, but explicitly states that either element may be sufficient to deem a combination a collocation. In addition, it is oriented towards learning and so is not merely about the textual identification of collocations, but shows concern for the ultimate motivation for many researchers' interest in collocation. As this definition of collocation is broadly in tune with the psycholinguistic evidence and is clear regarding the interaction of the two criteria, it is therefore endorsed and adopted in this thesis.

Having explored issues surrounding the definition of collocation, the next chapter will look at research into L2 learners' knowledge of collocations, considering what this research has revealed and the methodological issues involved. It must be borne in mind that much of this research does not define collocation as above, and indeed definitions of collocation vary across the studies. This means caution is necessary when comparing studies and reaching conclusions on their findings.

Chapter 3 L2 learners' productive knowledge of collocations

3.1 Introduction

This chapter examines studies of L2 learners' productive knowledge of collocations in order to answer two questions: (1) What has been discovered about learners' productive knowledge of collocations? (2) How can learners' productive knowledge of collocations be investigated? Such studies can be divided into two broad types: learner corpus studies and elicitation studies. Learner corpus studies tabulate the presence or absence of collocations and evaluate the use of collocations in corpora of learner language. Elicitation studies use instruments specifically designed to elicit learners' knowledge of particular collocations, for example, through gap-fill or translation tasks.

In accordance with the first question above, Section 3.2 reviews and evaluates a number of learner corpus studies (Section 3.2.1) and elicitation studies (Section 3.2.2). Section 3.2.3 brings together the findings of these studies to determine what is currently known about L2 learners' productive knowledge of collocations. Then, in order to consider how learners' productive knowledge of collocations can be investigated, Section 3.3 describes the problems and challenges for research in this area and evaluates the methods used in the studies reviewed. Section 3.4 concludes that a method is needed that avoids certain pitfalls diagnosed in Section 3.3, and one research instrument is identified and adopted as the starting place for empirical investigations.

3.2 Studies of L2 learners' productive knowledge of collocations

3.2.1 Learner corpus studies

Studies based on corpora of language produced by learners have investigated learners' deployment and use of collocations. The following discussion evaluates four major representative studies of this type. As stated previously, the aims of these reviews are twofold: to establish what is known about L2 learners' productive knowledge of collocations and to evaluate the methods used in these investigations.

Howarth (1998b) compared a corpus of advanced learner writing with a corpus of L1-user writing. The former included essays from 10 Master's degree students

from a variety of L1 backgrounds and totalled 25,000 words. The latter consisted of texts from the Lancaster-Oslo-Bergen corpus and academic writing by Leeds University staff totalling 238,000 words. Howarth looked at verb + noun complement combinations, analysing verbs that occurred frequently in the L1-user corpus and all verbs in the learner corpus. Adopting a phraseological approach to collocation, Howarth classified each combination as either a free combination, a restricted collocation or an idiom, primarily on the basis of the substitutability of the component words (see Section 2.2.1 for details of Howarth's phraseological continuum). In the learner corpus, 69% of the combinations were free combinations, 24% restricted collocations and 1% idioms (with 6% classified as errors). In the L1-user corpus, 62% were free combinations, 33% restricted collocations and 5% idioms. In addition, among the 10 learners, Howarth found no correlation between the proportion of collocations used and two measures of proficiency (a language test taken on admission to the university, the learners' academic grade at the end of the course). Howarth submits that the lower proportion of collocation use among the learners suggests either less knowledge of collocations or problems in deploying their knowledge of collocations.

Three problems can be identified in Howarth's study. The first is the comparability of the corpora, in that the L1-user corpus consisted of published texts by, in Howarth's own words, "mature, fully competent writers" (p. 166), whereas the learner texts were written by Master's degree students. More recent studies have suggested that writer expertise affects formulaicity, something Howarth does not seem to have considered. For instance, O'Donnell, Römer and Ellis (2013) looked at formulaic language in 15 corpora distinguished by level of expertise (undergraduate, postgraduate and expert) and by background (L1 user and L2 user), and found differences across the levels of expertise on three out of four measures of formulaicity (see also Neff van Aertselaer (2008) who made a similar finding). Howarth's results may therefore reflect differences in expertise between the writers, rather than (or in addition to) their status as L1 users or L2 learners/users.

A second problem is the reliability of the classification of combinations as free combinations, restricted collocations or idioms. Howarth reported that distinguishing between the first two in particular was difficult (see Section 2.2.1 above for why this may be so) and yet there is no data on intra-rater reliability, nor, it seems, was a

second rater used, so there can be no calculation of inter-rater reliability.

Third, the study does not fully consider variability within each corpus. Howarth reported that the L1-user corpus was treated as a single body of writing "based on an assumption that the forty or more authors of those texts were all mature, fully competent writers, and that shared phraseological norms could be identified from their overall combined performance" (p. 166). The learners' texts, in contrast, were analysed individually, since "it was necessary to take into account the considerable variability between learners in their use of conventional language forms" (p. 166). The results, however, concentrate on the overall learner corpus figures, means and standard deviations are not presented, and the only information on variability given is the range among the learners for each category. Interestingly, these ranges (58-74% for free combinations, 16-33% for restricted collocations and 0-4% for idioms) indicate that at least some of the learners had percentages similar to the overall L1-user results. It is therefore possible that there is quite some degree of overlap between individual learners and L1 users, and it is hard to be certain whether there is a genuine difference between the learners and the L1 users, especially considering the earlier point about the effects of writer expertise. Howarth's study therefore highlights three challenges faced by learner corpus research: obtaining comparable corpora, reliably identifying word combinations as collocations and dealing with variability within corpora.

Nesselhauf's (2005) learner corpus was considerably larger than Howarth's, totalling 154,191 words and consisting of 318 essays by 207 German L1 university-level learners of English. This means the language used by any one individual has only a small impact on the overall results. The study also avoided one problem with Howarth's work by not seeking to compare this corpus with an L1-user corpus; instead, the study focused on evaluating collocations in the corpus and identifying types and possible sources of errors. Taking a phraseological approach, 2,082 instances were identified in the corpus where either a collocation-like combination occurred or a collocation should have occurred. "Acceptable" collocations among the 2,082 were identified by reference to L1-user norms. First, four dictionaries and the BNC were used to identify "acceptable" collocations. Then, the remaining collocations were presented to two L1-user judges. If these judges disagreed, two further judges were consulted. Of the 2,082 instances, 748 (36%) were judged

questionable or unacceptable, which Nesselhauf sees as demonstrating the difficulty of collocation even for relatively advanced learners. Examining these problematic instances, in almost half of the cases the problem was with the verb (e.g. *<u>*make*</u> *an experience*), in around one in five cases the noun [1] (e.g. **give the children a* <u>*possibility*</u> *to play*), and in one third other elements such as determiners and noun complementation (e.g. **put* <u>*an*</u> *enormous pressure on*). Approximately 70% of the problematic instances were considered to display possible L2 influence (i.e. confusion among near synonyms or phonologically similar words) and 50% showed possible L1 transfer. Thus, a considerable number were analysed as showing both L2 and L1 influence.

In addition, Nesselhauf looked at other factors that may cause collocations to be challenging for learners. First examined were the proportions of less restricted and more restricted collocations that were used acceptably: the former being acceptable 63% of the time, the latter 75% of the time. Thus, more restricted collocations appear to be easier for these learners. Second, Nesselhauf looked at the use of free combinations and collocations: 76% of the free combinations were used acceptably, and 60% of the collocations. This "indicates that collocations are more often deviant than free combinations, but by no means radically so" (p. 208). A third analysis looked at congruence (i.e. whether the English collocation has a word-for-word translation equivalent in the L1, German), finding that 73% of the congruent collocations were used acceptably, and 50% of the non-congruent collocations. This suggests that congruence is important, but not dominant: many congruent collocations were problematic and many non-congruent collocations were unproblematic. Descriptions of non-congruent collocations as the heart of learners' problems with collocation (e.g. Bahns, 1993) do not then seem appropriate. Concluding her study, Nesselhauf argues that learners do make use of many collocations, but seem to operate word-by-word, rather than chunk-by-chunk, with a lexicon of words which have links between them.

Nesselhauf's study is extensive, impressively detailed and has been widely cited.

[1] As will be explained in Section 4.5.5, Nesselhauf argues that in verb + noun collocations the noun is first selected based on its meaning and the verb then selected in accordance with the noun. It may therefore be asked how there can be problems with the noun in a collocation: If the noun is selected on the basis of faulty knowledge of its meaning, the problem would seem to be with knowledge of the noun rather than with knowledge of the collocation. Nesselhauf does not seem to address this point.

There is nevertheless a clear problem with the study, which was also seen in Howarth's study: the reliability of the categorisation of combinations as "acceptable" or problematic. Multiple raters were employed, but there was no calculation of inter-rater reliability, nor even a basic figure for agreement among the judges. The process was also highly complex, with multiple possible ways in which a combination could eventually be assigned to a particular classification.

Durrant and Schmitt's (2009) study featured four corpora of learner writing, produced by learners from various L1 backgrounds, and four corpora of L1 writing, with each corpus consisting of 12 texts. The study followed the frequency tradition and first identified all adjective + noun and noun + noun combinations in the 96 texts: a total of 10,839 combinations. The frequency, *t*-score and MI score for each combination was found in the BNC. The study then examined the proportion of combinations with different levels of *t*-score and MI score in the various corpora. The main findings were: (1) L1 writers use more low-frequency collocations than learners, (2) learners use at least as many or overuse collocations with high *t*-scores in comparison with L1 writers, (3) learners underuse collocations with high MI scores compared with L1 writers. The authors therefore conclude that learners acquire a lot of highly frequent collocations, but high MI collocations, since they are generally of lower frequency, appear to take longer to acquire.

Two features of Durrant and Schmitt's study are of note. First, the authors looked at the proportion of combinations with different levels of *t*-score and MI score in the corpora: that is, the percentage of the total combinations within eight different bands of *t*-score and MI score. This gives a more nuanced picture of the type of collocations used in the various corpora and contrasts with many other studies that have adopted a simple threshold for designating collocation. Siyanova and Schmitt (2008), for example, conducted a rather similar study, but used only a threshold for MI and *t*-scores and, perhaps as a consequence, did not reveal any differences between the L1-user and learner corpora examined. The second notable feature is that figures were calculated for individual texts, rather than for each corpus as a whole. This allowed means and standard deviations to be calculated, in turn allowing inferential statistics to be used to compare the corpora.

Granger and Bestgen (2014) conducted a follow-up to Durrant and Schmitt's study which provides further support for the results. They used two corpora of

learner writing of a similar length and type, distinguished by proficiency level (intermediate and advanced). Applying Durrant and Schmitt's method, they found that the intermediate essays contained a smaller proportion of high MI collocations and a larger proportion of high *t*-score collocations than the advanced essays (see also Bestgen and Granger (2014), which extended the methodology by looking at the mean *t*-score and mean MI score for each pair of words in a text). These findings, in combination with Durrant and Schmitt's, suggest collocational development may involve highly frequent collocations (which *t*-scores tend to identify) gradually becoming less important, with less frequent but highly associated collocations (identified by MI scores) becoming more important. This pattern of results accords with Ellis, Simpson-Vlach and Maynard (2008), who, as reported in Section 2.4.1, found that L1 users were sensitive in processing tasks to lexical bundles differentiated by MI score, while L2 users were sensitive to bundles differentiated by frequency. Ellis et al. attributed their results to the effects of absolute frequency and relative frequency on learning, and it may be that progression towards the type of units identified by MI scores is a process which takes place over a considerable period of development and requires enormous amounts of exposure to the language.

There are, however, two issues with Durrant and Schmitt's study. The first is the comparability of the corpora. While the study attempted to match learner and L1 corpora, the authors admit that though the corpora resembled each other in a number of ways, there were nonetheless differences between them in terms of the conditions in which the texts were written and the audience the texts were written for. It is not entirely clear whether action was taken to evaluate the comparability of the corpora as a first step in the analysis, and it is unfortunate that individual results for each corpus for the proportion of combinations in each *t*-score and MI score range are not given. The one exception is that the average number of adjective + noun and noun + noun combinations per text is given for each corpus, and this reveals quite considerable differences within the learner corpora and within the L1 corpora. It is therefore possible that there were differences between the corpora besides the L1-user versus learner difference the study focuses on, and that any such differences may have confounded the results highlighted in the study.

The second issue, noted by Gablasova, Brezina and McEnery (2017), is that as MI highlights infrequent combinations, the constituent words in these combinations

may themselves be infrequent. Thus, it may not be that learners knew the constituents but did not know the collocation, they may simply have not known the constituents. In other words, learners' limited use of high MI collocations may have been due to lack of word knowledge, not collocation knowledge: the assessment of learners' knowledge of collocations was confounded by the issue of knowledge of words. This issue certainly needs to be considered. However, it should be pointed out that the tendency of MI to highlight infrequent combinations is well known and it is recommended (Clear, 1993; Manning & Schütze, 1999) that this problem be mitigated by only calculating MI scores for combinations that reach a minimum co-occurrence frequency. This recommendation was indeed followed by Durrant and Schmitt.

Laufer and Waldman (2011) used a corpus of 759 essays by Israeli school and university learners of English totalling 291,049 words. The corpus was sub-divided into three levels according to the school year of the learners. This corpus was compared with a 324,304-word corpus of writing by UK school and university students and US university students. The researchers began by finding frequently occurring nouns in the L1-user corpus, then identifying verbs which combined with each of these nouns. These verb + noun combinations were considered collocations if they appeared in either of two collocations dictionaries (the *BBI Combinatory Dictionary of English* (Benson, Benson, & Ilson, 1997) and the *LTP Dictionary of Selected Collocations* (Hill & Morgan, 1997)). In the learner corpus, the frequently occurring nouns from the L1-user corpus were searched for, verbs combining with each noun again identified, and the combinations judged as collocations in the same way. Laufer and Waldman had an L1 user identify "errors" in the learner verb + noun combinations, which were then reviewed by two L1 Hebrew speakers for L1 influence. The study found that: (1) Verb + noun collocations accounted for 10.2% of the total noun occurrences in the L1-user corpus and 5.9% of the noun occurrences in the learner corpus; a significant difference. (2) Among the three learner groups, the higher group produced significantly more collocations than the lower group. (3) For all three learner groups, around a third of the collocations produced were deemed deviant (i.e. the intended combination (as determined from the context of the essay) should have been a collocation, but in the view of the L1-user judge one of its components was incorrect). (4) There were no differences

between the three groups in the number of deviant collocations that exhibited L1 influence.

Three problems can be seen in Laufer and Waldman's study. First, the correspondence between the study's definition of collocation and its operationalisation is rather questionable. Collocation was defined as "habitually occurring lexical combinations that are characterized by restricted co-occurrence of elements and relative transparency of meaning" (p. 658), and this was operationalised by referring to two dictionaries. Yet there is no attempt to demonstrate that these dictionaries are based on the same or a similar conceptualisation of collocation. Both dictionaries follow the phraseological tradition, but both have been criticised by Nesselhauf (2005) for including free combinations, while the introduction to the LTP dictionary emphasises that it is a selection of collocations and does not aim to be comprehensive.

A second problem is the categorisation of combinations as deviant, and, for those categorised as such, as influenced by the L1. For the first classification, there was a single rater, while for the second there were two raters and inter-rater reliability is stated. It is not clear why one rater was deemed sufficient for one categorisation task, but not for the other. A further issue is that the second categorisation task featured two raters looking for signs of L1-Hebrew influence. Yet the study states that some of the learner essays (it is not stated how many) were by L1-Arabic speakers. The combinations from these essays should surely have been excluded from this analysis.

A final problem is that the starting point for the investigation was nouns which were frequent in the L1-user corpus. This in effect biased the study: there is an internal criterion of sorts for the L1 users, but an external criterion for the learners. It is not clear why nouns frequent in the learner corpus were not used in the same way, nor why an external list of frequent nouns was not used. It is particularly odd in that the authors state that their interest lies in "how learners use the collocations of the most frequent nouns that are familiar to them" (Laufer & Waldman, 2011, p. 659). Yet there was no attempt to establish to what extent these nouns were in fact familiar to the learners. Clearly, many of the nouns were known to many of the learners, since they did occur in the learner corpus, but it may be noted that after standardizing for the length of each corpus, the learners' use of the nouns was around 20% below

that of the L1 users. Also notable is that while most of the nouns are highly ranked in word frequency lists, some are not, and it would be interesting to know how much collocation use there was within each corpus for nouns at different frequency levels. In sum, the learners' lower level of use of collocations may in part be because of limited knowledge of these nouns.

As mentioned previously, the findings of the studies above are pulled together and reviewed in Section 3.2.3, and the main issues with the research identified above are discussed in Section 3.3.1.

3.2.2 Elicitation studies

Having reviewed and evaluated a number of learner corpus studies, this section looks at several elicitation studies likewise. Elicitation studies use an instrument, of various designs, to probe learners' knowledge of collocations. These studies have investigated the extent of learners' productive knowledge of collocations and in some cases have attempted to explore how this knowledge may develop.

Two early elicitation studies have been widely cited, but suffer from a number of problems. Bahns and Eldaw (1993) used two measures of productive collocation knowledge: a sentence-translation task and a cloze task. Unfortunately, the former was unconstrained and responses did not necessarily need to include the target collocations. In addition, there is very little statistical information given about the two tests, so their reliability cannot be confirmed; the basis for the selection of items is unclear; only overall figures are provided for the tests rather than the participants' mean scores; the study does not consider confounding factors that could have affected responses, such as the syntactic complexity of the sentences; and it compares knowledge of lexical words and knowledge of collocations, but with no control for the frequency of either. Bonk (2000) used a 50-item collocations test consisting of three sub-tests, only the first of which, a cloze test featuring verb + noun collocations, was a test of productive knowledge of collocations. Unfortunately, while Bonk carried out a variety of analyses of his collocations test, the focus is on the overall test, rather than this sub-test, and so little further information on it is available.

More recent elicitation studies have begun to overcome some of the problems seen in earlier studies. Revier (2009) designed an instrument which consists of a

sentence context with a gap for the whole collocation, next to which is a three-by-three matrix containing three verbs, three determiners and three nouns. Participants select one word of each type to produce the collocation. The test has 45 items, 15 each in three semantic categories: transparent (verb and noun both in core or literal sense); semi-transparent (verb in extended sense, noun in core or literal sense); and non-transparent (neither verb nor noun in core or literal sense). The collocations each include one of 15 highly polysemous verbs, and both the noun constituent and the collocation as a whole are frequent in the BNC.

Revier's study included 56 Danish L1 participants: 10th and 11th grade schoolchildren and first-year university students. Overall mean scores (out of 45) were 22.5 (*SD* = 7.7), scores appear to have been normally distributed, and reliability was good at .89. The university students had significantly higher scores than the 10th and 11th grade schoolchildren, with no difference between these latter two. There were also significant differences between the scores for the three item types: transparent items being significantly easier than semi-transparent, which in turn were significantly easier than non-transparent. Looking at these differences within each year group, there is a picture of emerging competence with each type of collocation.

Revier's study provides some interesting data, and, with 45 items, clear criteria for the selection of items and good reliability, the instrument is superior to those used in the early studies. The primary problem, as Revier says himself, is that despite setting out to measure learners' productive knowledge of collocations, his instrument could be described as a receptive test as much as a productive test. Nevertheless, the study makes two useful findings: knowledge of collocations appears to correlate with proficiency to some extent and a greater degree of opacity appears to correlate with difficulty.

Barfield (2009a) introduced an instrument named *LexCombi*, which was based on the *Lex30* (Meara & Fitzpatrick, 2000) test format, a test of productive vocabulary size. *Lex30* presents 30 stimulus words to the test-taker and asks for three associations for each, with a 30-second limit per item. *LexCombi* likewise presents 30 stimuli and gives a 30-second limit per item, but asks for three collocations. *LexCombi*'s cues are all frequent nouns and were selected after piloting found they distinguished well between L1-user responses and responses from advanced Japanese users of English. Barfield scored responses against a list of "appropriate"

collocations for each item, created using the *Oxford Collocations Dictionary* (1st edition) (Crowther, Dignen, & Lea, 2002) and *Collins Wordbanks Online* (HarperCollins, 2004).

LexCombi was trialled with 89 Japanese university students of low-intermediate to advanced proficiency. The mean number of "appropriate" responses (out of 90; three responses × 30 items) was 37.93 (SD = 9.74), the scores were normally distributed and reliability was .78. A Rasch analysis showed that the 30 items were well spread in terms of their difficulty. *LexCombi* scores had a correlation (Pearson's) of .57, $p < .001$, with the participants' TOEIC scores. There were clear differences in *LexCombi* scores between higher and lower proficiency participants, particularly in the number of collocations of different word classes. This led Barfield to propose that knowledge of collocations may develop first in terms of adjective + noun connections, before later verbal collocations become a major area of development.

Barfield's *LexCombi* has the advantage of being highly efficient, in that 90 collocations may be elicited from learners in just 15 minutes. It also stands out from other elicitation studies in featuring neither cloze nor translation tasks, but being more open. *LexCombi* is something of a hybrid in that it allows learners to produce language somewhat freely, but within a controlled environment. It may therefore overcome the narrowness of some instruments. One issue with it, however, is the scoring of responses. The lists of "appropriate" collocations were developed from sources that seem to have been selected mainly since they were conveniently available, and it may be asked whether these sources are an appropriate means of judging learners' responses.

González Fernández and Schmitt (2015) used a productive form-recall test. Each item had a sentence in the L1 (Spanish) and an English equivalent with two blanks for the collocation, with the first letter of each word provided. The test included 50 collocations, of a range of types, intentionally selected so as to have a wide range of frequency, *t*-scores and MI scores in the COCA. All collocations were incongruent with Spanish (i.e. they did not have a word-for-word Spanish translation equivalent), and constituent words were from the most frequent 5,000 words of English.

The test was given to 108 Spanish L1 participants. The mean score (out of 50) was 28.29 (*SD* = 9.74) and scores were normally distributed (test reliability was not

reported). The item scores had correlations (Spearman's) of .45 ($p < .01$) with corpus frequency, .41 ($p < .01$) with *t*-scores and -.16 ($p > .01$) with MI scores. Participants also completed a language background and use questionnaire, and collocations test scores had significant correlations with self-rated proficiency, years of English study and a measure of exposure to English. The study concludes that "L2 learners typically know a substantial number of collocations, providing some evidence to counteract the notion that collocations are too hard for learners" (González Fernández & Schmitt, 2015, p. 114).

González Fernández and Schmitt's study is unusual in that it involved multiple types of collocation and made no distinction between these types in its analysis. It is also unusual in describing as collocations combinations with low frequency of occurrence, *t*-scores and/or MI scores: many studies would describe these as combinations rather than collocations. From the data presented, it appears that it was mainly due to these low-frequency combinations that the item scores had a correlation of .45 with corpus frequency. It may therefore be that frequency has an impact on the acquisition of combinations, but any combination over a certain threshold (which could perhaps be said to mark collocations from other combinations) has a similar likelihood of being known. The correlation between the item scores and the COCA-based MI scores of -.16 is also interesting, given the earlier discussion of learners' relatively poor knowledge of collocations with high MI scores. However, although there was a large range in the MI scores among the 45 collocations tested, only one had an MI score below 3.00 (a score often used as a threshold in identifying collocations). It is not therefore clear whether MI scores have no association with knowledge of collocations, as the study suggests, since almost all the items had relatively high MI scores. One other issue with the study is that the combinations included constituent words from the most frequent 5,000 words of English. Given that some participants were beginners, it is likely that some of the component words were unknown to some of the participants. This raises the issue of whether knowledge of collocations was confounded by knowledge of individual words.

3.2.3 L2 learners' productive knowledge of collocations: Summarising the findings

This section brings together the findings of learner corpus and elicitation studies on L2 learners' productive knowledge of collocations, drawing both on the studies reviewed above and others in order to answer the question: *What has been discovered about learners' productive knowledge of collocations?* It should be noted once more that since there is no established definition of collocation, different studies make use of different definitions, complicating comparisons between them. The findings are in four areas.

First, several studies have found that L2 learners/users use fewer collocations in their writing as compared with L1 users (Fan, 2009; Granger, 1998; Howarth, 1998b; Laufer & Waldman, 2011). Other studies (O'Donnell, et al., 2013; Siyanova & Schmitt, 2008), however, found no difference in the number used, while it should be noted that in the case of the studies just cited that were examined in Section 3.2.1 (i.e. Howarth, 1998b; Laufer & Waldman, 2011), there were questions about the reliability of the findings. On balance, it is perhaps likely that learners use somewhat fewer collocations, but the extent of any difference is unclear.

There are also studies that offer a more nuanced picture of collocation use, finding that while L2 users use fewer strongly associated collocations (as identified by high MI scores), they use as many or more collocations of high frequency (as identified by high *t*-scores) (Bestgen & Granger, 2014; Durrant & Schmitt, 2009). This finding is reinforced by Ellis et al.'s (2008) processing study and by González Fernández and Schmitt's (2015) elicitation study which similarly suggest L2 learners/users favour more frequent, rather than more strongly associated, collocations. While Section 3.2.1 noted criticisms of the use of MI scores since they emphasise low-frequency combinations, in all the studies cited above, except the Bestgen and Granger study, combinations with very low co-occurrence frequencies were omitted and thus major distortions of MI scores were avoided.

A second area of interest has been the relationship between productive knowledge of collocations and general L2 proficiency. Several studies have found a positive relationship (Barfield, 2009a; González Fernández & Schmitt, 2015; Revier, 2009), while some have not (Howarth, 1998a; Nesselhauf, 2005). These last two studies both focused on advanced learners, and so the lack of a relationship may be

due to limited variance among the learners. There are also, however, questions about the studies finding a positive relationship: González Fernández and Schmitt relied on self-reported proficiency, while Revier's study had no measure of proficiency as such and assumed that years of schooling equated with proficiency level. Intuitively, a relationship between proficiency and knowledge of collocations might be anticipated since a beginner would be expected to have very little knowledge of collocations and an advanced learner to have rather a lot and there must be some progression between these two states. However, the evidence at present is rather weak (see also Section 10.3).

Other studies show different findings for different types of collocation. Looking at lexical bundles, studies have found a negative relationship between the number of bundles used and proficiency (Adolphs & Durrow, 2004; Groom, 2009; though see Lenko-Szymanska, 2014, for contrasting results), which has been explained as resulting from learners' gradually improving ability to vary and inflect bundles. Looking at collocations identified by *t*-scores and collocations identified by MI scores, intermediate learners have been found to use more of the former in comparison with advanced learners, while underusing the latter (Granger & Bestgen, 2014), which mirrors comparisons of learners and L1 users cited above.

These findings regarding learners' use of collocations versus L1-user use and comparisons of learners at different proficiency levels perhaps allow a tentative conclusion that knowledge of collocations develops in the direction of becoming more like that of L1 users. Section 2.3.6 discussed criticisms of the assumption that L1 users are the appropriate model for language learning. This tentative conclusion could be seen as suggesting that, notwithstanding the criticisms, learning does mean moving towards L1-user norms. Alternatively, it may be that, since the "native-speaker model" is so dominant currently in language teaching practices, learners do move towards L1-user norms at present, but this would not necessarily be the case were teaching practices different. A third possibility is that while movement towards L1-user norms takes place as learners acquire the essential features of the language, the ultimate destination for learning is not L1-user equivalence but some other norm, similar to but not identical to L1-user norms (see also Sections 10.6 and 11.4).

A third issue of interest has been the frequency of learners' "errors" with collocations and the extent to which these errors are attributable to L1 influence.

Both Laufer and Waldman (2011) and Nesselhauf (2005) found that around a third of collocations used by learners are problematic (though Howarth (1998b) classified only 6% of the verb + noun combinations in his corpus as errors), while it has been estimated that two-thirds of errors involve the selection of the collocation components and one-third other elements such as determiners and number (Nesselhauf, 2005; Wang & Shaw, 2008). L1 influence appears to be a major factor in these errors (Fan, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005), though intra-lingual L2 errors are also significant (Fan, 2009; Nesselhauf, 2005), while one study (Wang & Shaw, 2008) reports that advanced learners of very different L1 backgrounds used and mis-used collocations in similar ways, suggesting the L1 is of less importance. It seems learners do acquire collocations, but they may not always be acquired perfectly. This may be because adult L2 learners are analytic in their approach and focus on words rather than phrases (Wray, 2002, 2004).

A final area of interest is factors that may affect the difficulty of collocations for learners. There are findings pertaining to four factors:

- Following on from the above point, it has been found that learners tend to use L2 collocations that are congruent with L1 collocations (Granger, 1998), and that congruent collocations are more often used acceptably than non-congruent (Nesselhauf, 2005). Congruence has also been found to affect learners' processing of collocations (Wolter & Gyllstad, 2011, 2013; Yamashita & Jiang, 2010).
- Restricted collocations have been found to be used acceptably somewhat less often than free combinations (Nesselhauf, 2005). Wang and Shaw (2008), however, concluded that advanced learners have problems with common collocations whether they are restricted or free. It is therefore currently unclear whether restrictedness affects difficulty.
- Collocations that are more restricted were found to be more often used acceptably than less restricted collocations in learner writing (Nesselhauf, 2005). In combination with the previous point, this suggests a complex interaction between degree of restrictedness and acceptable use. However, given that few studies have investigated such issues, it must still be considered unsettled whether or how the degree of restrictedness affects difficulty.

- Learners of various proficiency levels have been found to produce adjective + noun collocations, while more advanced learners increasingly produce verb + noun collocations also (Barfield, 2009a). Relatively few studies involve more than one type of collocation, so there is little evidence to support or dispute this single finding.

The above summary shows that research on learners' productive knowledge of collocations has made some progress, and in some areas findings have been made across multiple studies. This research must still, however, be considered to be at an early stage: the findings often seem rather fragile and methodological issues are very much still being worked through. The following section therefore reviews problems seen in learner corpus and elicitation studies.

3.3 Researching L2 learners' productive knowledge of collocations: Problems and challenges

Learner corpus and elicitation-based studies face a number of problems and challenges, some of which are common to both types of study. This section describes problems with learner corpus studies first, and then discusses elicitation studies.

3.3.1 Problems and challenges with learner corpus research

There are five principal problems with learner corpus research on L2 learners' productive knowledge of collocations: (1) the difficulty of collecting writing samples; (2) variability within corpora; (3) the comparability of corpora; (4) the reliability of classifications; (5) the confounding of knowledge of collocations by knowledge of words.

The first problem is the difficulty of collecting writing samples from learners. In many circumstances, it is not feasible to collect a substantial enough piece of writing from a learner to reflect their knowledge of collocations: that is, a collocation may not appear despite it being known (Nesselhauf, 2005). With short pieces of writing in particular, the effects of the specific writing task, both in terms of the task itself and topic (see Forsberg & Fant, 2010; Gablasova, et al., 2017) come to the fore. This is of course precisely why learners' writing is usually treated as a collective (i.e. described as a corpus), rather than investigated on a text-by-text basis. However, one consequence is that longitudinal research, in which data is collected on several occasions from learners and individual gains and losses compared between times, is

very difficult (though see Adolphs & Durrow (2004) and Bell (2009) for longitudinal case studies). A further consequence is that studies are often narrow in that they involve only advanced learners; a focus which has affected perceptions of collocation. There seems to be an element of circularity in the fact that, as Barfield (2009a) points out, many studies have focused on advanced learners, and then, finding errors in their collocation use, declared collocations to be an area of language that is only beginning to be mastered at an advanced stage. If there were research involving learners across a range of proficiency levels, collocation may instead be seen as something that begins to develop at an early stage, with learning continuing through to advanced levels and beyond, rather like vocabulary knowledge in general.

The second problem, very much related to the first, is that variability within corpora is seldom explored. Although some studies (e.g. Durrant & Schmitt, 2009; O'Donnell, et al., 2013) have looked at internal variability, most learner corpus studies do not, and simply produce results for the corpus as a whole. This means that internal variability remains hidden and unexplored, and also means inferential statistics cannot be used, making findings weaker and more uncertain.

The third is the comparability of corpora. For full comparability, the contributors to the corpora must be of similar levels of writing expertise and experience, as O'Donnell, Römer and Ellis (2013) and Neff van Aertselaer (2008) have shown, while the writing itself must be at least of the same type, and should ideally have been produced under the same task conditions and be on the same topics. A basic problem, as Durrant and Schmitt (2009) point out, is that the types of writing often done by learners are not often done by L1 users. Gablasova, Brezina and McEnery (2017) note that a particular problem is the use of general corpora as a reference point for a learner corpus (which is typically of one particular genre, such as argumentative essays). They argue that just as single words vary in frequency across genres, so do collocations, and so these two types of corpora should not be compared. These problems make it disappointing that learner corpus studies do not seem to take steps to evaluate the comparability of the corpora first before embarking on their primary analyses.

The fourth problem is the reliability of classifications. These may include classifying combinations as collocations or otherwise and classifying collocations as "acceptable" or "deviant". Frequency-oriented research relies on statistical measures

for classifications, but Gablasova, Brezina and McEnery (2017) are critical of how these are used, arguing that the measures are often poorly understood in terms of the mathematical reasoning behind them, the scale on which they work and their practical effects. Using these measures also involves researchers in making a number of decisions which are essentially arbitrary (see Section 2.2.2). Phraseological studies often make use of human raters, yet despite the fact that much research in applied linguistics involves making classifications and there are well-established procedures for the use of multiple raters and the calculation of inter-rater reliability, in many learner corpus studies these procedures are curiously absent. One particular issue with making classifications is the use of L1 users as the reference point. Laufer and Waldman's (2011) study has a particular problem in this regard as the study design biases the findings against the learners, but there is a wider problem in collocation research with the use of L1-user norms, as discussed in Section 2.3.6.

A final problem is that knowledge of collocations can be confounded by knowledge of words, an issue raised regarding Laufer and Waldman (2011) and Durrant and Schmitt (2009). That is, if the intention is to test knowledge of a collocation, but learners do not know the component words in that collocation, it will likely be found that the learners do not know the collocation. However, it can be argued that this does not really show a lack of collocation knowledge. To show that, it has to be demonstrated that the learners know the component words, but not that they can be used in combination (i.e. that they collocate). Wang and Shaw (2008) also note that the question of word knowledge can distort views of errors: "collocation errors" in the eyes of some may in fact reflect limited word knowledge. They give the example of "do damages", which may reflect limited knowledge of the word *damage* (i.e. that it is uncountable with this sense), rather than limited knowledge of the collocation itself.

3.3.2 Problems and challenges with elicitation research

Elicitation studies likewise face a number of challenges. Five problems are explained below: (1) task design; (2) task validity; (3) item selection; (4) the reliability of classifications; (5) the confounding of knowledge of collocations by knowledge of words.

Perhaps the biggest challenge for elicitation studies may be task design. Many

elicitation tasks involve a sentence context in which the target collocation is embedded, and so a number of potentially confounding factors which could affect responses must be avoided. These include the syntactic complexity of the sentence context and the frequency of the words and word combinations in the sentence context. Another issue is that many elicitation tasks target particular collocations and so must provide sufficient prompts and guidance to ensure that those particular collocations are elicited, while at the same time ensuring that participants are actually producing the collocations. Some formats, such as Revier's (2009), may be too constrained and may be considered receptive tests rather than productive measures, while others, such as Bahns and Eldaw's (1993) are not constrained enough in that participants can respond quite legitimately without using the targeted collocations.

A second issue, related to that above, is task validity and particularly the question of whether an elicitation task mirrors free production of collocations. González Fernández and Schmitt (2015), for example, note of their study that while participants "are able to produce the written form of a substantial number of collocations . . . it is a matter of speculation the degree to which the participants would be able to employ this knowledge in their own free writing and speaking" (p. 106). Mollin (2009) suggests, in discussing word-association-like instruments, of which Barfield's (2009a) is an example, that some elicitation tasks may encourage more analytic processing and so differ from normal language production. It would seem necessary then, for the validation of an elicitation task, to explore the extent to which the task is akin to free production. If this is not undertaken, it is unclear exactly what a task is eliciting.

Third, there is the issue of how items can be selected in a principled manner. Early studies give very little information on item selection. More recent studies are generally better in this regard, often explaining various criteria that had to be met. There is nonetheless often little information on how collocations were selected from among those that met the criteria.

A fourth issue, also raised regarding learner corpus studies, is the reliability of classifications. For most elicitation studies, the key issue is determining what counts as a collocation. Many studies follow the frequency tradition, and so the central issues are the suitability of the reference corpus with respect to the expected

participants and the informed use of statistical measures of collocation.

Finally, there is another problem shared with learner corpus studies: the possibility that the measurement of knowledge of collocations is confounded by knowledge of words. This may be an issue both with regard to the sentence contexts used, as already pointed out above, and also with regard to the collocations themselves. That is, as was noted regarding González Fernández and Schmitt (2015), responses may be governed as much by knowledge or otherwise of the component words in a collocation as by knowledge of the collocation itself.

3.4 Moving forward

Section 3.2.3 showed that research on L2 learners' productive knowledge of collocations has made some progress. This research, however, is still at an early stage and is limited by methodological issues. On the one hand, learner corpus studies can provide useful insights into learners' knowledge and use of collocations, but this approach is not really suitable for longitudinal research and the fact that it is only really viable with learners of rather high proficiency levels means that cross-sectional research designs are limited also. Elicitation studies could overcome these shortcomings, but most studies to date face questions regarding the elicitation task and what it measures.

One elicitation instrument that may have potential, however, is Barfield's (2009a) *LexCombi*. *LexCombi*: (1) is extremely simple in design, with no sentence context to complicate matters, and no question of it possibly being a receptive task; (2) has an element of free production to it, so avoids the narrowness of targeting a specific set of collocations, but at the same time is more constrained than a writing task; (3) is highly efficient in the generation of data, allowing a lot of collocations to be elicited in a relatively short period of time; and (4) can be used with learners with a range of proficiency levels, since the task is so simple and the cue words are of high frequency, opening up greater possibilities for research designs.

There are of course issues with the instrument that remain to be explored. However, notwithstanding these issues, *LexCombi* would appear to be a potentially useful instrument for exploring learners' productive knowledge of collocations, and as will become clear in the remainder of this thesis, there is substantial scope for developing the instrument further in productive ways. This sort of iterative

development does not seem to have taken place with any of the elicitation instruments reviewed above, but has occurred, for example, with the Yes/No vocabulary size test (see, among other studies, Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Huibregtse, Admiraal, & Meara, 2002; Meara & Buxton, 1987; Mochida & Harrington, 2006). In this case, an initial study led to a series of further studies looking at different aspects of how the test works, with whom it works, how the data it provides should be treated, and more. This may be an ongoing process and can perhaps never reach an end point, but it does lead to instruments that are better understood, both in terms of what they can tell us and their limitations.

This thesis will therefore take forward *LexCombi* as a potentially useful means of exploring L2 learners' productive knowledge of collocations. It will seek to investigate issues with the instrument, and will carry out a series of investigations of different aspects of the instrument. The thesis, then, seeks to answer three questions:

- What strengths and weaknesses does *LexCombi* show as a tool for measuring L2 learners' productive knowledge of collocations and how can it be improved?
- What insights into the development of productive collocation knowledge may *LexCombi* provide?
- What wider issues for our understanding of collocation are raised by the development of *LexCombi* and an initial exploration of *LexCombi* data?

Chapter 4 presents an initial trial of *LexCombi*. This was a partial replication of Barfield's (2009a) study, intended to begin the exploration of how the instrument performs. Chapter 5 goes on to further explore *LexCombi* and how learners interact with it. Chapters 4 and 5 reveal a problem with the format of *LexCombi* and Chapter 6 therefore trials an adaptation to the format which sought to better steer participants towards producing collocations. Chapter 7 explores the scoring of *LexCombi*, an issue raised above in reviewing Barfield's initial *LexCombi* study. Chapter 8 trials new cues, in order to address issues with the original set. Chapter 9 reports on a trial and evaluation of *LexCombi* in its final form. Chapter 10 draws together the findings from the empirical studies, to consider what can be learned about collocation knowledge using *LexCombi*. Chapter 11 reflects on the overall significance of this work for our understanding of collocation knowledge more generally, before Chapter

12 describes the conclusions of the thesis.

Chapter 4 Trialling *LexCombi*

4.1 Introduction

The aim of this chapter is to report on a trial of *LexCombi*, an instrument developed by Barfield (2009a) with the aim of eliciting learners' productive knowledge of collocations. This trial and the issues that arose from it provide the rationale and context for the main research studies that follow in subsequent chapters.

LexCombi was identified as a promising instrument in Section 3.4. *LexCombi* presents 30 cue words to participants and asks for three collocates for each cue, giving a 30-second time limit per cue. *LexCombi* is, then, extremely simple in design, and is more open than many elicitation instruments in that participants are free to give any collocates for the cues, rather than specific collocations being targeted. At the same time it is more constrained and focused than free production, which makes *LexCombi* highly efficient: up to 90 collocations can be collected from a participant in just 15 minutes. In addition, *LexCombi* appears suitable for use with learners of varying proficiency. The simplicity of the task and the fact that its cue words are of high frequency mean it can be used with less proficient learners, while more proficient learners are also able to demonstrate their knowledge of collocations.

There are two issues with *LexCombi*, however. The first is how best to determine whether a response to *LexCombi* is a collocate or not: in other words, there is a potential problem with scoring the responses. Barfield's scoring was based on a previously compiled list of collocates for each item. Each response was checked against this list, and a participant's *LexCombi* score was the total number of responses which were on the lists, the maximum score being 90.[1] Barfield made his lists by taking collocates from the *Oxford Collocations Dictionary* (1st edition) (Crowther, et al., 2002) and from *Collins Wordbanks Online* (HarperCollins, 2004). However, the motivation for choosing these sources is unclear, and their appropriacy may be questioned.

The second issue concerns the validity (in the technical sense) of the *LexCombi*

[1] Barfield refers to collocates on the lists as "appropriate collocates"; several other researchers use the term "acceptable" in similar contexts. Yet it is not always clearly stated to whom the collocates are appropriate or acceptable. This thesis will henceforth use the term "canonical", which is intended to designate only that a collocate is on the scoring list.

task. Barfield's aim was to develop an instrument that allows insights into learners' ability to produce collocations. However, although it does have an element of free production to it, the task is clearly somewhat different from free production. There is a need, therefore, for an exploration of whether learners have knowledge of how the responses they give to *LexCombi* are used in combination with the cues.

Sections 4.2-4.4 describe the method used in the trial of *LexCombi* and show the results, comparing them with Barfield's findings. Section 4.5 takes a fresh look in the light of this trial at the issues with *LexCombi* identified above (scoring and task validity), and reports on three further issues which arose in the course of the trial.

4.2 Method

The study sought to trial *LexCombi* by replicating Barfield's study as far as possible. This was aided by Barfield's willingness to provide some of his files. *LexCombi* was thus used with the same 30 cues in the same two variant orders, the same examples were given, and the same instructions for participants were used, though translated from English into Japanese to ensure full comprehension. The instructions and *LexCombi* itself are shown in Appendix A.

A similar group of participants to Barfield's was also used: 77 Japanese university students. The participants were volunteers from a total of 84 learners in three classes that I had been teaching for an academic year. All were aged 18-20 and were majoring in subjects other than English, though had at least two English classes per week. The data was collected immediately after the final classes of the academic year; those willing and able to participate remaining in the classrooms during their own time. Unfortunately, no proficiency measure was available for these participants, but I would rate them as generally low-intermediate, and a small number had TOEIC scores in the 500-600 range, approximately B1 on the CEFR scale (Educational Testing Service, n.d.).

After a brief oral introduction to the aims of the study, participants were given the test booklets, with the two variant orders of cues assigned randomly. As Appendix A shows, the 30 noun cues were listed down the left side of the page and three spaces were provided next to each cue for the responses. Participants were given time to read the instructions, then three practice items were attempted, and possible responses were reviewed as a group. Participants then completed *LexCombi*,

looking at one cue at a time, with an oral instruction to move onto the next cue every 30 seconds. The entire process took just over 20 minutes.

The responses were typed up in full in an Excel file (with the original test booklets retained for reference), then reviewed and cleaned up as detailed in Table 4.1. This was necessary to deal with two issues: multi-word responses and misspelled responses.

Barfield's intention with *LexCombi* was to elicit single-word responses, the desire being to "strip away the grammatical, syntactic and contextual features" and focus on the "basic lexical task of producing appropriate collocates" (p. 96). However, the instructions, as devised by Barfield and adopted in this study, did not specify that single-word responses should be given, and, indeed, the examples of collocations in the instructions included *borrow a* and *go to*, which may have

Table 4.1: Procedures for cleaning up response data.

Procedure	Cue	Example response	Cleaned up response
The word nearest the cue was taken as the response with the exception of determiners, link verbs and *sb*/*sth*. If, however, these are the only words given as a response, they are accepted as responses.	CAR	drive a car	drive
	EXPERIENCE	experience is fun	fun
	QUESTION	ask sb a question	ask
	POWER	power of love	of
	COUNTRY	my country	my
If a response included words both to the left and right of the cue, the word on the left was taken as the response. If there were intervening articles or link verbs, the first principle above applied, and the valid word nearest the cue was taken as the response.	HEALTH	keep health good	keep
	EXAMPLE	have a example of sth	of
Obvious misspellings were corrected.	CHILD	child seet	seat
Numerals were changed to words, and the abbreviations *sb* and *sth* changed to *somebody* and *something*.	QUESTION	1	one

suggested that multi-word responses were permissible. Barfield (personal communication) reports that in his study there were few multi-word responses. In this study, however, there were many such responses. Some were of a simple nature: for example, for the cue QUESTION, the responses *have a* and *ask a*. In other cases, they were less so: for example, for HEALTH, one response was *do something for your*.

In cases where multiple words were given as a response, an obvious issue arises about how to score them. One option would be to score each word in a response in relation to the cue, but that would overlook the relationships between the response words occurring together. Another option would be to regard the words jointly as the collocation. This has some merit, as the multi-word responses did typically have a single semantic function in conjunction with the cue. However, a separate scoring protocol would be needed that could be accommodated alongside the single-word response protocol. A third option is to select one of the words as the core collocate. The advantages of this include the retention of a simple scoring system, and the consistency, within the studies, of dealing only with two-word collocations. This third option was chosen and consequently the multi-word responses necessitated decisions about which word to take as the response.

Barfield's approach was to focus on lexical words: he reduced multi-word responses "to the main lexical element (e.g. *take care of* became *care*), [and] phrasal verbs were scored for the base verb (for instance *come* for *come up with*)" (p. 98-99). In this study, a slightly different approach was taken (see Table 4.1), in which, with the exception of determiners, link verbs and *sb*/*sth*, neither lexical nor function words were prioritised and the guiding principle was to accept the nearest word to the cue as the response.

This was for two reasons. First, under Barfield's method, it is unclear what approach to take when no lexical words were present. There were numerous responses in this study consisting of a function word alone or even two function words together. Second, learners' knowledge of collocating function words is of intrinsic interest as it gives an indication of their understanding of the grammatical patterning of cues. Hunston and Francis (2000) provide evidence of the importance of such knowledge. In their terms, for example, the cue POWER has the pattern **N of n**. Where participants provided responses such as *of love*, *of things* and *of something*,

the *of*, which implies knowledge of this pattern, is of more interest than their suggestions for words that could fill the relatively open noun phrase slot. Thus the procedures detailed in Table 4.1 ensured that in the above examples *of* was taken as the response.

Misspelled responses were treated as follows. Apparently misspelled responses which formed an English word were not corrected. Responses which did not form an English word were corrected when the intended response was obvious, but not when there was any ambiguity over the intended response. For example, the response *seet* for the cue CHILD was corrected to *seat*, but the response *plice* for the cue LAW was not corrected, since there was doubt regarding whether the intended response was *police* or *price*.

A second difference with Barfield's study concerns the scoring of responses. As mentioned above, Barfield used the *Oxford Collocations Dictionary* (1st edition) (Crowther, et al., 2002) and *Collins Wordbanks Online* (HarperCollins, 2004) to create lists of collocates for each cue, accepting any collocate from either source as canonical. As explained in Section 2.5, this thesis has adopted Durrant's definition of collocation, which sees collocations as word combinations that require specific knowledge to be understood or produced, or that facilitate fluency due to their frequency. For this trial of *LexCombi*, however, the scoring of responses was not guided by Durrant's definition, as the intention was to replicate Barfield's study as far as possible. Unfortunately, at the time this trial was conducted the *Collins Wordbanks Online* website was no longer available, meaning that the exact lists of canonical responses that Barfield used could not be replicated. In addition, access to a copy of the first edition of the *Oxford Collocations Dictionary* could not be obtained (the dictionary being no longer on sale). Consequently, the second edition of the *Oxford Collocations Dictionary* (OCD2) (McIntosh, 2009) was used instead to judge the canonicity of the responses.

Using the OCD2 lists, the participants' responses were scored automatically using a specially written program named CollCheck (Imao & Brown, 2014) developed by a colleague and myself. CollCheck checks each response against the list of collocates for that cue, disregarding letter case, and tabulates the number of responses on the lists for each participant. This total, which is the *LexCombi* score, reflects the extent to which a participant responded in a "canonical" fashion: that is,

in line with the existing set of recognised collocates provided by the dictionary.

4.3 Results

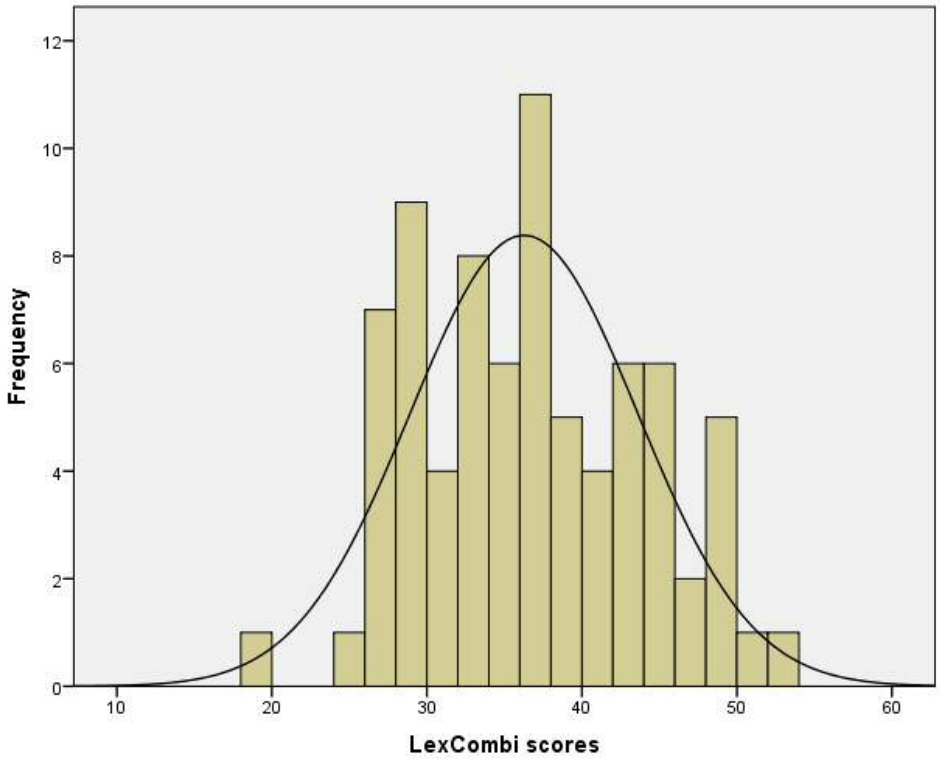
The descriptive statistics for the *LexCombi* scores, that is the number of canonical collocates provided by participants, are shown in Table 4.2. The results are set alongside Barfield's for comparison. Figure 4.1 shows the distribution of scores in the current study.

Table 4.2: Descriptive statistics for the *LexCombi* scores.

	Current study	Barfield (2009a)
N	77	89
Mean	36.26	37.93
SD	7.33	9.74
Minimum	19	17
Maximum	53	61

Note. Maximum score = 90.

Figure 4.1: Distribution of *LexCombi* scores in the current study.



4.4 Discussion

The results above demonstrate that *LexCombi*, as intended, elicits a considerable number of canonical collocates from learners in a short period of time. In the current study *LexCombi* has produced scores that have a reasonable distribution and that differentiate the participants. Despite the differences between the lists of canonical collocates used in the current study and in Barfield's, the scores are broadly consistent with Barfield's results.

4.5 Issues with *LexCombi*

In reviewing Barfield's *LexCombi* study, two issues (its scoring and task validity) were identified, as described in Section 4.1. In the course of administering *LexCombi* and scoring participants' responses, three further issues with *LexCombi* arose. These concerns suggest that any interpretation of *LexCombi* data will be compromised by unresolved questions regarding what exactly has been collected and how. Therefore, instead of proceeding with an analysis of the present data, the next sections consider these issues. The following sections make it clear that modifications to *LexCombi* are required, and that new data will need to be collected with a revised version of the instrument before useful inferences can be made from *LexCombi* data. The next sections, then, address five questions concerning *LexCombi* and the data it elicits:

1. What is the most appropriate point of reference for categorising responses as collocates or not?
2. What collocation knowledge can a cue-response actually reflect?
3. Are participants responding with collocations in particular, as desired and instructed, or with associations in general?
4. What do missing responses signify?
5. What is the impact on an analysis of cues being interpreted as verbs rather than, as intended, nouns?

4.5.1 The reference point for categorising responses as collocations

The first question is about the most appropriate reference point for categorising responses as collocates. That is, to produce a *LexCombi* score, a decision must be made with respect to each response given by a participant as to whether it is a collocate of the cue or not. As explained previously, Barfield created a list of

canonical collocates for each cue and checked whether each response was on the list or not. Barfield's lists were created by combining collocations listed in the first edition of the *Oxford Collocations Dictionary* (OCD1) (Crowther, et al., 2002) with corpus searches for collocations using *Collins Wordbanks Online* (HarperCollins, 2004), while in the current study the second edition of the *Oxford Collocations Dictionary* (OCD2) (McIntosh, 2009) was used.

This section first considers the appropriateness of these sources and discusses the criteria used to determine the content of the lists of canonical collocates. It then examines individual responses which were given by a substantial number of participants, since these recurrent responses would appear to have some currency for the participant group. Of particular interest are those that were not on the lists, and the question of whether they should have been counted as "canonical" or not.

Regarding the appropriateness of the sources, Section 2.3.6 discussed the language variety used in judging learners' collocations. To briefly re-iterate, there have been a number of criticisms of the use of L1-user norms, particularly with respect to corpus-based work but also more generally. L1-user norms are criticised for causing learners to feel inadequate (Widdowson, 2000, 2003), for failing to recognise that learners are bilinguals and so unable to become monolinguals in the target language (Grosjean, 1989), and for not acknowledging learners' accomplishments in learning to use their L2 (V. Cook, 1999).

Barfield's two sources are both based on L1-user corpora, and the source used in the current study is based on a corpus mostly consisting of L1-user texts. More specifically, the OCD1 was based on the 100-million-word BNC, entirely composed of British English texts. *Collins Wordbanks Online* enabled searches of 56 million words from British and American texts in a 4:1 ratio. The OCD2 made use of the Oxford English Corpus. This is a very large corpus of almost 2 billion words at the time the dictionary was compiled, 80% of which consists of British and American texts with small proportions from a number of countries where English is the main or an important language (Oxford Dictionaries, n.d.). Interestingly, under the heading "Why is collocation important?", both editions of the *Oxford Collocations Dictionary* give one reason as: "For the student, choosing the right collocation will make his speech and writing sound much more natural, more native-speaker-like, even when basic intelligibility does not seem to be at issue" (p. vii). Depending on

one's point of view, this can be seen as an endorsement of the idea that learners should follow L1-user norms or as an acknowledgement that many teachers and learners hold this view.

Criticism of the "native-speaker model" has led to calls for a re-orientation towards a model based on English as a Lingua Franca or English as an International Language (see, for example, Mauranen, 2011; Seidlhofer, 2005). Nesselhauf (2005), however, has pointed out the difficulty of replacing L1-user norms, since it is difficult to describe an alternative or even ascertain the existence of one. Recent ELF research acknowledges this issue, and suggests that ELF research is less about establishing new conventional forms and more about the processes and strategies speakers use to communicate successfully (Cogo & Dewey, 2012).

Determining the most appropriate sources from which to identify canonical collocates is not then an easy task. However, especially in light of the fact that one of Barfield's aims with *LexCombi* was to overcome the deficit view of collocation and focus on what learners know rather than on what they do not know, an aim that this thesis shares, judging their responses solely against L1-user norms does not seem appropriate.

The second area of concern with the scoring is the criteria used to determine which words are included on the lists. Of the two sources used by Barfield (2009a), one, *Collins Wordbanks Online* (HarperCollins, 2004), provided ranked lists of up to a 100 collocates for a word, either in terms of *t*-scores or MI scores (see Section 2.2.2). Decisions were thus required about the choice of measure (i.e. *t*-scores or MI scores) and about a threshold to determine which results would be included. Barfield (personal communication) used *t*-scores, and set the threshold at a *t*-score of 2.0. This threshold is conventional (Hunston, 2002), but its basis is questionable. *T*-scores (the collocation measure) are derived from the *t* test (a statistical confidence test) and are meant to indicate whether the two words co-occur significantly more often than would be expected based on the individual frequency of each component word. A *t*-score of 2.0 is considered "significant" because using the *t* test, a statistic over 1.96 indicates significance. However, Evert (2008) points out that the calculation of *t*-scores for word combinations does not meet the assumptions necessary for a valid test of significance (see also Section 7.2.2).

In using his second source, the *Oxford Collocations Dictionary* (1st edition) (the

second edition of which was used in the current study), Barfield was not required to make any decisions, criteria for the make-up of the lists having already been selected by the dictionary editors. The introductions to the first edition (Crowther, et al., 2002) and second edition (McIntosh, 2009) of the *Oxford Collocations Dictionary* show that they were compiled in similar ways, but in neither is the exact process of compilation explained. Both mention that "the main source" (p. viii) of data was a corpus (as noted, the BNC and the Oxford English Corpus respectively) and that compilers were able to "check how frequently any given combination occurred, in how many (and what kind of) sources, and in what particular contexts" (p. viii). Both also state that "totally free combinations are excluded and so, for the most part, are idioms" (p. viii), and that the collocations included were restricted to particular patterns: e.g. in the case of verb entries, adverb + verb, verb + verb and verb + preposition, with short phrases including the headword also appearing.

The broader issue here is not so much the precise decisions taken in Barfield's study and in this one, but rather how researchers can make principled decisions in this area. Corpus linguistics has given us a range of measures and some conventions have arisen regarding thresholds for what are described as "significant" scores, but the selection of one measure over another or a slightly higher or lower threshold can remain unexplained. As Davies and Gardner (2010) admit regarding one of these measures, "using MI is sometimes more an art than a science" (p. 6) in that, to quote Sinclair, Jones and Daley (1970) once more (see Section 2.2.2), "although the significance tests in themselves are objective, they rest on essentially arbitrary choices made by their users" (p. 72).

It should also be noted that decisions taken about sources and criteria for their use, often with practical considerations in mind, imply a particular conceptualisation of collocation. Thus, for example, while both editions of the *Oxford Collocations Dictionary* claim that the approach taken was "pragmatic, rather than theoretical" (p. viii), it is apparent from the details above that the approach resulted in the inclusion of particular types of collocates and reflected a particular notion of collocation.

In order to better evaluate the appropriateness of the scoring method used in the current study, which in essence followed Barfield, individual responses (both canonical and non-canonical) which were given by a substantial number of participants were examined. This examination looked at responses given by 10% or

more of the participants (hereafter referred to as "recurrent responses"), such as *play* for the cue ROLE and *make* for the cue DECISION. In the current study, with 77 participants, any response given by eight or more participants was therefore a recurrent response. There were 145 recurrent responses in all, an average of almost five for each of the 30 cues. These responses accounted for 2,507 (45%) of the participants' 5,566 responses. Two questions were asked of the recurrent responses: (1) What proportion were canonical? (2) What is the nature of those which were deemed non-canonical by the current scoring criteria?

The majority of the recurrent responses were canonical collocations (i.e. they appear in the OCD2): 92 (63%) of the 145. For a typical cue, three or four of its recurrent responses were on the OCD2 list, and one or two were not. There was, though, a good deal of variation. For example, the cue VALUE had five recurrent responses, all canonical, whereas BODY had six, but only one was canonical. Interestingly, the more participants that gave a recurrent response, the more likely it was to be canonical. There were 93 responses which were given by between 10% and 20% of the participants, and so only narrowly passed the 10% threshold for being considered recurrent, of which 51 were canonical. In contrast, there were seven responses given by 50% or more of the participants, every one of which was canonical. As a group, these low-intermediate proficiency learners had then a strong tendency to produce canonical collocates.

The second question above is about the nature of the recurrent responses which were deemed non-canonical. An examination of the 53 non-canonical recurrent responses showed there to be four types. First, some simply revealed the limitations of the OCD2. For example, the recurrent responses *my* for the cue FAMILY and *why* for REASON do not appear as collocates in the OCD2. Yet in both the COCA and the BNC these responses are important collocates, occurring with high frequency and having high MI scores. Such collocations were presumably omitted from the dictionary since they do not fit into the categorisation scheme. This lends support to Barfield's decision to use two sources for his collocates lists, as such omissions would be less likely, and suggests that new scoring lists for *LexCombi* may benefit from considering multiple sources.

Second, some non-canonical recurrent responses seemed to derive very much from the particular linguistic environment of this group of participants. This includes

both the wider environment of English as it is used within Japan and the participants' experience of English in the classroom. One example of the former was the response *note* for the cue DEATH. The recurrence of this response will be easily understood by those familiar with contemporary Japan; it likely came from the title of an extremely popular manga and film series *Death Note* (Japanese: デスノート *desu nōto*). An example of the latter was the response *recorder* for the cue VOICE. As already noted, the data was collected in the final class of an academic year in which the participants' main exposure to English would likely have been in their two weekly classes. In these classes the participants had used *voice recorders* several times. This combination was thus a prominent part of their recent English experience.

Third, some of the non-canonical recurrent responses seemed to come from compounds: for example, two recurrent responses were *news* for the cue PAPER, and *man* for POLICE. These pairs of words appear together very infrequently in corpora, but are of course frequent as single orthographic words (i.e. as compounds). Given the variation with which compound nouns are written, with some written as single words, some as hyphenated words and some as two separate words, and given that compounds may be acquired aurally in any case, it is perhaps not surprising that learners would struggle to be aware of where word boundaries lie. Alternatively, such responses could be seen as evidence not so much of unfamiliarity with conventions, but of a tendency for analysis. This tendency has been noted in work on L2 phraseology (Wray, 2002), but responses of this type may indicate that learners also analyse orthographic words at times. That is to say, even though a learner may have seen *newspaper* as a single orthographic word many times, they may have analysed it as a two-word phrase. In fact, it could be the case that while, for example, *news* and *paper* are strongly associated in a learner's mind, there is no holistic storage of the word *newspaper*. Certainly, there is evidence that L1 users are sensitive to the internal components of compounds (Libben, 2005, 2014).

As Section 2.3.3 noted, Granger and Paquot (2008) observed that though research on collocation is based on the fact that the units consist of two words, the ambiguity of the concept of 'word' means that this basis is not particularly clear. They point out that this issue is rarely discussed in the literature and that the way researchers treat compounds of different types, that is, whether they are included or

excluded as phraseological units, varies a great deal. The examples discussed above give some indication that the difficulties surrounding this issue for researchers may be matched in terms of learners' knowledge of these words.

Finally, there were some recurrent responses that from the perspective of L1-user norms are non-standard. For example, the responses *solve* for the cue ISSUE and *keep* for HEALTH. The former does occur occasionally in the COCA, but the more conventional *resolve* appears with vastly greater frequency and produces a strong MI score. This response may have resulted from the participants mislearning the collocational phrase or perhaps viewing *solve* and *resolve* as equivalents as indeed both can be translated into Japanese 解決する (*kaiketsusuru*). The latter example, *keep*, co-occurs with *health* with extremely low frequency in the COCA, and it seems likely here that the issue is translation from the L1, *to keep one's health*, meaning to protect or look after one's health, is a literal translation of the Japanese phrase 健康を守る (*kenkou o mamoru*).

Considering all the above, the recurrent responses which are not canonical seemed, in many cases, to be reasonable responses. Some are deemed non-canonical due to the approach and decisions taken by the dictionary's editors, while it seems odd to disregard responses derived from the local linguistic environment or the participants' knowledge of compounds. Only a relatively small number are simply non-standard from an L1-user perspective. It therefore seems that the scoring of *LexCombi* requires modification and that a somewhat broader view of collocation may be necessary in order to fully capture the productive knowledge of learners. This, and the other issues above regarding *LexCombi* scoring, will be pursued in Chapter 7.

4.5.2 Knowledge of how cues and responses are used together

The second question is what collocation knowledge a cue-response actually reflects. In seeking single-word responses, Barfield's (2009a) aim, as noted earlier, was to "strip away the grammatical, syntactic and contextual features, and reduce . . . the measure of learners' productive L2 collocation knowledge to a basic lexical task of producing appropriate collocates" (p. 96). The intention was to make the elicitation task simple and quick, thereby allowing more data, and data from a wider range of participants, to be collected. This approach is only reasonable, however, if it can be

demonstrated that a learner giving a single-word response to a cue has some knowledge of how that combination of words is used. If this is not the case, *LexCombi* cannot really be said to be eliciting learners' productive knowledge of collocations.

When a participant, as desired, provided a single-word response, there was no means of judging whether that participant had knowledge of how the cue and response are used together. However, as noted in Section 4.2, the participants in the current study frequently gave multi-word responses, and among these responses the cue word itself often featured. These multi-word responses including the cue provided an insight into how participants believed the cue and response are used together.

In many cases, these responses indicated that the participants were able to use the response and cue in standard ways: responses such as *drive a car* and *a new car* for the cue CAR, and *break a law* and *law school* for the cue LAW. There were also, however, some responses that suggested the participant was less certain of the use of the cue-response combination. For example, for the cue DECISION, one response was *have a decision right*. *Right* (which was taken as the response in accordance with the procedures in Table 4.1) does appear in the OCD2 as a collocate for DECISION, and thus this response was canonical. The full response, however, seems somewhat irregular, and indeed, while the words *decision right* appear together in the COCA, the majority of occurrences are phrases such as *make a decision right now*/*away* (there are, in addition, many examples of the phrase *right decision*, and it is likely this that leads to *right* being listed as a collocate of *decision* in the OCD2).

This incidental evidence that a single-word response, marked canonical, might actually conceal a lack of knowledge about the way a cue and response go together demonstrates the need for a different study design in order to establish participants' capacity to combine cues and responses grammatically. Chapter 5 reports such a study.

4.5.3 Are responses collocations or associations?

The next question is whether participants are responding with collocations in particular, as desired, or with associations in general. The issue is that, despite being

instructed to provide collocates, some participants gave what appear to be non-collocational associations for the cues. For example, for the cue FAMILY, one participant gave the responses *mother*, *father* and *brother*, while for the cue HEALTH one response was *WHO* (i.e. the World Health Organization). There were not, it should be said, a large number of clear instances like the above, and indeed the participants involved in the above examples appear to have provided collocates for the other cues.

The small number of such cases suggests that perhaps this issue can be considered unfortunate but acceptable, a type of noise in the system. However, if participants are on occasion providing associations more generally rather than collocations in particular, they are potentially missing out on opportunities to display their knowledge of collocations, and *LexCombi* is eliciting a distorted view of their knowledge of collocations. This issue is therefore returned to in Chapters 5 and 6.

4.5.4 The significance of missing responses

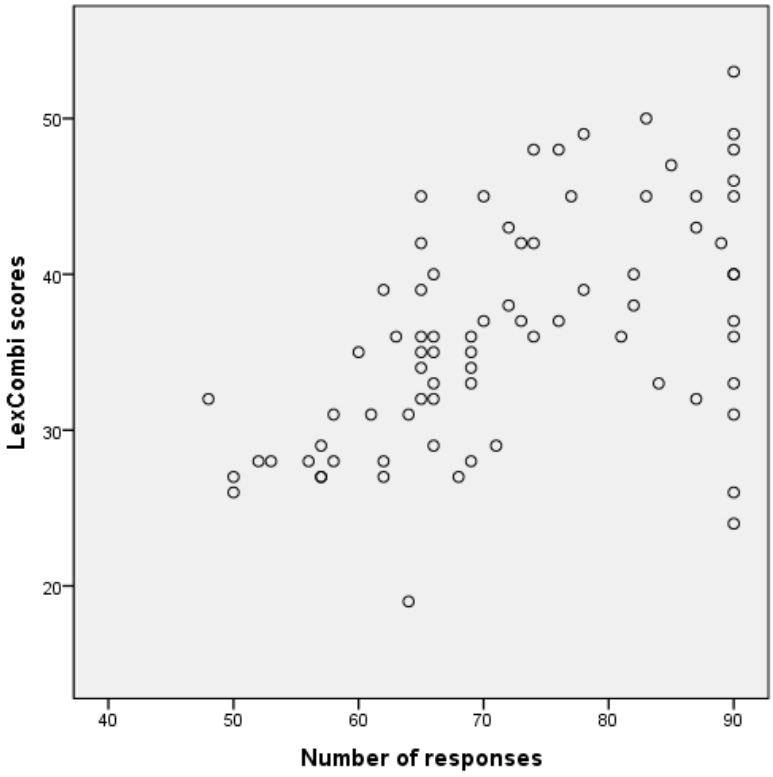
The fourth question is what is signified by missing responses. In the current study, just 13 of the 77 participants, 17%, gave three responses for each of the 30 cues as requested and some participants gave only two responses for the majority of the cues. In all, there were 5,566 responses, 80% of a possible 6,930, and thus 1,364 missing responses. This could signal a lack of motivation to complete *LexCombi*, an issue discussed by Nation (2007) as a considerable problem for the validity and reliability of research instruments. For ethical reasons, the data was collected in the participants' own time, and it was made clear that participation was voluntary, that the data was being collected for research purposes with no bearing on grades, and that only those who wished to take part should do so. Nevertheless, since only seven potential participants chose not to take part, it is possible that some of those who participated felt some pressure to do so, and may have had limited engagement with the task. There was, however, no evidence of test fatigue, which might be expected if the participants were not fully engaged in the task. Dividing each participants' responses in two by taking the first 15 cues and the second 15, a Wilcoxon signed ranks test showed no significant difference between the number of responses given on the two halves: $z = -1.548$, $p = .122$.

Alternatively, the number of missing responses might reflect the participants'

proficiency. That is, participants may have at times given less than a full set of three responses for a particular cue simply because they could not think of three responses for that cue. Certainly, some of the cues received fewer responses than others, suggesting that participants lacked responses for these cues in particular. Four cues (DEATH, GOVERNMENT, LAW and VALUE) received an average of less than two responses per participant, whereas there was an average of 2.4 responses per participant for the 30 cues overall. It is also noteworthy, however, that even the cue that elicited the most responses, CAR, received only 2.8 responses per participant. A certain level of missing responses seems, then, to be normal.

Missing responses certainly affect *LexCombi* scores, since if a response is not given, there is no possibility of it being a scoring response, and there was a moderate correlation (Pearson's) between the number of responses given and *LexCombi* scores of .53 (significance not calculated since the number of responses data was not normally distributed). However, as the scatterplot in Figure 4.2 makes clear, a high number of responses did not necessarily lead to a high *LexCombi* score because not all responses were canonical and thus point-scoring.

Figure 4.2: Scatterplot of *LexCombi* scores and the number of responses.



Ongoing monitoring of the number of missing responses would therefore be merited, as would further consideration of the *LexCombi* cues. The issue of missing responses is returned to in Chapter 6, and the issue of *LexCombi*'s cues is pursued in Chapter 8.

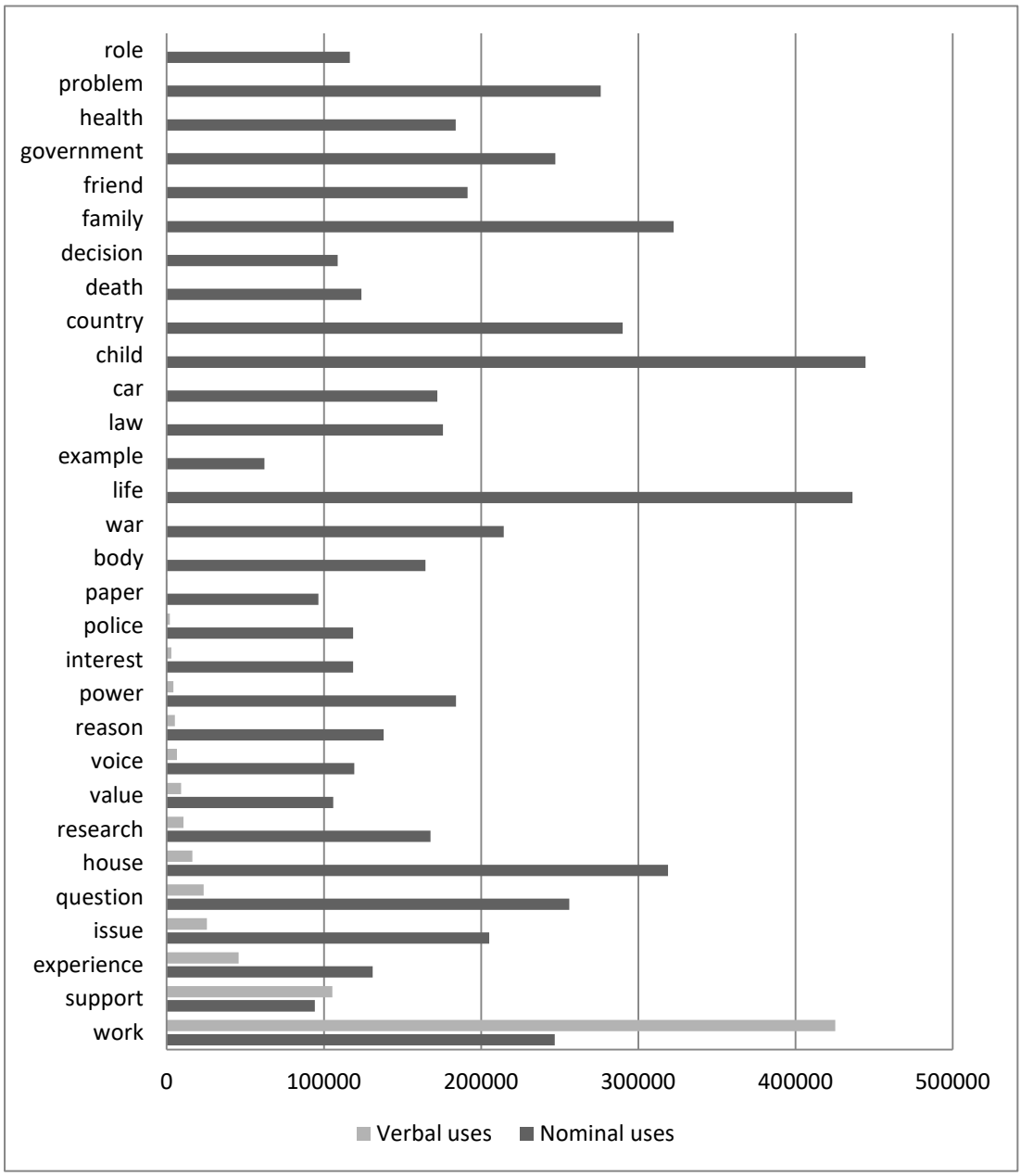
4.5.5 Cues interpreted as verbs rather than as nouns

A final issue is the word class of the cues, the question being what the impact is of cues being interpreted as verbs rather than, as intended, nouns. Barfield selected his 30 cues as nouns, but at least 14 also have regular verbal uses, and participants occasionally appear to have responded to them as verbs. This was observed to be an issue with the cues SUPPORT (with responses such as *him* and *somebody*) and WORK (*for*, *with*), and to some extent with EXPERIENCE (*something*) and RESEARCH (*something*). It is not, however, always easy to determine whether a participant's response invokes a nominal use or a verbal use of a cue. For example, a participant who gives the response *hard* for the cue WORK may have a verbal use in mind, *I work hard*, or a nominal use, *This is hard work*. Alternatively, the participant may have no particular use in mind and the response may have resulted from *hard* occurring with *work* in both ways along with others.

The *LexCombi* instructions used by Barfield and thus also adopted in this trial do not mention the cues being nouns. Potentially, therefore, revising the instructions could resolve this issue. However, it may be that certain cues are better known to learners as verbs than as nouns, and so a superior approach may be to replace these words with cues that are exclusively or almost exclusively used as nouns.

To gauge which of the cues may be problematic in this way, corpus searches were conducted for nominal and verbal uses of all 30 cues. Figure 4.3 presents the results of searches for the noun lemma and the verb lemma of each. There are no verbal uses (or only a very small number of seemingly erroneous uses) for many of the cues; a number have between 2% and 10% verbal uses; one (EXPERIENCE) has 25%; while for two cues (SUPPORT, WORK) verbal uses actually outnumber nominal uses. Thus these results largely coincide with the observations above. Interestingly, with regard to words which are regularly used as nouns and as verbs, Teddiman (2012) has found that L1 users have generally reliable intuitions about which use is more frequent. The differences in the participants' responses to cues

Figure 4.3: Frequency of nominal and verbal uses of *LexCombi* cues in the COCA (Davies, 2008-).



which are primarily used as nouns and those which also have significant verbal uses suggest that learners may also have intuitions in this area.

Despite the general correspondence between the corpus data and the way learners seem to have responded, for one cue there was not a correspondence: nominal uses of RESEARCH are much more common than verbal uses, yet participants sometimes seemed to think of verbal uses. This may be because RESEARCH was simply a difficult cue for many participants. RESEARCH received

an average of just over two responses per participant (fifth lowest among the 30 cues) and at the same time attracted 78 response types (third highest among the cues). This suggests that the participants as a group were not clear on the uses of RESEARCH and struggled to know how to respond to it.

A final question on this issue must be whether it matters if participants responded with nominal uses or verbal uses or to some amalgamation of the two. On one level it probably does not. Participants in most cases likely did not consciously consider the part of speech of the cue when responding to it, and might not do so even if the instructions were revised to mention that the cues are nouns. In the interests of getting as much information as possible about what learners know, perhaps any type of response should be welcomed.

On the other hand, there are some suggestions that nouns are central to collocation. Nesselhauf (2003, 2005), following Hausmann (1991, 1999), describes nouns as the starting point for verb + noun collocations, suggesting that the noun is selected first on the basis of its meaning and the verb is then selected on the basis of the noun. Indeed, for Nesselhauf, it is the very fact that verb selection is driven by the noun rather than by its usual sense that makes the combination a collocation. The introduction to the *LTP Dictionary of Selected Collocations* (Hill & Morgan, 1997) meanwhile states that "the noun is often the most important word in a sentence because it is 'what you are talking about'. The other words are built around the noun" (p. 8), and the *Oxford Collocations Dictionary* (2nd edition) similarly argues:

> When framing their ideas, people generally start from a noun. You might think of *rain* and want to know which adjective best describes rain when a lot falls in a short time. You would be unlikely to start with the adjective *heavy* and wonder what you could describe with it (*rain*, *breathing*, *damage*, *gunfire*?) Similarly, you might be looking for the verb to use when you do what you need to do in response to a challenge. But you would not choose *meet* and then choose what to meet (*a challenge*, *an acquaintance*, *your death*, *the expense*). (McIntosh, 2009, p. vi).

It may be important then to focus on nouns as the foundation for collocation knowledge.

Regardless, if *LexCombi* respondents provide responses that invoke verbal uses

of a cue, it is undesirable in a practical sense. The scoring of responses is based on a list of collocates for nominal uses of the cues, and thus responses that invoke verbal uses are unlikely to be deemed canonical. Respondents giving such responses are therefore likely to achieve lower scores than they might otherwise do so, and miss an opportunity to display the breadth of their collocation knowledge. One option might be to base the scoring lists on all uses of a cue, rather than restrict it to nominal uses, but this would complicate the construct of *LexCombi* and so will not be pursued. Instead, cues that show a tendency to invoke verbal uses of a cue will be considered problematic. This and other issues with Barfield's *LexCombi* cues are taken up in Chapter 8.

4.6 Taking *LexCombi* forward

This chapter has reported on an initial trial of *LexCombi* and explored how it performs with learners. Based on data from 77 Japanese L1 university students, *LexCombi* was found to be able to produce scores with a reasonable distribution and to differentiate participants. However, in examining and trialling *LexCombi*, a number of issues arose.

The first issue discussed was the scoring of *LexCombi*. The sources selected for the identification of canonical collocations were highlighted as one issue, as were decisions on the criteria used to define collocations. The effects of the current method of scoring *LexCombi* were then brought to light by examining responses that recurred frequently among the learners. The majority of these recurrent responses were found to be canonical. Those that were not, however, were also seen to be reasonable responses in many cases. It thus seems that a broader view of collocation may be required, and therefore the scoring of *LexCombi* will be examined in detail in Chapter 7.

Second, it was noted that multi-word responses occasionally appear to indicate that participants lack knowledge of how a response can be used in combination with a cue. If this is often the case, it raises issues of validity for *LexCombi* in that it becomes rather unclear what exactly *LexCombi* scores represent. Consequently, Chapter 5 will look into this issue in more detail.

Third, it was found that some participants on occasion gave responses that appear to be associations for the cues rather than collocations. If *LexCombi* does not

clearly guide participants to give collocations, as desired, participants may miss out on opportunities to display their knowledge of collocations. This issue will therefore be examined in greater depth in Chapters 5 and 6.

Fourth, there were around 20% missing responses, which may reflect the participants' motivation to complete *LexCombi* or possibly a lack of knowledge of collocates for certain cues. Evidence of test fatigue was not found, which perhaps points more to the latter, while it seems that a certain level of missing responses may be normal in that there were 5% missing responses even for the cue that received the most responses. It was also seen that while the number of responses correlated with a participant's *LexCombi* score, a high number of responses did not always result in a high *LexCombi* score since responses may or may not have been canonical. The issue of *LexCombi*'s cues will be pursued in Chapter 8, while the number of missing responses will continue to be monitored (see Chapter 6).

Finally, while Barfield selected his 30 cues as nouns, some have verbal uses and participants' responses can invoke verbal rather than nominal uses of the cues. Since the scoring of *LexCombi* is based on the nominal uses, this again means that some participants forgo opportunities to display their knowledge of collocations. A full examination of the *LexCombi* cues will therefore be undertaken in Chapter 8.

Chapter 5 How learners interact with *LexCombi*

5.1 Introduction

Chapter 4 presented an initial trial of *LexCombi* and discussed a number of issues concerning the instrument. This chapter reports two studies aimed at gaining further insights into some of those issues, and in doing so explores how learners interact with *LexCombi*.

In the first study (the full-phrase responses study), a modified version of *LexCombi* was used in which learners were requested to give full-phrase responses, rather than single words. As discussed in Section 4.5.2, one question in the initial trial of *LexCombi* was whether participants giving a response to a cue are able to use that cue-response combination. This is linked to the broader issue of task validity and the extent to which learners' performance on *LexCombi* mirrors their performance in free production.

Barfield's aim with *LexCombi* was for respondents to focus on the basic task of producing words that go together with the cue. By not requiring any indication of how the responses and cues are used together grammatically or semantically, his intention was to make the elicitation task simple and quick. This allows much more data to be collected in a given period of time, and enables data to be collected from learners with a wide range of proficiency levels. We saw in Section 4.5.2, however, that a participant may be able to give a response without a clear understanding of how their response and the cue are used together. If this is the case, it may be questioned whether *LexCombi*'s ultimate purpose, to elicit learners' productive knowledge of collocations, is fulfilled. As pointed out in Section 3.3.2, task validity is a key issue for elicitation instruments and *LexCombi* would be a more valid instrument if it can be shown that when learners give a response to a cue, we can reasonably assume that they have knowledge of how the two words are used in combination.

In Chapter 4's initial trial of *LexCombi*, it was found that participants frequently gave multi-word responses, and that these responses often include the cue itself. These multi-word responses including the cue allowed some insight into the issue of whether those participants were able to use responses and cues together. It was seen

that many participants apparently did have knowledge of the use of the cue-response combinations. However, in some cases, despite the response being canonical, the phrase given suggested that the participant's knowledge of the cue-response combination was limited. For example, in the response *have a decision right* for the cue DECISION, *right*, which was taken as the response, is canonical, yet the full phrase seems irregular (see Section 4.5.2).

The full-phrase responses study therefore sought to investigate the following question:

- Are learners giving a single-word response to a cue able to use that cue-response combination in a manner consistent with regular usage?

One way of approaching this question would have been to have learners complete *LexCombi*, and then have them demonstrate that they are able to use their responses with the cues. Unfortunately, since the time available for data collection was limited, this approach could not be taken. Instead, participants' knowledge of the use of the cue-response combinations was investigated by modifying the *LexCombi* instructions so as to elicit full-phrase responses rather than single-word responses to the cues. This modified form of *LexCombi* was administered to learners equivalent to those in Chapter 4. The intention was to generate phrases that could be judged for "regularity" (see Section 5.3.2). If those phrases were found to be largely regular, and if those phrases included within them many of the same single-word responses obtained in the Chapter 4 study, it could suggest that learners giving single-word responses, as in Chapter 4, are able to use the cue-response combinations.

The second study reported here (the think-alouds study) was concerned with the processes learners engage in when completing *LexCombi*. In the development of measurement tools, the importance of listening to respondents has been described as a crucial step, in that it helps developers to better understand the construct that the measure taps into, thereby assisting in establishing its validity (Wilson, 2005). This study was then concerned with the following question:

- What processes do learners engage in when producing responses to *LexCombi*?

There are a number of methods which may be used to seek insights into learners' thinking as they respond to test items, such as exit interviews and think-alouds

(Wilson, 2005), and the latter was chosen for use in this study. This was because the think-aloud procedure was felt to offer the potential of more direct access to the conscious processes employed by respondents as opposed to an after-the-fact report which may deviate towards description and justification. Thus, in this study think-aloud protocols were elicited from a small number of participants as they completed *LexCombi*. The think-alouds were conducted with an open mind as to what might be revealed, though with a particular interest in certain issues. One was the role of the L1 in producing responses; another was the issue outlined in Section 4.5.3 regarding learners who appeared to give some responses which are more general associations rather than collocations in particular.

5.2 Full-phrase responses: Method

Two instruments were used in this study: a modified version of *LexCombi* and a Yes/No vocabulary test. *LexCombi* was modified in two ways. First, the *LexCombi* instructions were altered so that, rather than "three words", participants were asked to write "three phrases that you most expect to see in conjunction with the word". In addition, short phrases, with the cue word included, were provided as examples, in place of the single-word examples in the original version. Participants also completed three practice items, allowing a further general check that full-phrase responses including the cue were being provided. The second change to the instructions was to state that all 30 cues are nouns. This change was made since, as reported in Section 4.5.5, some participants in the initial trial of *LexCombi* appeared to respond to some of the cues as verbs. In all other respects, the instrument, and its administration, was identical to that used in the Chapter 4 study: the same 30 cues were used in the same two variant orders, and the same time given, 30 seconds per cue, for participants to give up to three responses. Appendix B shows the modified instructions, which may be compared with the original instructions in Appendix A.

Yes/No vocabulary tests provide an estimate of vocabulary size. These tests present a list of words and respondents are required to mark each word they know or can use. The words are sampled from different levels of a frequency list to provide a broad representation of vocabulary. In addition, pseudowords are included to allow an estimation of the extent to which respondents may be guessing. Eyckmans, Van de Velde, van Hout and Boers (2007) raised concerns about the varying propensity

of learners to guess on Yes/No tests; Shillaw (1999), however, found that Japanese learners generally show little tendency to make guesses. In addition, any participant displaying a high level of guessing can be excluded.

Yes/No tests have been found to be highly reliable and have strong correlations with various aspects of proficiency (Milton, 2009; see also Section 10.3.1). The Yes/No test was used in this study as a proxy measure for general proficiency.

The test given was an *X_Lex* Yes/No test (Meara & Milton, 2003), the actual test being the first *X_Lex* test provided in Milton (2009, p. 254). This test includes samples of 20 words from each of the first five 1000-word frequency bands, along with 20 pseudowords, and is given in Appendix C. Since each real word on the test represents 50 words on the frequency list, the Yes/No scores are calculated by multiplying the number of real words marked by 50 and then making an adjustment for the number of pseudowords marked. Several different methods have been developed for making this adjustment (Mochida & Harrington, 2006), but the simple method recommended by Milton of subtracting 250 for each pseudoword marked was followed.

The participants were 40 Japanese university students who volunteered from two classes taught by myself. These two classes are a parallel cohort to two of the three classes used in the Chapter 4 study, the Chapter 4 study and the full-phrase responses study being conducted a year apart though with the same class streams used. This means the participants in the two studies were from different intakes, but were very similar as they were at the same stage in their university studies, were approximately the same age, were majoring in the same subjects, had similar academic backgrounds and were of similar proficiency levels. This close similarity supported the planned comparison of the phrases elicited in the current study with the single-word responses obtained in Chapter 4's study. The participants were between 18 and 20 years of age and were nearing the end of their first year of university. All were majoring in subjects other than English, though had at least two required English classes per week. The participants completed the Yes/No test at the end of a regular class period, and then completed the modified *LexCombi* the following week. Data from five participants were excluded: four did not fully complete one of the two instruments, and one showed a high level of guessing on the Yes/No test. The Yes/No test showed the remaining 35 participants to have an estimated mean

vocabulary size of 3,876 words (*SD* = 373), with scores ranging from 2,950 to 4,500. According to Milton's (2009, p. 181) guide to *X_Lex* scores and proficiency, the participants ranged, then, from elementary to advanced proficiency.

After briefly presenting the basic results for the current study, the analysis is divided into two parts. First, the extent to which the data in the current study may be comparable to that collected in the Chapter 4 study is considered. Second, the regularity of the responses is evaluated.

5.3 Full-phrase responses: Results

The 35 participants gave an average of 2.2 responses per cue, though none proved able to give three responses for all 30 cues as requested. In total, there were 2,303 responses, 73% of a possible 3,150.

Table 5.1 presents the *LexCombi* scores for the current study. These scores were produced as in Chapter 4: that is, the responses were reduced to a single word, following the procedures given in Table 4.1, and *LexCombi* scores calculated using the CollCheck (Imao & Brown, 2014) program on the basis of the OCD2 lists of canonical collocates.

Table 5.1: *LexCombi* scores (i.e. number of canonical collocates) in the current study.

	Current study
N	35
Mean	35.86
SD	9.334
Minimum	20
Maximum	55

Note. Maximum score = 90.

5.3.1 The impact of the modification

The first step in the analysis was to consider the extent to which participants' responses in the current study were similar to those in the Chapter 4 study. This was necessary because of the possibility that modifying the *LexCombi* instructions had altered the nature of the task, which was not the intention. Requiring participants to write phrases rather than single words within the same 30-second time limit may

have made it more difficult for learners to give a complete set of responses. On the other hand, writing phrases may have had an enabling effect on participants by encouraging them to think of concrete examples of how the cue words are used.

Any impact of the modification was checked for in three ways. First, the number of responses given was compared. This comparison was made between the participants in the current study and those participants drawn from the equivalent classes in the Chapter 4 study (N = 57), rather than the full complement of 77 participants in that study. Since a number of participants in the Chapter 4 study gave a full set of 90 responses, the distribution of the number of responses was not normal and so a Mann-Whitney U test was used. This revealed a significant difference in the number of responses between the two studies: Chapter 4 *Median* = 69, Current study *Median* = 65; $U = 751.5$, $z = -1.981$, $p < .05$.

Secondly, the scores achieved by participants in the current study and in the Chapter 4 study were compared. Table 5.2 gives the results for the current study (repeated from above), alongside those from the Chapter 4 study. An independent *t*-test was used to compare the scores and revealed no significant difference between them: $t(59.675) = -0.557$, $p = .58$.

Table 5.2: *LexCombi* scores in the current study and for the parallel classes in the Chapter 4 study.

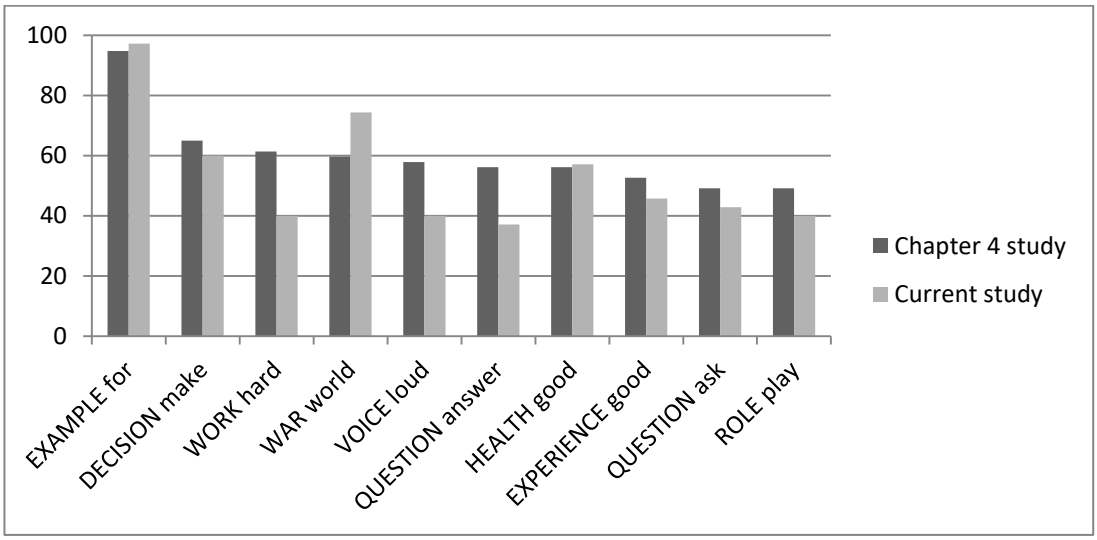
	Current study	Chapter 4 study
N	35	57
Mean	35.86	34.82
SD	9.334	7.366
Minimum	20	19
Maximum	55	50

Note. Maximum score = 90.

The final way in which the impact of the modification was judged was to look at frequently recurring responses. Figure 5.1 shows the 10 most frequently given responses from the Chapter 4 study and the proportion of participants who gave these responses. Alongside are the proportion of participants who gave these responses in the current study. As can be seen, while there were some differences, all

10 of the responses also appeared frequently in the current study's responses.

Figure 5.1: The 10 most frequently given responses in the parallel classes in the Chapter 4 study and their frequency in the current study.



Note. Bars represent the percentage of participants that gave each response. Caps show the cue, lower case the response.

The above was an attempt to establish whether the modifications made to *LexCombi* had an impact on the responses of participants. To sum up, (1) there were significantly fewer responses in the current study than in the Chapter 4 study, which could be due to the additional words that had to be written in the modified version; (2) there was no significant difference in the number of canonical collocates, which suggests that the underlying processes of selection were similar; and (3) frequently recurring responses were given by a broadly similar proportion of participants, which also suggests the same underlying processes of selection. It seems reasonable to conclude that the modifications had some impact on participants' responses, though not as to invalidate a more detailed comparison, as reported below.

5.3.2 Regular English usage

The "regularity" of the phrasal responses was evaluated using a combination of intuition, reference to the *Oxford Collocations Dictionary* (2nd edition) (McIntosh, 2009) and consultation of the COCA (Davies, 2008-). This analysis was conducted with those responses in the current study that were deemed canonical after being reduced to a single-word response. Each such phrasal response was categorised as

"regular", meaning an ordinary or usual use, or "non-regular".

Determining whether an instance of language is "regular" is, of course, not a simple matter. However, for a large proportion of the responses, it was felt that a cautious application of intuition was sufficiently reliable to consider a response "regular". This was the case, for example, for responses such as *ask a question*, *play a role* and *go to work* to the cues QUESTION, ROLE and WORK respectively. Only a small proportion of the responses required further deliberation. For each response, therefore, there was a two-step process: (1) Does intuition suggest this phrase is a regular use of English? If so, mark the response as "regular". If not, proceed to step two; (2) Does the OCD2 or COCA data bring to light uses that suggest the phrase may be a regular use of English? If so, mark the response as "regular". If not, mark the phrase as "non-regular". For example, the response *death to* to the cue DEATH initially struck me as somewhat odd, but the COCA revealed it to be quite common, often as the rabble-rousing cry of incitement *death to* + [a group of people], and so this phrase was categorised as "regular".

Step two of this process was not always simple, however. A phrase that does not appear in either the OCD2 or the COCA is not necessarily non-regular; and where phrases do occur in the COCA, there is no easy answer to the question of how many occurrences are required to deem a phrase "regular". The heart of the problem is that when designating an instance of language as "non-regular", what we are trying to do is find the lack of something, the absence of a phrase from the language, and it is not possible to find positive evidence that puts a response in this category. Recognising this fact, the consultation of other sources in Step two was a matter of making judgements, with the sources acting as support for those judgements. At times, the sources simply revealed regular uses that intuition had somehow overlooked, as with the *death to* example above. More difficult cases were responses that seem incomplete. One example was the response *have my car* for the cue CAR, which the COCA shows to occur within longer phrases such as *have my car* + past participle (e.g. *have my car stolen*) and *have my car* + prepositional phrase (e.g. *have my car nearby*). However, *have my car* is incomplete as a phrase by itself, and it is possible that the intended meaning here was *have a car*/*own a car*/*have my own car*. This particular example is also potentially linked to a Japanese coinage マイカー (*maikaa* = literally, *my car*), meaning *to have one's own car*, which is used in expressions such

as マイカー持っているの？ *maikaa motteiruno* = literally, *do you have my car?*), meaning *Do you have your own car?* Nevertheless, since it was not possible to interrogate the participants as to their intentions, and not wishing to second-guess participants' intentions, phrases such as these were marked as regular.

In working with the responses, it also became apparent that one can be stricter or more lax in categorising the responses. There were some responses which seemed essentially regular, but had some minor problem, most commonly missing or mistaken articles (e.g. *it is big problem*, *a important decision*). If the central issue is whether learners can use the substantive lexical item, rather than whether they have mastered function words and morphology, such problems should perhaps be overlooked. Thus, in the final analysis, the responses were classified as either "regular use", "regular use by lax standards" and "non-regular use". Table 5.3 gives the results. As can be seen, 84% of the responses were judged to be "regular use", or 94% when inconsequential errors were permitted.

Table 5.3: Regularity of phrasal responses which were canonical responses.

Response phrases of regular use	1,053	84%
Additional response phrases of regular use by lax standards	125	10%
Response phrases of non-regular use	77	6%
Response phrases deemed canonical when reduced to a single-word response.	1,255	100%

5.4 Full-phrase responses: Discussion

The full-phrase responses study set out to investigate whether learners giving a single-word response to a cue are able to use that cue-response combination in a manner consistent with regular usage, and the results above provide an insight into the regularity of participants' phrasal responses. To summarise, the analysis of 1,255 responses showed 84% to be "regular", or 94% by more lax standards. Thus, the vast majority of the canonical responses were regular. It was also seen that the responses to the modified version of *LexCombi* were broadly similar to those given by parallel

groups of learners in the Chapter 4 study, particularly in that responses recurring within each group of participants were given by a similar proportion of participants in each case.

It may therefore be supposed that the Chapter 4 participants, who mostly gave single-word responses to the cues, did have knowledge of how those responses and cues are used in combination. It should be acknowledged that this supposition depends on an analogy between the two sets of participants and the two sets of responses. Nonetheless, it does seem reasonable to suggest, on the basis of the regularity of the responses in the current study, that the Chapter 4 participants were probably able to use the cue-response combinations. This finding may be seen as a step towards demonstrating that learners' performance on *LexCombi* broadly mirrors their performance in free production, and substantiates the claim that *LexCombi* is a measure of learners' productive knowledge of collocations.

5.5 Think-alouds: Method

Individual think-aloud sessions were conducted with the original *LexCombi* format as used in Chapter 4, with a separate group of participants from those involved in the full-phrase responses study. The participants were six Japanese university students who volunteered from a class taught by myself. As before, they were aged 18 or 19 and were nearing the end of their first year of university. All were language majors, but in languages other than English, though they had two required English classes per week. Immediately prior to the think-aloud session, the participants completed the same Yes/No test as used in the full-phrase responses study (Appendix C). One of the six proved unable to complete the think-aloud task, and so was removed from the study. The five remaining participants had an estimated mean vocabulary size of 3,880 words (*SD* = 507) and the scores ranged from 3,200 to 4,400.

After the same explanation and practice as used in the Chapter 4 study, participants were asked to think-aloud while responding to the cues as follows:

> As you try to think of collocations, I would like you to think aloud. What I mean by think aloud is that I want you to say out loud everything that you say to yourself silently. You can speak in Japanese or in English or use both. I will record what you say. But please do not talk to me. Just act as if you are alone in the room speaking to yourself.

> If you are silent for any length of time, I will remind you to keep talking aloud. Do you understand what I want you to do?

These instructions were modelled on Ericsson and Simon (1984, p. 376), and were provided in writing in Japanese, while I, as the researcher, used both English and Japanese in introducing the tasks to further emphasise that both languages could be used. Following advice in Ericsson and Simon (1987) and Gass and Mackey (2000), participants first completed three cues as a warm-up and during the experiment itself were given the simple reminder "Keep talking" if they lapsed into silence. The think-alouds were recorded using a digital recorder.

The analysis of the think-aloud data followed standard qualitative content analysis procedures. Dörnyei (2007) describes four stages of qualitative content analysis involving transcribing, pre-coding and coding, growing ideas and interpreting the data, and these were the basic stages followed. Yet Dörnyei notes that these stages merge with each other and often take place concurrently, and this was found to be the case here. Richards (2003) describes the identification of categories as the central feature of qualitative inquiry, and the search for categories regarding how the participants generated responses was the primary goal.

The analysis proceeded as follows. The think-aloud recordings were transcribed, with the transcriptions subsequently checked by a L1 user of Japanese. The high quality of the recordings and the fact that all the participants, as is common in Japan, used a pencil to write their responses meant it was possible to hear on the recordings the sounds of the pencil on the paper and establish exactly when each response had been written, so this information was also incorporated into the transcriptions. Coding of the data took place in an iterative process whereby first, as the transcriptions were being made, I made notes of interesting features, and then, as I reviewed the transcriptions several times, added further notes while extending and developing the original notes. This led to the identification of three categories of how responses were produced (see Section 5.6).

The identification of the three categories was partly a deductive process and partly inductive. One point of interest in deciding to conduct think-alouds was the role of the L1 in generating responses, so uses of Japanese by the participants immediately drew attention. The second category explained below was also not an entirely new departure. While dealing with the Chapter 4 data, I had been struck by

pairs of antonyms given successively as responses. In repeatedly reviewing the think-aloud transcriptions, it became clear that this may be part of a broader grouping.

Having identified three categories, the entire set of transcriptions was then reviewed again, using both the audio recordings and the written transcriptions, and each response was placed in one of the three categories, though a number of responses remained unclassified.

5.6 Think-alouds: Results

The three categories identified are *Active use of the L1*, *Chaining* and *Spontaneous production of a response*. The number of responses produced in each way is shown in Table 5.4, and descriptions of the three categories follow.

Table 5.4: Number and percentage of responses (in brackets) in each category.

	Participant					Total
	001	003	004	005	006	
Active use of the L1	11 (24)	0 (0)	1 (1)	6 (10)	2 (2)	20 (6)
Chaining	3 (7)	2 (4)	3 (3)	0 (0)	4 (5)	12 (4)
Spontaneous production	13 (29)	29 (56)	78 (90)	38 (63)	77 (92)	235 (72)
Unclassified responses	18 (40)	21 (40)	5 (6)	16 (27)	1 (1)	61 (19)
Total	45	52	87	60	84	328

5.6.1 Active use of the L1

Four of the five participants at times seemed to use their L1, Japanese, to help generate responses. This is an example for the cue PROBLEM (Italicised words are Japanese, translations follow in brackets, square brackets show other comments):

Participant 005: Problem *wa* (= How about~?). Big problem [Writes *big*]. [6 seconds] *mondai o kaiketsusuru* (= solve a problem). Solve problem [Writes *solve*]. Solve.

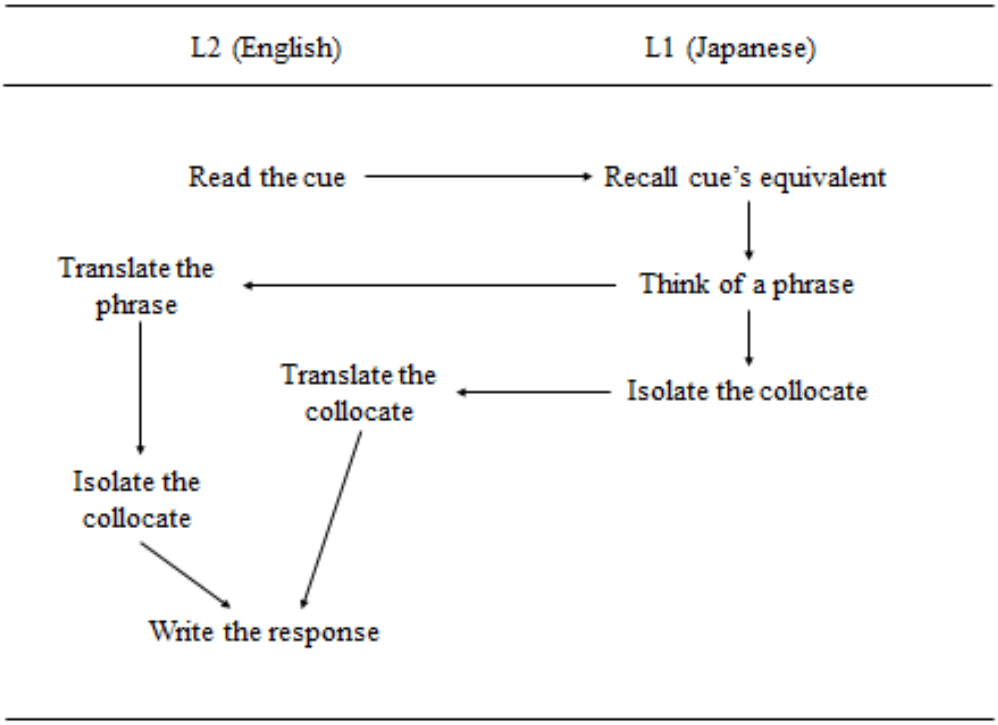
Here, after first giving one response, *big*, with no hesitation, the participant pauses for a considerable time before giving a Japanese phrase, which she then translates into English. Only this second response was classified as *Active use of the L1*.

It is of course not possible to know what the participants were doing on these occasions, but it can reasonably be inferred that the process leading to these responses was to read the cue, recall the cue's equivalent in Japanese, think of a collocation in Japanese, translate it into English, isolate the collocate and finally write the response. Not all these steps may be verbalised, however, and as shown in Figure 5.2, there are two possible points at which translation into English may take place.

The process can, moreover, break down, as seems to occur in this example for the cue GOVERNMENT:

Participant 001: Government. [5 seconds] *seifu* (= government), *seifu* (= government). [4 seconds] *seifu* (=

Figure 5.2: The assumed process of generating responses through active use of the L1.



government). *seifu o hihansuru* (= criticise the government). [2 seconds] *kougeki* (= attack). Attack [Writes *attack*].

Here, it appears that the participant first thinks of one Japanese collocation, *seifu o hihansuru* (= criticise the government), and then, realising that she lacks an English equivalent for the verb *hihansuru* (= criticise), comes up with a synonym *kougeki* (= attack) for which she does know an English equivalent.

As can be seen in the examples above, use of the L1 seems, for the most part, to be more an actively deployed strategy when the participant is struggling to produce responses, rather than a normal means of producing a response.

5.6.2 Chaining

Another way responses were produced was for chains of responses to be given; that is, one response prompted another. This most often occurred with opposites, as in this example for the cue VOICE:

Participant 001: Loud (Writes *loud*). [5 seconds] Um . . . loud *dakara* (= so) small [Writes *small*].

This participant immediately gives *loud* as a first response to the cue, and then moves to its opposite, in this context *small*, to generate a second response. Only this second response is classified in the *Chaining* category. Some of the responses placed in this category were not so explicitly linked by the participants, as in the following example for the cue CAR:

Participant 004: New car [Writes *new*]. Old car [Writes *old*].

Here, the participant makes no explicit link between the two responses, but from the audio recording they appear to be linked: "Old car" is said immediately after "New car" while the word *new* is still being written. The two responses sound very much like one unit. In addition to the total of 12 responses classified as *Chaining*, on eight other occasions the participants gave opposites as responses immediately following each other, but with some silence or other comments in between. Since on these occasions the two responses did not appear to be one unit, they remained in the unclassified category.

One participant also seemed to chain responses in a different way. For example,

for the cue DEATH:

> Participant 006: Death. Hmm. Suddenly [Writes *suddenly*]. Suddenly death is sad [Writes *sad*].

This participant seems to almost build a narrative, using her first response and the cue itself to generate a further response. Two other participants produced a pair of responses that seemed to follow on from one another in a similar way, but as these participants did not link the responses explicitly, these responses remained in the unclassified category.

5.6.3 Spontaneous production of a response

The third, and much more common, way for responses to be produced was for the response to seemingly occur spontaneously to the participant. For example, for the cues EXAMPLE and FRIEND:

> Participant 003: Example. For [Writes *for*].
>
> Participant 004: Friend. Best friend [Writes *best*].

Responses were placed in this category on the basis that there was no evidence of other processes occurring and due to the speed with which the response was given. Specifically, the responses in this category were produced after a silence of less than three seconds after the previous comment or action. Those responses that came after a silence of more than three seconds are included in the unclassified category. The cut-off of three seconds was selected after repeatedly listening to the think-aloud recordings and considering what appeared to be a significant gap compared to the usual rhythm with which the participants thought aloud. In fact, most of the responses in the spontaneous production category were given after no clear silence at all or a silence much shorter than three seconds.

It must be acknowledged, however, that the evidence on which responses were placed in this category is weaker than that for the other two categories. The reasoning was that a silence of more than three seconds suggests that the participants may have been engaged in other thought processes, but failed to verbalise their thoughts as requested. This is, however, not necessarily the case. Nor is it necessarily the case that a response after a silence of less than three seconds precludes the possibility of other thought processes being engaged in. Indeed, it may be

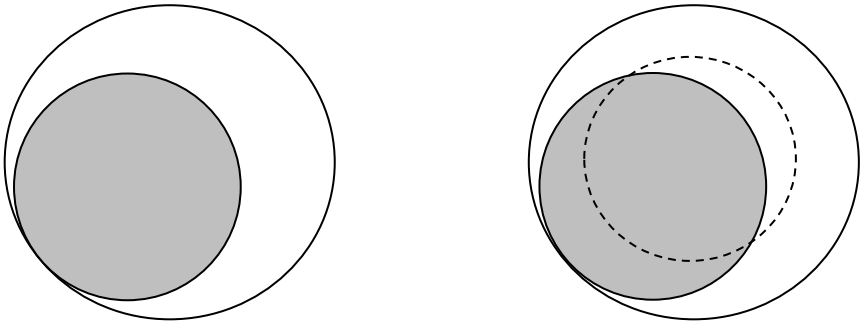
inappropriate to consider the *Spontaneous production* grouping a category at all. Richards (2003, p. 276), in discussing the practice of categorisation, describes four essential features of a category, one of which is that it is "analytically useful" which he explains as "When used, does it [the category] contribute anything to understanding? e.g. No, it's far too wide and too crude" (p. 276). It may be the case that this category is not "analytically useful", given that it contains over 70% of the responses and there is no information about what was actually taking place when these responses were produced.

Nevertheless, the spontaneity of these responses is interesting. While there is a possibility that other, unspoken processes were at times occurring, in many cases it is not clear what better explanation there is than that they were genuinely spontaneous. Furthermore, the majority of the responses in the *Chaining* category above were also spontaneous, the only difference being we can see what prompted the response in those cases. In essence, the responses could be divided into just two categories: those produced spontaneously and those produced through the conscious application of a strategy. It seems then that for a large proportion of the responses, the response was produced through automatic processes that were not available to conscious thought. Thus, the think-aloud procedure was unable to shed light on those processes.

5.7 Think-alouds: Discussion

This study sought to uncover the processes learners engage in when producing responses to *LexCombi*. Based on the above results, in the majority of cases, *LexCombi* seems to be eliciting responses from participants much like a traditional word association test is intended to; that is, spontaneous responses with little reflection or consideration. Indeed, this is not surprising since, although *LexCombi* is intended to elicit collocations, it is in fact identical in appearance to a word-association instrument, *Lex30* (Meara & Fitzpatrick, 2000), on which it was based. From a participant's viewpoint *Lex30* and *LexCombi* differ only in their instructions. This then raises the question of whether *LexCombi* is actually eliciting collocations or whether it stimulates a broader range of associations. In word association research, collocations are considered one type of association, so it could be that in responding to *LexCombi* participants are in fact giving associations for the items, some of which happen to be collocations. Figure 5.3 shows this line of thinking

Figure 5.3: Collocations (grey) as a subset of all possible word associations (left); Items elicited by *LexCombi* (broken circle), encompassing some collocations and some non-collocational associations (right).



diagrammatically. Certainly, in dealing with the Chapter 4 data, it appeared that some participants on some occasions had given non-collocational associations as responses, as was noted in Section 4.5.3.

Some evidence for the idea that *LexCombi* may elicit associations in general can be found in the think-aloud data. First, two participants indicated difficulties in focusing on collocations alone. Participant 003 did not verbalise these difficulties during the think-aloud itself, but afterwards reported wondering whether possible responses that occurred to her were collocations or not. Participant 001, responding to the cue COUNTRY, commented:

> Participant 001: Country. Country. *aa, taigigo shika omoitsukanai* (= Ah, I can only think of opposites). *uun* (= Hmm).

This participant then seems to recognise that antonyms are inappropriate responses.

There are also other instances of non-collocational associations occurring to participants in the protocols. In response to the cues RESEARCH and SUPPORT respectively, we find:

> Participant 006: *ato wa* (= What else)? Research. Researcher. Research. Uh, school. School [Writes *school*].
>
> Participant 005: Support. Support *nanika* (= what)? Support. Supporter. Support company [Writes *company*]. Support.

Both participants come up with a different form of the cue word, but do not seem to consider it a valid response and quickly move on. On another instance a participant explicitly rejects a form-based association:

> Participant 006: Interest. Um, interested in *wa dame nandayone* (= is no good). Interest. Book [Writes *book*].

However, there were also responses given by the think-aloud participants that appear to be non-collocational associations (Table 5.5). These responses are considered non-collocational associations for a variety of reasons. First, and most basically, none are canonical collocates. Responses (a) and (d) appear to be classic paradigmatic responses [1]. Response (b) is considered a form-based response. Responses (c), (e) and (f) appear to be conceptual associations. Finally, all the responses bar (f) were spoken separately from the cue; that is, they were not said alongside the cue as a phrase.

Table 5.5: Examples of responses that appear to be non-collocational associations.

	Participant	Cue	Response
(a)	001	WAR	peace
(b)	003	INTEREST	be interested in
(c)	004	ISSUE	soon
(d)	006	ISSUE	problem
(e)		GOVERNMENT	problem
(f)		LAW	sentence

In most cases, these responses were given without particular comment, but in two cases notable comments were made:

> Participant 001: War *dakara* (= so) . . . peace [Writes *peace*].
>
> Participant 006: Law. Law. Law. Law? *houritsu* (= law). Law sentence [Writes *sentence*]. *imi ga wakaranaiya* (= I don't know what it means).

Participant 001 appears to clearly give an antonym as a response, despite elsewhere

[1] Response (a) could alternatively be seen as a syntagmatic association deriving from the famous novel, but it was considered unlikely that the participant was familiar with the novel's English title.

recognising that antonyms are inappropriate, as was noted earlier. Participant 006 seems to indicate that she does not know what her response, *sentence*, means in this context; she does, however, seem aware that it has some conceptual connection with the cue.

To summarise, the think-aloud data suggests that non-collocational associations occurred to *LexCombi* participants as potential responses with some regularity. All five participants showed some evidence of this. Some of these ideas were, however, passed over by participants; generally, it would seem, form-based associations. Others, however, in most cases meaning-based associations, were written down as responses.

5.8 Discussion and conclusions

This chapter has reported on two studies that used different methods to explore how learners interact with *LexCombi*: asking learners to respond with full phrases rather than single words (the full-phrase responses study) and asking learners to think aloud as they responded to *LexCombi* in its original format (the think-alouds study). Through this exploration, two issues have been examined: first, whether learners giving a single-word response to a cue are able to use that cue-response combination in a manner consistent with regular usage, and second, what processes learners engage in when producing responses to *LexCombi*.

To re-cap the findings, regarding the first issue, it was seen that the vast majority of the phrasal responses exhibited regular use. Further, responses that recurred frequently in the Chapter 4 study were also found to be used by many participants in the full-phrase responses study. Given the very similar profiles of the learners in the two studies, it was suggested that the learners who gave these single-word recurring responses in the Chapter 4 study were aware of how these responses are used in combination with the cue. It was thus tentatively concluded that learners giving single-word responses to *LexCombi* do have a reasonable knowledge of how the cues and their responses are used in combination.

Regarding the second issue, the think-alouds suggested three different processes involved in producing *LexCombi* responses. One of these, active use of the L1, appears to be a strategy actively deployed by participants to generate responses. The other two appear to be less controlled. In some cases, it seems that the chaining of

responses takes place. However, by far the largest number of responses were apparently produced spontaneously, without any conscious processes involved. This raised the question of whether *LexCombi* elicits collocations, as desired, or associations in general, a good number of which happen to be collocations. There were a number of instances in the think-aloud data which suggest that the latter may be the case. To be clear, most of the participants most of the time appeared to be attempting to give collocates as responses. However, it was concluded that participants were sometimes giving associations more generally.

The point, then, is not that *LexCombi* fails to elicit collocations, it is that it may not exclusively elicit collocations. This is an issue because *LexCombi* asks for three and only three responses, and because there is a 30-second time limit. Those who give a non-collocational association thus waste an opportunity to display their knowledge of collocations.

Interestingly, the modified instructions for *LexCombi* used in the full-phrase responses study (which asked participants to give a phrase rather than a single word) appeared to guide participants more concretely towards giving syntagmatic responses. However, the trade-off was messy data which it is difficult to deal with. The procedures for reducing responses to single words (Table 4.1) are an attempt to handle such data in a systematic and consistent way, but there must be a question as to how changes to these procedures could affect the results. The original *LexCombi* format does then have the considerable advantage of efficiently eliciting a large quantity of data in a format more easily analysable. What would be ideal is a format that preserves these advantages while more definitively eliciting collocations specifically. One format that may achieve this is presented in Figure 5.4. This format would seek to elicit single words from participants as in the original *LexCombi*, but requiring participants to place the response either before or after the item would, it is hoped, cause them to think syntagmatically. That is, participants would be encouraged to think about the linear arrangement of words in text and this should result in collocational responses being provided. Chapter 6 seeks to take forward this proposed format and trial it.

Figure 5.4: A possible alternative format for *LexCombi*.

Write down three collocations for each of these words. Write your response either before or after the item as appropriate.

_____ house _____	_____ house _____	_____ house _____
_____ issue _____	_____ issue _____	_____ issue _____
_____ research _____	_____ research _____	_____ research _____
_____ power _____	_____ power _____	_____ power _____
_____ question _____	_____ question _____	_____ question _____
_____ voice _____	_____ voice _____	_____ voice _____

Chapter 6 Revising the test format

6.1 Introduction

This chapter seeks to determine whether a proposed adapted format of *LexCombi* (see Section 5.8) is more effective in eliciting collocations from learners of English than the original format. The chapter reports on a comparison between data collected with the adapted *LexCombi* format and data previously collected with the original *LexCombi* format (reported in Chapter 4).

This trial of the adapted format was motivated by a concern that the original format did not sufficiently guide participants towards providing collocates for the cues as opposed to associations in general. In Chapter 4, reporting on my initial trial of *LexCombi*, it was noted that some responses seemed to be associations of the cues, rather than collocations. For instance, for the cue FAMILY, one participant responded with *mother*, *father* and *brother*, which, as hyponyms of the cue, have strong semantic relations with it, but are not collocates of it.

Chapter 5 presented further evidence that participants may at times provide associations as responses, instead of collocations specifically. An analysis of think-aloud protocols from learners completing *LexCombi* first found that the majority of responses were provided quickly and without intervening comments, which could indicate a spontaneous direct link from cue to response, without the use of conscious processing. Thus, *LexCombi* appeared to elicit responses in a manner akin to word association. Second, all five think-aloud participants at times generated non-collocational associations, and though some of these were rejected by the participants as unsuitable, some were written down as responses.

The above findings are perhaps unsurprising when one considers, as mentioned previously, that *LexCombi* was explicitly modelled on a word association test: *Lex30* (Meara & Fitzpatrick, 2000). Collocations are of course one major category of associations, so *LexCombi* does allow the collection of collocational data. The key point is that simply asking participants to produce collocations may not be enough to elicit response sets consisting exclusively of collocations.

Chapter 5 therefore proposed that *LexCombi* be adapted so that the design of the instrument, rather than the instructions alone, drive the elicitation of the desired data.

LexCombi in its original format is identical in appearance to *Lex30*, with only the instructions differing between the two. To assume that *LexCombi* exclusively elicits collocations from learners places a lot of faith in the instructions, and this is an especially pertinent issue given that many learners are unfamiliar with the term "collocation".

The importance of task design, rather than task instructions alone, is shown by Peters (2009). In her study of vocabulary learning, two groups of learners were given different task instructions, while receiving identical materials. The intention was that the groups would attend to different aspects of the materials, but the study found no significant differences in vocabulary learning between the groups. Peters also collected qualitative data on learners' strategy use in completing their tasks, and this showed that the learners in the two groups seemed to have approached their tasks in the same way, despite receiving different instructions.

A further issue with instructions for experimental tasks is the possibility that, while some participants attend carefully to the instructions and follow them faithfully, others are less inclined to do so. An instrument which guides learners in a particular direction through its design is therefore superior in ensuring that more participants are consistently performing the task as desired.

Thus, it was thought that an adapted format for *LexCombi*, designed to guide learners towards producing collocations, may have advantages. The adapted format, along with the original format, is shown in Figure 6.1. The adapted *LexCombi* design directs participants to enter each response either to the left or to the right of the cue. The intention was that this would encourage participants to think in terms of how words follow one another in space (in written language) and time (in spoken language) and thus guide participants towards providing collocational responses in particular. This intention was reinforced by adapting the explanation of collocations given in the instructions from the original "words that you most expect to see together with the word if you were to read it" to "something that would be used either before or after the word".

The trial of the adapted format was therefore intended to answer the following question:

- Is the adapted LexCombi format more successful in eliciting collocations from learners than the original format?

Figure 6.1: The original format of *LexCombi* (above) and the adapted format (below).

house	1	2	3
issue	1	2	3
research	1	2	3
power	1	2	3

_____ house _____	_____ house _____	_____ house _____
_____ issue _____	_____ issue _____	_____ issue _____
_____ research _____	_____ research _____	_____ research _____
_____ power _____	_____ power _____	_____ power _____

6.2 Method

Due to the limited availability of participants, it was not possible for a single group of participants to complete both the adapted *LexCombi* format and the original format. Instead, the two formats are compared by using data collected in my Chapter 4 study, which used the original format of *LexCombi*, and new data collected for this study using the adapted format.

The new data set consists of responses to the adapted *LexCombi* and a Yes/No vocabulary size test. Other than the change to the format and the explanation of collocation described above, the adapted *LexCombi* was kept as similar as possible to the original to enable their comparison. The same 30 cues were used, presented in the same two variant orders, the same time-limit of 30 seconds per cue was used and the same examples were provided. The instructions for the adapted format were also kept as similar as possible to the original. The only changes made were: (a) to instruct participants to "Write each collocation next to the word, either to the left or

the right as appropriate," (b) to illustrate how to complete the items, and (c) to inform participants that all 30 cues are nouns. The complete instrument is shown in Appendix D.

The Yes/No test was the same as in Chapter 5 (see Appendix C). In the complete data set that was collected (involving 134 participants, as explained below) there were some signs of ceiling effects for the Yes/No test. This is not an issue, however, for the portion of the data set examined in this chapter.

The Chapter 4 data set consisted of responses to the original *LexCombi* from 77 Japanese university students drawn from three classes taught by myself. For the purposes of making comparisons with the adapted format, only participants from two of those three classes were used. These 57 participants were between 18 and 20 years of age and were nearing the end of their first year of university. All were majoring in subjects other than English, but had two required English classes per week. Data collection took place at the end of regular class time, in the final week of the academic year.

Data for the adapted format of *LexCombi* came from a total of 134 participants from six classes. Nine participants were excluded, one because the tasks were not fully completed and eight due to high levels of guessing on the Yes/No test, leaving 125 participants. For the present analysis, only the data from two of the six classes is used, these two classes being parallel to the Chapter 4 classes (the full data set including all 125 participants is made use of in Chapters 8 and 10). After removing participants as explained above, there were 47 participants from these two classes. The participant demographic for these participants was very similar to the Chapter 4 participants: the participants on both occasions being similar in age, and being from similar backgrounds. Data collection again took place at the end of regular class time in the final two weeks of the academic year. On the first occasion, the purposes of the research were explained and learners were invited to participate. Participants then completed the Yes/No vocabulary test. On the second occasion, participants completed the adapted *LexCombi*.

The Chapter 4 participants and the participants providing new data for this current study are parallel in that the classes from which the participants were drawn are the same streams in the university's system and from the same academic departments. In each year the students had been placed in those classes by the same

internal university placement test (it was unfortunately not possible, for ethical reasons, to obtain the university placement test scores for the two groups of participants), and the data was collected from the learners at the same point in their university studies. The sole difference between them is that they are different intakes of students, the 2009 intake and 2011 intake respectively. The proficiency of the 47 participants from the new data set may be estimated by the Yes/No vocabulary test scores. Their estimated mean vocabulary size was 3,805 words (*SD* = 398). The Chapter 4 participants did not complete the Yes/No test, so no estimate of their vocabulary size is available. However, as a teacher of these four classes, I judged them to be very similar in terms of proficiency.

Section 6.3 first presents the descriptive statistics for the adapted format of *LexCombi* and then reports two comparisons of the original and adapted *LexCombi* formats. The first comparison was of the number of canonical collocates given by participants. To enable this, the adapted *LexCombi* data was treated in the same way as the Chapter 4 data: that is, the responses were typed up in full and cleaned up in accordance with the procedures in Table 4.1; then *LexCombi* scores were calculated using the CollCheck (Imao & Brown, 2014) program to check each response against the OCD2 lists. The second comparison of the two formats treated the responses as word association data and applied a word association classification scheme to the data in order to determine whether there was a difference in the number of responses classified as collocational.

6.3 Results

The 47 participants who completed the adapted format of *LexCombi* gave an average of 2.5 responses per cue, with seven (15%) giving three responses for all 30 cues as requested. There were 3,554 responses in all, 84% of a possible 4,230. The *LexCombi* scores for these participants are presented below (see Table 6.1).

6.3.1 Comparing the formats: *LexCombi* scores

The two formats of *LexCombi* were first compared in terms of the number of canonical responses given by participants completing each format. The *LexCombi* scores for the two sets of data are shown in Table 6.1. The scores were very similar, and a *t*-test found no difference between them: $t(102) = -0.296$, $p = .77$.

Table 6.1: Descriptive statistics for the original *LexCombi* scores and adapted *LexCombi* scores.

	Original *LexCombi* format	Adapted *LexCombi* format
N	57	47
Mean	34.82	34.38
SD	7.366	7.815
Minimum	19	14
Maximum	50	51

Note. Maximum score = 90.

6.3.2 Comparing the formats: A word association classification

In the second comparison of the two formats, the responses were treated as word association data and a word association classification scheme was applied. This comparison of the two formats involved taking a different perspective from before on how participants may approach *LexCombi*. In calculating *LexCombi* scores (as presented in Section 6.3.1), we began with the assumption that participants had provided collocations as responses and then asked how many were canonical. This second comparison had a different starting point, treating the responses as word associations and asking how many could be classified as collocational.

For this second comparison, then, two raters applied a word association classification scheme (Fitzpatrick, 2007) to a sample of the responses from the Chapter 4 data and from the adapted *LexCombi* data. Samples, rather than the full data sets, were used since the full sets were of unequal size and because the full sets (over 9,000 responses) would have made the task unfeasibly large for the raters. The samples were created as follows: first, from the Chapter 4 data set, I took the responses to 5 cues from 6 randomly chosen participants; then, the responses to 5 more cues from another 6 randomly chosen participants were selected, and so on for all 30 cues. This meant that for each individual cue there were up to 18 responses (6 participants × 3 responses each), and the responses of 36 participants (6 participants × 6 groups of 5 responses) contributed to the sample. The sample consisted then of 540 responses (18 responses × 30 cues), 10.5% of the full data set. This same process was repeated for the adapted *LexCombi* format data set, producing a second sample of 540 responses, this time representing 12.8% of the full data set. In total

then, the two raters classified 1,080 responses.

Fitzpatrick's (2007) word association classification scheme (Table 6.2) was used since it was judged to be comprehensive and relatively easy to apply while crucially separating collocational, here called position-based, associations from other types.

Table 6.2: Fitzpatrick's (2007) word association scheme.

Category	Subcategory	Definition
Meaning-based association	Defining synonym	y means the same as x e.g. SUPPORT *help*
	Specific synonym	y can mean x in some specific contexts e.g. REASON *cause*
	Lexical set/context related	x and y belong to the same lexical set/are coordinates/meronyms/superordinates/provide context e.g. CAR *vehicle*; CAR *train*; CAR *engine*; CHILD *adult*
	Conceptual association	x and y have some other conceptual link e.g. DEATH *sad*
Position-based association	Consecutive xy collocation	y follows x directly (includes compounds) e.g. HOUSE *work*
	Consecutive yx collocation	y precedes x directly (includes compounds) e.g. VOICE *loud*
	Other collocational	y follows/precedes x in a phrase with word(s) between them e.g. PAPER *piece*
Form-based association	Change of affix	y is x plus or minus an affix e.g. INTEREST *interested*
	Similar form not meaning	y looks or sounds similar to x but has no clear meaning link; or y is an associate of a word with a similar form to x e.g. POWER *tower*; LAW *salary* (*salary* is an associate of *low*, similar in form to LAW)
Erratic association	No link/blank	y has no decipherable link to x or no response was given e.g. VALUE *foods*

Note. x = the cue; y = the response. In the examples, the cue is in caps, the response in italics.

The training of the two raters involved first presenting the classification scheme as shown in the table. In this table, the categories and definitions are taken from Fitzpatrick, while the examples are drawn from *LexCombi* data outside of the samples. The raters were then given a further set of examples, also from outside the samples, on which to try out the classification scheme, before being shown my own classification of those examples and given an opportunity to ask questions.

The responses for the two formats were transcribed and presented to the raters in identical format, with the data from the two conditions combined and appearing as a single data set. The raters were told that the responses were word associations and were not informed that the data was elicited with two different formats of an instrument. After the rating task, the raters were debriefed on these points. Each response was presented exactly as the participant had given it. The raters were asked to ignore obvious misspellings and where a multi-word response had been given, the entire response was shown with the word to be classified underlined. The raters used the 10 sub-categories to classify the responses, since it was felt that focusing on the more tightly defined sub-categories would make the task easier. The raters reported some difficulty in classifying responses as meaning-based or position-based. Specifically, they reported difficulty with a small minority of responses which on the one hand were conceptually related to the cue, while on the other hand, though not seeming to be especially strong collocates, could co-occur with the cue. This difficulty with word association classification has long been noted (Meara, 1983; Wolter, 2001). Nevertheless, there was a high level of inter-rater reliability, with the two raters agreeing on the classification 87% of the time, with a Cohen's kappa of .71. Responses which received conflicting categorisations were resolved by myself, again while blind to which data set the response was from.

Table 6.3 shows the results of the raters' classification of the two samples of responses. A Pearson chi-square found an overall significant difference between the two *LexCombi* formats: $\chi^2(2) = 19.421$, $p < .001$. The standardised residuals revealed that among the three categories, there was a significant difference only for the number of missing responses. Thus, while the adapted *LexCombi* format elicited slightly fewer non-position-based responses and more position-based responses, these figures did not reach significance.

Table 6.3: Distribution of responses rated as position-based or otherwise.

	Responses rated as position-based	Responses rated as non-position-based	Missing responses
Original *LexCombi* format	354	60	126
Adapted *LexCombi* format	417	48	75

6.3.3 Multi-word responses

As Section 4.2 noted, the original *LexCombi* format elicited a considerable number of multi-word responses, despite Barfield's intention being to elicit single-word responses. In working with the data collected with the adapted *LexCombi* format, it became evident that it had encouraged participants to give single-word responses, as desired, more often. In the same samples used above, consisting of 540 responses each from the Chapter 4 data and the data collected for this study, there were 133 multi-word responses in the Chapter 4 sample, but only 37 in the sample from the current data, a significant difference ($\chi^2(1) = 64.339$, $p < .001$).

6.4 Discussion

The question that the study reported in this chapter sought to answer was whether the adapted *LexCombi* format is more successful in eliciting collocations from learners than the original *LexCombi* format. The motivation for adapting the format of *LexCombi* was a concern that, when responding to the original format, learners may fail to comprehend, forget or simply ignore the instruction to provide collocations and may instead provide associations more generally. The question was addressed by comparing data elicited with the original *LexCombi* format (reported in Chapter 4) with data elicited with the adapted *LexCombi* format. Two different groups of participants were, then, involved. The two sets of data were compared by examining the number of canonical responses (i.e. the *LexCombi* scores, based on the OCD2 lists of collocates) given to each *LexCombi* format, and by looking at how two raters viewed responses to each format in terms of a word association classification scheme. The first comparison operated on the assumption that participants were acting as requested and attempting to provide collocates in response to the cues. The

second comparison made no such assumption and instead treated the responses as word associations, since the original *LexCombi* format does not itself steer participants towards providing collocate responses.

In terms of the first comparison, the canonicity of the responses, no difference was found between the two formats. It should be noted, however, that all four of the classes from which participants were drawn had had considerable exposure to the idea of collocation and to the learning of collocations through the year-long English courses they were just completing when data collection took place. These participants may have been better informed regarding collocation than many learners and were thus perhaps more likely to understand and follow the *LexCombi* instructions successfully.

The second comparison, in contrast, did find a difference between the two formats. Firstly, the adapted *LexCombi* elicited significantly fewer missing responses. Secondly, when the responses were categorised, the balance was towards somewhat fewer non-position-based, and more position-based, responses though the difference was not statistically significant. The categorisations were of course based on human judgement, but, as noted above, the degree of inter-rater reliability between the two raters was high .

One explanation for these findings may be that, in responding to *LexCombi*, individual tendencies led some participants to move away from the *LexCombi* instructions regardless of the format. Fitzpatrick (2007, 2009) has found that, in giving word associations, individuals appear to favour certain types of associations, with some individuals predominantly giving associations judged to be position-based, for example, while others tend to give associations judged to be meaning-based.

It should also be emphasised that, with both the original and adapted formats, the majority of responses were rated as position-based (Table 6.3). As said before, responding with other types of associations appeared to occur only with some participants, some of the time.

The experiment reported in this chapter indicated, therefore, a difference between the two formats under one comparison, but not under the other. One potential explanation for this difference between the results of the comparisons might be that there was a quirk in the sample used in the second comparison. This

does not, however, appear to be the case: in terms of *LexCombi* scores, each sample was similar to the full data set from which it was drawn. An alternative interpretation of the findings regarding the two formats may be that the physical design of the adapted format, by encouraging learners to think about which words can occur either before or after the cue, makes the task somewhat easier for learners: hence the significantly smaller number of missing responses. This increase in the number of responses did not, however, lead to a greater number of canonical collocates being provided. That is, the additional responses elicited by the adapted format were mostly position-based, but were not canonical collocates.

6.5 Choosing a format

This chapter has investigated whether an adapted format of *LexCombi* is superior to the original format. The adaptation of the format was motivated by the observation that some participants, some of the time, seemed to respond with associations generally rather than collocations in particular. To move forward with *LexCombi*, a choice must therefore be made between the two formats.

The results above showed that there was no significant difference in the number of canonical collocates elicited by each format. Similarly, in the analysis of the raters' classifications, there was no significant difference in the number of unwanted non-position-based associations.

There were though some differences between the two formats. First, the adapted *LexCombi* format resulted in significantly fewer missing responses. Second, there was a tendency, admittedly non-significant, for the adapted format to elicit more responses rated as position-based under a word association analysis. Third, the adapted format appeared to guide participants towards providing single-word responses as desired, rather than multi-word responses.

It should also be borne in mind that the participants in both conditions in this study were probably more aware of the concept of collocation than many learners. For learners unfamiliar with the concept of collocation, the adapted *LexCombi* format may prove more successful in that it makes explicit the desired syntagmatic nature of the responses with respect to the cue.

It can be concluded that the original *LexCombi* format has no particular advantages over the adapted format, while the adapted format does have some

advantages. Consequently, the adapted *LexCombi* format is considered superior to the original format, and will be used hereafter.

Having decided upon the format of *LexCombi*, the next issue that must be addressed is the scoring of *LexCombi*: that is, deciding how best to determine whether learners' responses to *LexCombi* are collocates or not. Chapter 7 describes how this issue was dealt with.

Chapter 7 Revising the scoring of responses

7.1 Introduction

The previous chapter considered an adapted format for *LexCombi*, prompted by a concern that it was eliciting associations in general rather than collocations in particular (see Section 4.5.3). Another issue raised in the initial trial of *LexCombi* was its scoring (Section 4.5.1). As explained in Section 4.2, in order to score *LexCombi*, Barfield (2009a) created a list of canonical collocates for each cue and then scored responses against these lists. Barfield's lists were compiled from two sources: the first edition of the *Oxford Collocations Dictionary* (OCD1) (Crowther, et al., 2002) and *Collins Wordbanks Online* (HarperCollins, 2004). In my own work thus far with *LexCombi*, I have taken the simpler approach of using a single source: the second edition of the *Oxford Collocations Dictionary* (OCD2) (McIntosh, 2009). Section 4.5.1 raised three concerns about the use of these dictionary and corpus sources.

To recap, the first issue is the appropriateness of the sources used in creating the lists. Although the OCD1, the OCD2 and *Collins Wordbanks Online* each drew on a sizeable corpus, in each case the corpus consisted, solely or primarily, of texts from British and American sources. As noted earlier, a number of commentators have critiqued the "native-speaker model" for language learning (G. Cook, 1998; V. Cook, 1999; Widdowson, 2000, 2003), and there have been calls for its replacement by a model based on English as a Lingua Franca (ELF) or English as an International Language (Mauranen, 2011; Seidlhofer, 2005). At the same time, scepticism has been expressed about its replacement, because of the difficulty of describing or even identifying an alternative (Nesselhauf, 2005). Decisions on which sources should be used certainly require careful deliberation.

The second issue is that the exact criteria used in creating the lists, and the reasoning behind the use of these criteria, are somewhat vague. When a researcher uses a collocations dictionary, the difficult issue of defining collocation is essentially delegated to the dictionary's editors, and yet the limited details provided by the dictionaries on their compilation means that there is not complete clarity regarding the decisions that were made or the bases for those decisions. In using his second

source, *Collins Wordbanks Online*, Barfield had to make decisions regarding which measure of collocation to use and how to treat the results. Barfield chose to use *t*-scores and followed convention in setting the threshold for *t*-scores at 2.0. Yet Evert (2008) argues that statistically there is no justification behind this convention (see Section 7.2.2).

The third issue is that the above methods may be too narrow in scope. Section 4.5.1 examined responses to cues which had been given by at least 10% of the learners in the study, but which were not canonical collocates. To summarise the findings, some of these recurring responses revealed limitations in the OCD2 (e.g. the responses *my* for the cue FAMILY and *why* for the cue REASON do not appear in the dictionary, despite high frequency and high MI scores in corpora, since the OCD2 lists only verbs, adjectives, nouns, prepositions and phrases for noun entries); some appeared to reflect the participants' exposure to English in a Japanese environment (e.g. the response *note* for the cue DEATH likely stemmed from the popularity of *Death Note* (Japanese: デスノート *desu nōto*), a manga and film series); some were component parts of compound words (e.g. *news* for the cue PAPER and *man* for the cue POLICE); and some, in terms of L1-user norms, were simply non-standard (e.g. *keep* for the cue HEALTH). In sum, the recurring responses that were not canonical collocates on the basis of the OCD2 often seem to be reasonable responses. It may therefore be advisable to adopt a wider definition of collocation in order to gain a fuller appreciation of learners' knowledge.

One further point about Barfield's study is that no definition of collocation was offered; there was only an implicit definition that emerged from the choices made with regard to *LexCombi*'s scoring. An explicit definition gives a firm basis for decisions to be made about scoring. In Chapter 3, this thesis adopted Durrant's (2014) definition of collocation as:

> "combinations of two words that are best learned as integral wholes or independent entities, rather than by the process of placing together their component parts, either because (i) they may not be understood or appropriately produced without specific knowledge, or (ii) they occur with sufficient frequency that their independent learning will facilitate fluency" (p. 448).

Going forward, therefore, any scoring approach adopted should be in coherence with

this definition. This means the scoring approach needs to accommodate both of Durrant's elements: collocations as defined by specialised uses of words and collocations as defined by frequency.

The overall question guiding the work reported in this chapter was:

- How should the canonicity of learners' responses to *LexCombi* be determined?

The chapter reports on my exploration of a number of approaches to scoring the adapted *LexCombi*: that is, determining whether each response is a collocate of the cue or not. It was not expected that the problems outlined above could be completely overcome. Rather, the intention was to trial a range of approaches, to be explicit about the criteria used, to consider the appropriateness of the approaches for learners and to take a wider view of collocation so that a more complete view of learners' productive knowledge of collocation could be obtained. This chapter first introduces the various approaches used (Section 7.2), before providing a detailed comparison of them. The approaches were compared in terms of:

- which words they do and do not consider to be collocates for each cue (Section 7.3);
- the scores produced when applied to responses from learners and their coverage of recurrent responses (Section 7.4);
- their measurement properties, as judged by a Rasch analysis of *LexCombi* scores (Section 7.5);
- their assumptions about the concept of collocation (Section 7.6).

The aim of these comparisons was to elucidate the characteristics of each approach, both in terms of their practical application and their theoretical implications, and ultimately to choose one scoring approach as the primary means of scoring the adapted *LexCombi*.

This exploration of *LexCombi* scoring was carried out concurrently with the selection of a new set of cues for *LexCombi*, with this cue trialling and selection work being reported in Chapter 8. Consequently, the comparisons of the scoring approaches are based on 70 potential cues for *LexCombi* which were under trial.

7.2 Scoring approaches

As seen in previous chapters, two possible approaches to scoring *LexCombi* are to:

(1) have judges evaluate each response as a collocate or otherwise; and (2) use a list of canonical collocates. The former approach may have an advantage over the latter in that responses that are acceptable but somehow missing from a list may be dealt with, but it also has a number of disadvantages. First, from a practical perspective, having judges evaluate responses is not scalable. That is, while it may be viable for a small study, once a larger number of participants are involved, it quickly becomes unmanageable. For example, a study with 100 participants would produce up to 9,000 responses. Second, there would be problems with reliability. In the unlikely case that a single judge were able to deal with all the responses, their judgements would likely not be fully consistent due to human error and the complex nature of language. In the more probable case that multiple judges were used, the difficulty of ensuring consistency in judgements of collocation would be all the greater.

In this study, therefore, the latter approach was taken. The issue then was the method used to produce the lists of canonical collocates. Four primary methods were trialled, along with two ways of combining these methods. The four primary methods are:

- dictionary-based;
- corpus-based;
- L1-user norms; and
- L2-user norms.

The two combinations are termed:

- multiply listed; and
- all lists combined.

7.2.1 Dictionary-based lists of collocates

Collocations dictionaries provide a particular view of collocation, and a number of previous studies have made use of them (e.g. Barfield, 2009a; Laufer & Waldman, 2011; Nesselhauf, 2005; Wang & Shaw, 2008). Three dictionaries of collocations were used and the collocates listed combined into a single list for each cue: that is, a collocate included in any of the three dictionaries was considered canonical. Since *LexCombi* asks participants to give single-word responses to cues, where the dictionaries give multi-word rather than single-word collocates (e.g. one dictionary has *go by* as a collocation for the cue NAME), each component word was listed

separately as a canonical collocate (i.e. *go* and *by* were each included in the list). The three dictionaries used were the *BBI Combinatory Dictionary of English* (BBI) (Benson, et al., 2009), the *Frequency Dictionary of Contemporary American English* (FDCAE) (Davies & Gardner, 2010) and the second edition of the *Oxford Collocations Dictionary* (OCD2) (McIntosh, 2009). These dictionaries were selected since they are current, collectively provide coverage of both British and American English, and include all 70 of the potential cue words. The dictionaries also provide contrasting approaches to collocation, as the information each gives on its compilation (see below) demonstrates. This was expected to result in lists which are quite wide-ranging and inclusive, in order to capture as fully as possible learners' productive knowledge of collocations.

The BBI reports that its collocations come:

> from the authors' intuition, supported by their reading of and listening to contemporary English and by consultation with . . . valued colleagues . . . Nowadays, our task is eased not only by the availability of corpuses of contemporary English (such as the British National Corpus) but also by the amazing resource of the Internet itself, which enables us to search in it for a word and find superb examples of that word in context . . . But the items that occur to us or that we find are passed through the filter of standard lists of complementation patterns in such works as *A Comprehensive Grammar of the English Language* by Randolph Quirk et al. and of collocations (as in the lists of Lexical Functions prepared by Igor Mel'čuk et al. . . .). So what users of BBI get is the product of native-speaker intuition *expanded* by our exposure to authentic English and then *refined* through the standard grids of phraseology and valency developed by outstanding scholars" (p. viii-ix).

The FDCAE describes its compilation thus:

> To find the collocates for a given word, a computer program searched the entire 385-million-word corpus [i.e. the COCA] and looked at each context in which that word occurred. In all cases, the context (or "span") of words was four words to the left and four words to the right of the "node word". The overall frequency of the collocates in each of

those contexts was then calculated, and the collocates were examined and rated by at least four native speakers. Obviously, common words such as *the*, *of*, *to*, etc. were usually the most frequent collocates. To filter out these words, we set a Mutual Information (MI) threshold of about 2.5" (p. 6).

The OCD2 informs us that in choosing which collocations to include:

> The approach taken was pragmatic, rather than theoretical. The questions asked were: is this a typical use of the language? Might a student of English want to express this idea? . . . The aim was to give the full range of collocation – from the fairly weak (*see a movie*, *an enjoyable experience*, *extremely complicated*), through the medium-strength (*see a doctor*, *direct equivalent*, *highly intelligent*) to the strongest and most restricted (*see reason*, *burning ambition*, *blindingly obvious*) . . . Totally free combinations are excluded and so, for the most part, are idioms . . . The first question (Is this a typical use of language?) required that all the collocations be drawn from reliable data. The main source used was the Oxford English Corpus . . . a database of almost two billion words of text in English taken from up-to-date sources from around the world . . . Compilers of the dictionary were able to check how frequently any given combination occurred, in how many (and what kind of) sources, and in what particular contexts . . . The second question asked (Might a student of English want to express this idea?) led to a focus on current English: language that students not only need to understand but can be expected to reproduce. Consideration was given to the kinds of texts that students might wish to write. Primary attention was given to what might be called 'moderately formal language' (p. v-vi).

Judged by these descriptions, the BBI comes very much from the phraseological approach to collocation, in which collocations are viewed as combinations of words with a degree of opacity and fixedness, and intuition seems to have played a considerable part in its compilation. The FDCAE is very much from the statistical tradition, in which the emphasis is on frequency and numerical findings in corpora. The OCD2 lies somewhere between these two, with its use of a corpus and corpus

tools seemingly tempered by a strong role for the compilers.

7.2.2 Corpus-based lists of collocates

The second of the four approaches to listing canonical collocates was corpus-based. This approach was explored since the majority of recent studies on collocation and L2 learners/users have largely been frequency-based (see Section 2.2.2), and because Durrant's definition of collocation, which guides this thesis, sees frequency as one basis for judging a word combination a collocation. The three dictionaries used in compiling the dictionary-based lists are each, to a greater or lesser extent, corpus-based. However, as the quotes from the dictionaries in Section 7.2.1 make clear, and as further clarified in Section 7.6 below, none of the dictionaries is purely corpus-based and the hand of the compilers can be seen in each. A corpus-based approach was therefore expected to produce somewhat different lists of collocates from the dictionary-based approach, and, as Section 7.3 below shows amply, this was indeed the case.

Corpus-based work is often viewed as being more objective. However, in order to produce lists of collocations using a corpus, a great number of decisions must be made, each of which may affect the results (as noted in Section 2.2.2). As McEnery and Hardie (2012) put it: "calculating collocation via statistical testing, uncontroversial in theory, becomes problematic in practice . . . there is in effect an inherent subjectivity in the determination of what is, and what is not, a collocate" (p. 127). Making these decisions is not easy since there is no means of independently validating the results: "All we can do is look at the result and decide whether it makes sense or not" (Barnbrook, et al., 2013, p. 89).

The first decision concerns the choice of corpus or corpora. As noted in Section 7.1, one concern in the current study was the question of selecting appropriate corpora for the learners concerned. It was thus explored whether an ELF corpus could be consulted. Unfortunately, probably the best such corpus available, the Vienna-Oxford International Corpus of English (VOICE, 2013), currently comprises approximately one million words, making it too small for detailed investigation of collocation.

Instead, two larger L1-user corpora were used. First, the COCA (Davies, 2008-) was selected, since American English is very much the focal point of English

education in Japan. The COCA is a very large corpus, but its spoken component consists entirely of broadcast material from television and radio programmes. In order to obtain better coverage of informal, everyday spoken language, the slightly older (early 1990s) BNC was also used. It may be noted that the COCA was also used by one of the three dictionaries (the FDCAE) which were employed in compiling the dictionary-based lists. However, the FDCAE was just one contributor to those lists, while the COCA was one of two corpora used in compiling the corpus-based lists. The expectation therefore was that the lists of collocates produced by the dictionary-based and corpus-based approaches would differ somewhat, and this was indeed the case (see Section 7.3).

In both the COCA and the BNC, separate searches were conducted of the spoken and written sections of the corpus. This was because in both corpora the written component is far larger than the spoken and so each is biased as a whole towards written language. In addition, Nation (2004, 2006b) has suggested that spoken corpora are more suitable models for language learning, Simpson-Vlach and Ellis's (2010) work on an academic formulae list found considerable differences between spoken formulae and written formulae with relatively little overlap between the two, and Gablasova, Brezina and McEnery (2017) have demonstrated differences in the strength of particular collocations in spoken and written corpora. By conducting separate searches of the spoken and written components of the corpora, it was considered that broader collocations lists could be obtained.

Secondly, decisions must be made about the analysis of the corpora. These include determining the span to be used (i.e. the number of words either side of the search term), choosing a measure of collocation, deciding how to enter the search term, and establishing how to treat the output that results from the previous choices. Decisions on these matters were reached after surveying a number of recent studies on collocation, reviewing corpus linguistics literature, and trialling different criteria.

- The span: Following Sinclair, Jones and Daley's (1970) pioneering work, a span of four words either side of the node became conventional, and many recent studies continue to follow this practice (e.g. Durrant, 2009; Groom, 2009; Walker, 2011; Wolter & Gyllstad, 2011). Mason (1997), however, showed that there is some variability in the influence words have on their lexical environments, and McEnery and Hardie (2012) demonstrate that

different spans can have dramatic effects on the output of an analysis, resulting not only in more or fewer collocates being identified, but different collocates being identified. Another factor in determining the appropriate span for my purposes is the nature of *LexCombi* in its adapted form, which might be thought to encourage responses that typically appear closer to the cue word. Thus, a trial was conducted using 10 of the cues and performing the same searches with spans of +/-2, +/-3 and +/-4. In contrast with McEnery and Hardie's findings, the choice of span made relatively little difference to the results. For example, a span of +/-4 as opposed to +/-3 produced a slightly longer list of collocates, but, apart from the few additional items on the +/-4 lists, almost all the items on those lists were the same. Accordingly, bearing in mind the desire for methods which take a broad view of collocation, it was decided to follow convention and use a +/-4 span in all cases.

- Measures of collocation: An array of measures of collocation have been proposed (see Evert (2008) and Manning and Schütze (1999) for reviews); however, in recent studies the most commonly used measures are:
    (1) frequency of co-occurrence (e.g. Durrant & Schmitt, 2009; González Fernández & Schmitt, 2015; McGee, 2009; Shin & Nation, 2008; Siyanova-Chanturia & Spina, 2015; Siyanova & Schmitt, 2008; Sonbul, 2015; Walker, 2011);
    (2) *t*-scores (e.g. Durrant & Schmitt, 2009; González Fernández & Schmitt, 2015; Groom, 2009; Walker, 2011; Wolter & Gyllstad, 2011);
    (3) MI (mutual information) scores (e.g. Durrant, 2009; Durrant & Schmitt, 2009; Ellis, et al., 2008; González Fernández & Schmitt, 2015; Groom, 2009; Siyanova & Schmitt, 2008; Sonbul, 2015; Wolter & Yamashita, 2015).

Measures of collocation can differ in terms of: (1) the statistical conceptualisations that lie behind them; and (2) the results they produce. Regarding the former, most collocation measures take account of both the frequency of the combination as found in the corpus (observed frequency) and the frequency we might expect to find given the frequencies of the

individual component words (expected frequency), but there are two basic types. Some, such as *t*-scores, are measures of certainty of collocation: that is, they ask "How unlikely is the null hypothesis that the words are independent?" (Evert, 2008, p. 1,228). The calculation for *t*-scores is observed frequency minus expected frequency, divided by the square root of observed frequency (see Barnbrook (1996) for details on this calculation). Others, such as MI scores, measure the strength of association, asking "How much does observed co-occurrence frequency exceed expected frequency?" (Evert, 2008, p. 1,228). Thus, MI scores are the ratio between observed frequency and expected frequency, with the results then log-transformed to produce more manageable numbers (see also Section 2.2.2).

Both types of measures have been criticised as being misconceived since the statistics are based on an assumption that words are combined at random when clearly they are not. This means that there can be no statistically meaningful cut-off point established for these tests (e.g. the usual practice of describing *t*-scores greater than 1.96 as significant is inappropriate; see Section 4.5.1). Nonetheless, these tests can still be considered useful in that they provide scores that allow collocations to be ranked (Evert, 2008; Kilgarriff, 2005; Manning & Schütze, 1999; Stubbs, 1995). If a cut-off point is then adopted, it must be recognised as a choice made by the researcher.

In terms of results (i.e. the collocations identified), of the most widely used measures, co-occurrence frequency and *t*-scores are said to produce somewhat similar results, while MI scores give a rather different picture (Barnbrook, 1996; Durrant & Schmitt, 2009). Research has also revealed differences between L1 users and L2 learners/users in their processing and use of collocations as defined by these measures (see Section 3.2.3). Collocations with high MI scores seem to be favoured by L1 users, while collocations with high frequency or high *t*-scores seem to be favoured by L2 learners/users (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis & Simpson-Vlach, 2009; Ellis, et al., 2008).

Given the capacity for different measures to pick up different aspects of collocation, two measures were used in the present study. First, frequency of

co-occurrence was selected as one measure since in being introduced to *LexCombi* learners are told that collocations are "pairs of words that are often used together". Second, to obtain an alternative view, and also because it has been suggested as capturing something important about our intuitions of collocation, MI was used.

These two measures have both faced criticism. Frequency of co-occurrence is criticised in that two highly frequent words will tend to co-occur simply because each is frequent, not because they necessarily have any association with each other (Barnbrook, 1996; Henriksen, 2013). MI is criticised for producing highly inflated scores for words that occur with low frequency in the corpus (Manning & Schütze, 1999). To deal with these problems, a threshold level for each measure was set using the other measure. That is, when searching for the most frequent collocates, only those that reached a minimum MI threshold were included, and when searching for the highest MI collocates, only those reaching a minimum frequency threshold were included. This was intended to ensure that the results of the frequency measure were not dominated by words of high frequency alone and that the results of the MI measure were not dominated by obscure words.

- Entering the search term: The next issue considered was what exactly we are searching for collocates of: that is, how the search term should be entered. Corpus interfaces allow items to be searched for in a variety of ways: (1) As an individual word form, in corpus terms a string of letters (e.g. collocates of the four-letter string *name* bounded on either side by spaces or a punctuation mark); (2) Where the corpus has been grammatically tagged, as an individual word form occurring as a particular part of speech (e.g. collocates of the string *name* when it appears as a noun); (3) As a lemma (e.g. collocates of the strings *name*, *names*, *named* or *naming* when each appears as a noun).

  As Section 2.3.1 reported, there is no consensus on which option is best. In one recent corpus-based study of collocations, lemmas were employed because "preliminary experiments indicated that syntagmatic associations often hold between lemmas rather than particular word forms"

(Michelbacher, et al., 2011, p. 252). Another study suggests that the choice of word forms or lemmas makes little difference (Durrant, 2014). However, a long line of corpus linguists argue in favour of word forms. The OSTI report (Sinclair, et al., 1970) showed that combining findings for the various forms of nouns or for the various forms of verbs provided no additional information on collocations (i.e. lemmatisation led to the identification of fewer collocations than looking at forms individually). Kjellmer (1994) did not carry out lemmatization for his dictionary of collocations since "our working hypothesis must be that it is the individual form of a word rather than the lemma to which it belongs that tends to co-occur with certain other forms" (p. xix). Taylor (2012) gives several examples of how inflections of verbs can each have their own sets of collocates. Consequently, the third option above, searching for lemmas, was ruled out, and indeed few recent studies focus on lemmas. The choice between options one and two was based on the fact that the *LexCombi* instructions inform participants that the cues are nouns. Thus, the second option above was chosen: that is, the cues were searched for as nouns, only in the base form.

- Treating the output: The final decision to be made was how to treat the results of the corpus searches. There are two options: a threshold score for canonicity can be established, or, from a ranked list of collocates, the top *n* results can be accepted as canonical (Evert, 2008). Despite it not being possible to establish thresholds for "significant" collocates, as discussed above, the decision was made to employ thresholds so that the collocates of each cue would have the same minimum values. For the MI-based searches, collocates were defined as combinations with MI scores of at least 3 (i.e. the collocation occurred 8 times more frequently than would be expected by chance, since $\log_2(8) = 3$) and the minimum frequency threshold was set at one occurrence every two million words (meaning that for the smallest corpus-component consulted, the BNC Spoken, at least five occurrences were required). For the frequency-based searches, collocates were defined as combinations with a frequency of at least one occurrence every million words and the minimum MI threshold was set at 1.58 (i.e. the collocation occurred 3 times more frequently than would be expected by chance, since

$\log_2(3) = 1.58$).

Accordingly, eight separate corpus searches were conducted for each cue. That is, the spoken and written components of both the COCA and the BNC were separately searched for MI-based and frequency-based collocations. The span in each case was +/-4; the search term was the base form of each word specified as a noun; and MI-based collocates had to have an MI score of > 3 with a frequency of > 1 occurrence per two million words, while frequency-based collocates had to have a frequency of > 1 occurrence per million words with an MI score of > 1.58.

7.2.3 L1-user norms

Norms-based scoring of *LexCombi* (i.e. using lists based on the responses of competent users of English performing the task) was pursued for several reasons. First, *LexCombi* is a descendant of a word association instrument, *Lex30* (Meara & Fitzpatrick, 2000), and in word association research the use of norms lists is one common approach (Fitzpatrick, Playfoot, Wray, & Wright, 2015). Second, it has been suggested that elicitation tasks are somewhat different from natural language production (Mollin, 2009), and therefore scoring based on others completing exactly the same task may provide a fairer reflection of possible achievement on the task. Thirdly, norms lists can be seen as turning things around such that rather than defining "collocation" and then extracting lists of collocates, with the norms lists, a definition emerges from the informants' responses. Finally, looking at norms for *LexCombi* introduces a different way of viewing collocation. The majority of studies on collocation treat it as a property of language; a phenomenon that is encountered when language is examined as if it were an artefact. This view is common to both the statistical approach to collocation and the phraseological approach. However, norms data provides an alternative way of viewing collocation, as a connection between items in the mental lexicon; that is, a psychological phenomenon.

There being no existing collocations norms data of the type described above, it was necessary to compile a data set. Norms data was collected from a total of 75 respondents. All respondents were British, reported English to be their first language and were either undergraduate students at a university in the UK or educated to at least Bachelor's degree level. Forty-two respondents were under 20 years of age, seven between 21 and 30, 18 between 31 and 40, one between 41 and 50, and two

between 51 and 60; five did not report their age. These respondents completed *LexCombi*, receiving the same instructions and examples (Appendix D) as learners (though in English rather than Japanese), the only difference being that they were asked to respond to 70 cues rather than only 30 since, as mentioned previously, there were 70 cues under trial.

Any response given by two or more respondents was accepted as a collocate. Lemmatisation of the responses was not employed, for two reasons. First, as discussed earlier, it has been found that different forms of a word have their own sets of collocates. Second, lemmatisation involves altering the responses of informants. Instead, the assumption was that there was likely a reason for the form in which a response was given by an informant, so altering that form would be inappropriate.

7.2.4 L2-user norms

In accordance with the desire for scoring approaches that are appropriate for the learners in question, it was also decided to trial L2-user norms-based scoring. This was motivated by criticisms of the use of L1-user norms (see Section 2.3.6), and by the idea that the reference point against which learners are judged should be as close as possible to the language they have been exposed to (see Section 2.3.7).

Data was collected in the same way as for the L1-user norms, again with 75 respondents. All respondents had Japanese as their first language, were proficient in English and were regular users of English. Informants were recruited in snowball fashion, with the majority known to me personally, or being one step removed. English proficiency was not confirmed formally, but attested to informally by the network of respondents. The respondents reported their use of English as follows: 50 reported speaking, reading and writing in English on at least a weekly basis, and 21 more reported doing one of these at least weekly (four respondents did not respond to this query). The respondents had spent an average of almost four years overseas, though there was considerable variability, with some having spent limited time abroad and some quite considerable periods of time. Time spent abroad was almost exclusively in English-speaking countries, though seven informants had spent time in non-English-speaking countries. The respondents reported a range of occupations: there were 22 academics, 18 English teachers, eight undergraduate or graduate students, seven housewives, seven office administrators, four teachers of subjects

other than English, three translators, one airport employee and one research assistant (four did not report their occupation). Eight participants were aged 21-30, 27 were 31-40, 16 were 41-50, 15 were 51-60, and five were over 60 (four did not give their age). As with the L1-user norms, the responses were not lemmatised and any response given by two or more of the informants was accepted as a collocate.

7.2.5 Multiply listed

The four primary approaches, dictionary-based lists, corpus-based lists, L1-user norms and L2-user norms, each resulted in lists with some idiosyncratic features. In order to eliminate these features and concentrate on the commonalities among the lists, a further set of lists was created, derived from the four primary approaches. A list for each cue was made by accepting words appearing in any two or more of the above four lists for that cue: that is, any words appearing in more than one of the lists.

7.2.6 All lists combined

In order to take the most wide-ranging and inclusive approach to collocation, one more set of lists was made, again derived from the four primary approaches. These lists were created by simply merging the four primary lists for each cue into a single list. That is, a collocate appearing in any one of the four lists was accepted as a collocate for the cue. These lists are referred to hereafter as ALC.

7.3 Comparing the content of the lists

The six sets of lists were first compared in terms of their content. Table 7.1 shows the descriptive statistics for the number of collocates per cue in each set of lists.

Table 7.1: Number of collocates in each set of lists for the 70 cues.

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Mean number of collocates	151.31	102.67	29.40	30.96	56.34	233.77
SD	55.094	53.037	5.648	6.028	17.905	72.922
Minimum	47	23	16	17	20	103
Maximum	374	286	41	43	112	421

Clearly, there is considerable variability in the number of collocates, both across the sets of lists and within some of the sets. The two norms-based sets of lists are, however, very similar, reflecting the similarities in their compilation. One other noticeable feature is that the ALC lists are considerably larger than even the largest of the four primary lists. This shows that the overlap between the four lists is relatively small (see Tables 7.3 and 7.4 below).

Table 7.2 shows the correlations between the number of collocates in each list for the cues. Some of the list sets are not normally distributed, so while Pearson's correlations are given, significance levels cannot be calculated. These figures show that a cue which has more collocations under one approach has some tendency to have more collocations under other approaches. The figures also demonstrate, however, that the lists are somewhat different from each other. Among the four primary approaches, the highest correlation is between the two norms-based approaches, but even this correlation is moderate. The number of collocations in the dictionary-based and corpus-based lists exhibit a correlation of .29, despite two of the three dictionaries being largely corpus-based.

Table 7.2: Correlations between the number of collocates in each list for the 70 cues.

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Dictionary-based		.29	.35	.44	.59	.85
Corpus-based			.33	.46	.83	.73
L1 norms				.61	.53	.44
L2 norms					.55	.59
Multiply listed						.82
ALC						

Next, Table 7.3 shows the mean number of collocates common to the lists per cue and the number of collocates unique to each list. Since, however, the mean length of the lists varies greatly (as shown in Table 7.1), Table 7.4 provides the mean percentage of collocates on each list which appear on the other lists from the perspective of each set of lists. Thus, reading across Table 7.4, the first row shows that on average the collocates overlapping between the dictionary-based and corpus-

Table 7.3: Mean number of collocates unique to each list and overlapping for the 70 cues.

A	Only in A	In both A and B	Only in B	B
Dictionary-based	113	38	64	Corpus-based
Dictionary-based	135	16	14	L1 norms
Dictionary-based	132	19	12	L2 norms
Corpus-based	64	38	113	Dictionary-based
Corpus-based	91	12	18	L1 norms
Corpus-based	90	13	18	L2 norms
L1 norms	14	16	135	Dictionary-based
L1 norms	18	12	91	Corpus-based
L1 norms	16	13	18	L2 norms
L2 norms	12	19	132	Dictionary-based
L2 norms	18	13	90	Corpus-based
L2 norms	18	13	16	L1 norms

Note. The table contains some repetition of data for ease of reference, e.g. the first and fourth rows present the same figures albeit in reverse order.

Table 7.4: Mean percentage of collocates on each list appearing in another list from the perspective of each list (see text for details).

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed
Dictionary-based		39.45 26.24	54.00 11.42	62.50 13.78	88.01 34.65
Corpus-based			39.48 12.87	41.66 14.33	71.54 43.35
L1 norms				42.89 45.15	39.59 72.11
L2 norms					45.10 78.05
Multiply listed					

based lists account for a mean of 26.24% of the dictionary-based lists and 39.45% of the corpus-based lists (the corpus-based figure being higher since the corpus-based lists contain fewer items).

The figures in Tables 7.3 and 7.4 confirm that there are considerable differences between the approaches. This is not just a matter of the sets of lists being of very different lengths. The dictionary-based lists are on average substantially longer than the other lists and yet there are still quite a number of collocates in the other three sets of lists that do not appear in the dictionary-based lists. The two norms-based list sets, meanwhile, are similar in length, but have a limited degree of overlap.

Finally, to give a more concrete impression of the overlaps and contrasts between the lists, Table 7.5 presents the canonical collocates for one cue, NAME. It can be seen that just six collocates appear in all four of the lists, and a further 15 occur in various combinations of three out of the four lists. Also striking is the number of items which appear in only one of the four lists. Several other features in Table 7.5 illustrate characteristics of the four primary lists:

- The six collocates which appear in the other three lists but not in the corpus-based list (*bad*, *family*, *first*, *good*, *of*, *second*) are all high-frequency words. This reflects the criteria used in compiling the corpus-based lists, specifically the minimum MI threshold used when carrying out frequency-based searches. These words do appear frequently with NAME in the corpora, but do not reach the MI threshold.
- The four collocates appearing in the other three lists but not in the dictionary-based list (*calling*, *my*, *nick*, *your*) are also instructive. *Calling* is not included in the dictionaries most likely because all three, for the most part, list only the base form of each lemma (see Section 7.6.1 below); thus *call* does appear in the dictionary-based list. Likewise, *my* and *your* are most probably absent from the dictionary-based list due to an editorial decision by the dictionary makers to omit personal pronouns. Finally, *nick* is presumably not in the dictionaries since it is considered to be part of a compound. Its presence in the other three lists is explained by two factors. It is in the corpus-based list largely due its use as a personal name, though there are some instances in the corpora of *nick name* as two orthographically separate items. The occurrence of *nick* in both norms lists meanwhile is part

of a general tendency for informants to provide parts of words alongside the cue. Another example of this can be seen further down the table: the collocate *sur*, which occurs in both norms lists, but in neither the dictionary-based nor corpus-based lists (see also Section 7.6.3).

- One collocate occurs in the other three lists but not in the L2 norms list: *christian*. This would appear to reflect the influence of cultural practices on language, and therefore also on the lists: *first name* is the preferred term in Japan and tends to be included in school textbooks (see also Section 7.6.6).
- Of the 69 collocates which appear exclusively in the corpus-based lists, a good number seem quite idiosyncratic. This is typical of the type of "noise" that corpus-based searches often produce. For many purposes, it would seem appropriate to remove some of these items. This was not done, however, since it was considered that these collocates are unlikely to affect the actual scoring of responses. For example, personal names occasionally appear in the corpus-based lists, but it was thought unlikely that a participant would give a personal name as a response to *LexCombi*.

To sum up, this section has compared the content of the scoring lists and has found that they differ in a number of ways. First, the mean size of the lists is quite different. Among the four primary sets of lists, the largest, dictionary-based, is

Table 7.5: Collocates appearing in the four primary lists for the cue NAME.

DB	CB	L1	L2	full last	maiden	middle	real	tag
DB	CB	L1		christian				
DB	CB		L2	famous	given	list	user	
DB		L1	L2	bad family	first	good	of	second
	CB	L1	L2	calling	my	nick	your	
DB	CB			address brand change character clear	code common familiar father file	forget give household mention name	pronounce proper recognition remember spell	under use write

DB		L1		badge				
DB			L2	big	by	call	in	plate
	CB	L1		game	her	his		
	CB		L2					
		L1	L2	card	foreign	sur		
DB				abandon acquire adopt and appear ask associated assume assumed attached band bear become bell besmirch birth blacken carry catch choose come	company damage decide dirty domain double-barrelled down drop enter faces fake false fancy for forward from generic geographic geographical get go	God's great have hear holy hyphenated immortalize imply invoke joint keep know known law legal lend make man married mean mud	obtain official on original out pass pen pen-name personal pet pick place print professional proprietary protect put recognizable register reveal ring	scientific share sign smear something sound stage street suggest sully synonymous take to trade true vain vernacular well-known with withhold
	CB			account agent ballot bearing bears billy bore called cell changed crow date david desmond	directory dolly dot field forgotten funny gave god heard hence hi implies include is	its jesus jonathan joyce kingdom letters lifespan linda logical lord mentioned module package person	please posh range rank registered russell sang sarah send singer sp special specified spelt	standing suggests thy town trusted using valid what whatever whose wife withheld workstation
		L1		baby	long	own	sake	their
			L2	after	beautiful	it	unique	value

Note. DB = dictionary-based; CB = corpus-based; L1 = L1 norms; L2 = L2 norms.

approximately five times greater than the smallest, L1 norms. Second, there is relatively little consistency in the size of the lists: there is in general only a moderate tendency for a cue which has a relatively large number of collocates under one list to also have a large number of collocates under a different list. Third, the number of overlapping collocates between the sets of lists is relatively small. Even the larger lists do not come close to containing all of the collocates in the smaller lists.

It seems therefore that the desire of this study to explore rather different approaches to defining collocational canonicity has been achieved. Nevertheless, the extent of the differences between the sets of lists is notable, and highlights how different varying conceptions of collocation are. This may mean that no single source can provide a definitive list of collocates for a given word and the use of multiple sources is therefore advisable.

7.4 Scoring responses with the lists

This section considers how each set of lists performs with respect to scoring responses from learners. To enable this, a new set of data using the adapted *LexCombi* was gathered from learners. In addition, the learners were given a Yes/No vocabulary test which served as a proxy measure of proficiency.

The Yes/No vocabulary test combined items from the first *X_Lex* test provided in Milton (2009) and items from the computer-based *Y_Lex* test (Meara & Miralpeix, 2006). The *X_Lex* test contains 20 items each from the 1K to 5K frequency levels along with 20 pseudowords. The *Y_Lex* program contains banks of items at the 6K to 10K frequency levels and a bank of pseudowords, from which 20 items were randomly chosen at each level along with 20 pseudowords. The 200 items and 40 pseudowords were then randomly sorted to create the Yes/No test, which was presented to participants in paper-and-pencil format. This test can be found in Appendix E. Following Milton (2009) and matching the approach taken with the *X_Lex* test in Chapters 5 and 6, the estimate of the participants' vocabulary knowledge was made by multiplying the number of real words checked by 50, since each real word on the test represents 50 words from its frequency level, and then subtracting 250 for each pseudoword checked. This combined *X_Lex*/*Y_Lex* was used here since there were indications of ceiling effects when the *X_Lex* test was

used in Chapter 6.

The adapted format of *LexCombi* was used and three versions were created featuring different sets of cues. Ten cues used in previous versions of *LexCombi* were given to all participants, along with three different sets of 20 further cues. In this way, each participant responded to 30 cues, while a total of 70 cues were trialled (This is why, as described earlier, norms data was collected for 70 cue words). Two forms of each version of *LexCombi* were used with the cues in different orders in each. Further details on the cues and their trialling are given in Chapter 8.

Data was collected from 273 Japanese university students at two universities. Fifty participants were excluded: one did not fully complete the two instruments; eight had non-Japanese L1s; and 41 showed high levels of guessing on the Yes/No test. Thus the final data set contained data from 223 participants. The Yes/No test showed these participants to have an estimated mean vocabulary size of 5,438 words (*SD* = 847), with scores ranging from 2,950 to 7,850.

The responses to *LexCombi* were typed up, with obvious misspellings corrected and multi-word responses reduced to a single word, as explained in Table 4.1. The responses were then scored automatically against each set of lists in turn using CollCheck (Imao & Brown, 2014).

Two aspects of how the lists perform were considered. First, recurrent responses (i.e. responses to a given cue that were produced by at least 10% of the participants) were identified and the number of recurrent responses which were canonical under each scoring approach was examined. This analysis was undertaken since, as mentioned previously, one motivation for examining *LexCombi*'s scoring was the fact that many of the recurrent responses in the Chapter 4 data seemed reasonable yet were not considered canonical using the OCD2 as the criteria for canonicity. Second, the *LexCombi* scores of participants were calculated under each scoring approach and an analysis undertaken to ascertain how the scoring approaches impact on learners' scores.

7.4.1 Recurrent responses

Since the participants did not respond to all 70 cues, the threshold for recurrence varied between cues. For the 10 cues which all 223 participants responded to, recurrent responses were those given by at least 23 participants. The other 60 cues

each had between 73 and 76 respondents, so recurrent responses were those given by at least eight participants.

Using these criteria, a total of 318 recurrent responses were found, an average of 4.5 per cue. The cue BALL had the most recurrent responses, with 11, while DEVELOPMENT had the least, just one. Table 7.6 shows the extent to which the recurrent responses are included in each set of lists. There is a good deal of variation among the four primary scoring approaches, the corpus-based lists including just 56% of the recurrent responses, while the L2-user norms lists include 91%. The ALC approach naturally has the highest coverage of the recurrent responses, there being just 13 (4%) not included in these lists.

As reviewed in the introduction to this chapter, Section 4.5.1 identified four types of recurrent responses which were not deemed canonical by the scoring criteria then being applied (the OCD2 lists): some were deemed non-canonical due to the limitations of the OCD2, some reflected the influence of the particular linguistic environment of the learners, some were word parts forming complex words with the cue, and some were simply non-standard from the perspective of English usage. Of the 13 recurrent responses in the current data set which are not canonical under ALC scoring, one reflects the particular linguistic environment of the learners (the response *Osaka* to the cue STATION, the name of a major station). The majority, however, are of the fourth type, non-standard usage, such as the response *six* for the cue SENSE and the response *make* for RELATIONSHIP.

Table 7.6: Canonical recurrent responses under each scoring approach.

	Canonical recurrent responses (as percentage of total recurrent responses)
Dictionary-based	224 (70)
Corpus-based	177 (56)
L1 norms	218 (69)
L2 norms	288 (91)
Multiply listed	269 (85)
ALC	305 (96)

Note. Maximum = 318.

7.4.2 *LexCombi* scores under each approach

In order to ascertain how the scoring approaches actually impact on scores, each participant's *LexCombi* score was calculated under each of the scoring approaches. Table 7.7 presents descriptive statistics for the scores. Comparing the mean *LexCombi* scores under each approach with the mean number of items in each set of lists (given in Table 7.1 above), it is apparent that there is no clear relationship between the two. Naturally, ALC results in the highest scores, but of the four primary scoring approaches, the highest scores are given by L2 norms scoring, despite these lists being on average one third of the length of the corpus-based lists, which awarded the lowest scores. It can also be seen that all six of the scoring approaches produce data with a reasonable range and distribution of scores.

Table 7.7: Descriptive statistics for *LexCombi* scores under six scoring approaches ($N = 223$).

	Mean	*SD*	Minimum	Maximum
Dictionary-based	37.57	9.207	16	61
Corpus-based	27.80	6.886	13	51
L1 norms	31.97	7.381	15	50
L2 norms	42.17	9.050	18	66
Multiply listed	40.88	8.712	19	64
ALC	52.93	10.657	25	82

In order to cast further light on the scores produced by the approaches, Table 7.8 provides the correlations between the participants' scores under each approach. In some cases the scores were not normally distributed, so though Pearson's correlations are shown, the significance of the correlations could not be calculated. These correlations are all in the moderate to strong range and suggest that the considerable differences between the lists shown in Tables 7.3-7.4 translate into less substantial differences when it comes to scoring.

This section has looked at how well the recurrent responses among a group of learners were covered by the scoring approaches and at the *LexCombi* scores achieved by those learners under the six approaches. While coverage of the recurrent

Table 7.8: Correlations between scores under different scoring approaches.

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Dictionary-based		.79	.64	.80	.88	.86
Corpus-based			.69	.75	.84	.79
L1 norms				.84	.85	.81
L2 norms					.93	.91
Multiply listed						.92
ALC						

responses was seen to vary a great deal among the approaches, under ALC scoring almost all the recurrent responses produced by the participants were judged canonical, the exception being those that are non-standard in terms of English usage. ALC scoring therefore seems to be effective in crediting learners appropriately for their responses.

With regard to *LexCombi* scores achieved under the approaches, all six produced scores with a reasonable range and distribution. In addition, the scores under the various approaches correlated with each other quite strongly. This indicates that the choice of scoring approach gives a different impression of the absolute ability of the participants to produce collocations, but makes less difference to how their ability appears relative to each other.

7.5 A Rasch analysis of the scoring approaches

Having seen the impact of each scoring approach on learners' *LexCombi* scores, this section presents a Rasch analysis of the scores in order to explore the measurement properties of the approaches. A Rasch analysis was conducted for two reasons. First, it provides a rich array of insights into the statistical properties of a data set, allowing the quality of the data to be judged in various ways and any problems with the data to be highlighted. In simple terms, Rasch analysis provides more tools with which to investigate a set of test scores than is provided by classical test statistics. Secondly, Rasch analysis is able to deal with data sets that are not complete. As explained previously and elaborated further below, the learners in the current study responded to one of three sets of cues which were under trial, and thus the data sets were

incomplete. Classical test statistics cannot be used with such data sets, but for a Rasch analysis, provided that a certain number of observations are available to allow probabilities to be calculated, gaps in the data do not cause major problems.

Rasch analysis is based on the calculation of probabilities (Bond & Fox, 2007; McNamara, 1996). For any test item, there is a probability of it being answered "correctly", and for any test participant there is a probability of he/she answering an item "correctly". Thus, test items and test participants are viewed as analogous and can be placed on a single measurement scale.

Rasch analysis produces measures in units termed "logits", or log-odds units. Logits are a true interval measure in that a difference of one logit at one point on the scale is identical to a difference of one logit at a different point on the scale. Differences as expressed in logits can also inform us of the probability of a particular score on a certain item. Thus, a participant with the same overall score in logits as the overall measure in logits for a particular item has a 50% chance of success on that item. A participant with an overall score one logit higher has a 73% chance of success on that item, while a participant with an overall score one logit lower has a 27% chance of success on the item. As will be seen below, this allows the identification of scores that were unlikely.

The Rasch analyses were conducted using Winsteps (Linacre, 2015), with a separate analysis of the *LexCombi* scores data carried out for each of the six scoring approaches. The analyses had to take account of the fact that for each *LexCombi* item there are four possible scores (i.e. a participant can give 0, 1, 2 or 3 canonical collocates), and allow for the possibility that the items may not behave identically to each other. This is achieved in Rasch analysis by using the Partial-Credit model. For some of the data sets, some scoring categories were not present in the data. For example, under corpus-based scoring, none of the participants achieved a score of 3 for the item POWER. In these cases a dummy record (i.e. a pretend person with a score of 3 for POWER) was added to the data set to force Winsteps to take account of the category. This dummy record was then excluded from consideration when conducting analyses.

Prior to the main analyses, it was necessary to confirm that participants who received different sets of cues were equivalent. As mentioned previously, each participant responded to 30 out of 70 cues under trial, the data sets having the

structure shown in Figure 7.1. Since all participants responded to 10 *LexCombi* cues in common, the responses to these items were examined to establish that there were no fundamental differences between the groups of participants. This was confirmed by carrying out Differential Item Functioning (DIF) analyses on each data set (i.e. each set of scores produced by the six scoring approaches). The DIF analyses compared the item measures produced by each group for the ten cues. For two of the six data sets, the DIF analysis showed there to be one item out of the 10 (the item REASON) for which there was a significant difference between two of the groups. However, while this difference was statistically significant, it had a very small effect on the Rasch measures and thus it was felt reasonable to proceed with the full analysis of the six data sets.

Figure 7.1: Structure of the data sets.

Cues 1-10 Responses from all 223 participants		
Cues 11-30 Responses from participants 001-74	Cues 31-50 Responses from participants 75-147	Cues 51-70 Responses from participants 148-223

Tables 7.9-7.12 show the results of the Rasch analyses. The first of these gives descriptive statistics for the persons and items. The mean person measure shows the performance of the participants relative to the *LexCombi* items in logits. The Winsteps program by default places the mean item measure at zero, with negative person measures indicating that the items were somewhat difficult for the participants while a positive figure indicates that the items were somewhat easy. It can be seen that it was more difficult to score well on *LexCombi* under corpus-based scoring, and considerably easier when ALC scoring was used.

Next, the standard deviation and range of persons and items are shown. At first glance, the range in the measures under the different scoring approaches appear quite small, from 2.16 to 3.48 for persons and from 2.53 to 3.79 for items. If *LexCombi* is to be used to measure learners with considerable differences in their ability to produce collocations, the range of the items is a crucial matter: a small range being akin to a short ruler, useful only for measuring certain things, whereas a longer

Table 7.9: Descriptive statistics for the persons and items in logits.

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Mean person measure	-0.43	-1.08	-0.81	-0.19	-0.25	0.39
Person *SD*	0.49	0.43	0.42	0.47	0.46	0.56
Range of persons	2.49	2.47	2.16	2.61	2.56	3.48
Mean item measure	0	0	0	0	0	0
Item *SD*	0.64	0.89	0.68	0.56	0.67	0.62
Range of items	2.88	3.79	2.84	2.53	2.85	2.77

ruler can measure a greater variety of things. However, the range figures in Table 7.9 show the range in the overall difficulty of the items (e.g. under dictionary-based scoring, the Rasch measure for the most difficult item was 1.45 and for the least difficult item -1.43, and the difference between those two figures is the range, 2.88). Winsteps, in addition to providing these overall measures of item difficulty, also gives the difficulty of achieving each *LexCombi* score for each item. That is, since there are four possible scores for each *LexCombi* item (0, 1, 2 and 3), the difficulty of scoring 1 as opposed to 0 can be calculated, and likewise the difficulty of scoring 2 rather than 1, and of scoring 3 rather than 2. For each *LexCombi* item, then, there are three difficulty measures (referred to as thresholds). For example, under dictionary-based scoring, these thresholds for the cue COUNTRY were -0.92, 0.15 and 0.95. Thus, this item alone had a range of 1.87 logits (i.e. the difference between the lower threshold of -0.92 and the upper threshold of 0.95). The actual length of the "ruler" under each scoring approach was therefore greater than the figures in Table 7.9 suggest, since the threshold for achieving a score of 1 rather than 0 for the least difficult items was somewhat easier than the overall measure of difficulty for the least difficult item, and the threshold for achieving a score of 3 as opposed to 2 for the most difficult items was rather more difficult than the overall measure of difficulty for the most difficult item. Under each scoring approach then, *LexCombi* is

able to measure a considerable range of abilities.

Table 7.10 shows various aspects of the extent to which the data sets match the Rasch model. The theoretical model on which Rasch analysis is based predicts that a participant is more likely to achieve a good score on an easier item than on a harder item (and that it is more likely that an item will be answered correctly by a participant of higher ability than one of lower ability). However, being based on probability, the model does not expect this relationship to always hold: that is, it anticipates a certain degree of variability in the data. Fit statistics give an indication of whether the variability in the data for individual items and persons is within

Table 7.10: Fit and dimensionality of the data sets (N = 223; K = 70).

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Fit						
Number of underfitting persons (Infit mean squares > 1.3)	27	28	22	25	25	24
Number of overfitting persons (Infit mean squares < 0.75)	31	34	35	30	36	27
Number of underfitting items (Infit mean squares > 1.3)	0	0	0	0	1	0
Number of overfitting items (Infit mean squares < 0.75)	0	0	0	0	0	0
Point-measure correlations						
Number of items with a negative point-measure correlation	0	1	2	0	1	0
Number of other items with a low point-measure correlation (< .2)	4	13	15	6	2	1
Dimensionality						
Number of possible additional dimensions	4	3	5	5	5	4

expected boundaries. Underfit refers to item or person data that is more varied than the model predicts; overfit to data less varied than predicted. Underfit is therefore of greater concern.

For example, under dictionary-based scoring, Participant 66 achieved a score of 3 for the item TRADE (with the responses *center*, *world* and *fair*). Participant 66 had an ability of -0.77 (indicating a relatively low ability to produce collocates in the dictionary-based lists as compared with the mean person measure for dictionary-based scoring of -0.43), while the item TRADE had a difficulty of 0.31 (as compared with the mean item measure of 0.00, indicating participants found it relatively difficult to give responses to this item that were in the dictionary-based list). Moreover, the difficulty of achieving a score of 3 on TRADE was 1.28. Considering the gap between Participant 66's ability level and the difficulty of scoring 3 on TRADE, the participant had only a 12% chance of getting a score of 3. To put it another way, Participant 66, based on his/her overall performance, would not have been expected to come up with three canonical (i.e. listed) responses for TRADE, which is a difficult item. The achievement nonetheless of this score is why this participant's data underfits the model.

Recommendations for fit criteria vary among scholars (Bond & Fox, 2007; Wilson, 2005), and the criteria used here (see Table 7.10) are somewhat conservative in that the range for the fit statistics deemed acceptable is smaller than some recommend. Table 7.10 gives the number of underfitting and overfitting persons and the number of underfitting and overfitting items under the criteria shown. There were a considerable number of misfitting persons under each of the scoring approaches, but no misfitting items under five of the six scoring approaches.

McNamara (1996) states that an item may be misfitting because: (1) it is badly constructed; or (2) it measures a different construct or trait to the rest of the items. A person may be misfitting due to: (1) performance factors, such as fatigue, boredom and inattention; (2) the person being somehow different to other persons (i.e. the group of persons is not homogenous in some way); or (3) surprising gaps in that person's knowledge of areas covered by the test.

That there was only one misfitting item under one scoring approach (the item BED under multiply listed scoring, for which some participants surprisingly achieved scores of 0 despite it being overall a relatively easy item) suggests that

there are few problems with the construction of the *LexCombi* items or with the construct being measured. The considerable number of misfitting persons may have resulted from all three of the factors McNamara mentions. The *LexCombi* data was collected from undergraduate participants who volunteered to take part. However, there was no particular incentive, such as course credits, to motivate participants and some may have found their concentration wavering. Nation (2007) has highlighted how the attitude of participants can have a considerable impact on research results. It is also possible that, as per the second factor, some participants differed from the group as a whole. In order to preserve anonymity, the only background information sought from participants was confirmation that their first language was Japanese, so this is difficult to judge. However, it is possible that some participants may, for example, have lived in an English-speaking country for some time, which may have given them superior knowledge of the collocates of some particular cues. Regarding McNamara's third suggestion, surprising gaps in a person's knowledge of areas covered by the test, the issue in fact appears to be surprising degrees of knowledge rather than gaps. Within the results, Winsteps provides lists of the most misfitting item scores, and the majority of these are scores of 3 on items of relatively high difficulty for the particular participant in question. That is, the participant was able to provide three canonical collocates for a cue which the participant group as a whole found difficult, as in the example involving Participant 66 and the cue TRADE above. Examples such as this may be evidence of idiosyncratic pathways towards the acquisition of collocations among learners.

Next, Table 7.10 shows the number of items with negative or low point-measure correlations (the correlation between scores on an individual item and on *LexCombi* as a whole; equivalent to item-total correlations in traditional test statistics). This again points to items that may involve a different construct from the items as a whole. Here, the corpus-based and L1 norms approaches stand out. It was more difficult to score well on *LexCombi* under these two approaches (Table 7.9), and the relatively large number of problematic items may be a consequence of the generally low scores under these approaches.

Finally, Table 7.10 shows details on the dimensionality of the data. The Rasch model assumes unidimensionality: that is, a test measures a single trait or construct. Fit statistics and point-measure correlations give some indication of whether this is

the case, while dimensionality statistics use factor analysis to indicate the possible presence of other traits in the data (Bond & Fox, 2007). All six scoring procedures appear to have had other possible dimensions in play and it is somewhat difficult to know what to make of these numbers. Winsteps shows the particular items that are involved in other possible dimensions, and Linacre (2014) advises that when further dimensions are indicated, the focus should be on the substance of the items in question. If nothing in the content of the items can be found, the dimensionality figures may be simply the result of the randomness any real data contains. Consequently, for each additional dimension that was found, the particular items involved were examined to see if any characteristics could be pinpointed that might distinguish them from others. This involved, for example, checking whether the items were loan words in Japanese or whether they were concrete nouns or abstract nouns. There were, however, no patterns that could be recognised among the sets of items identified by Winsteps.

Moving on to other aspects of the Rasch analysis, Table 7.11 shows the number of items with disordered categories, as determined by two methods. For each *LexCombi* item, there are four possible scores (described in Rasch analysis as categories of responses) that can be achieved (i.e. 0, 1, 2 and 3), and we would expect that they are progressively more difficult to achieve. That is, there is an expected order for these categories. Items with disordered categories are those which do not exhibit this expected order.

Table 7.11: Number of items with disordered categories ($K = 70$).

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Number of items with disordered Rasch-Andrich thresholds	6	4	6	3	3	2
Number of items for which the ability levels of persons in each category was disordered	20	21	23	24	25	22

The first method of identifying disordered categories is examining the Rasch-Andrich thresholds for each item. In simple terms, these thresholds show the difficulty of achieving each score: that is, the difficulty of achieving a score of 1 rather than 0, of achieving 2 rather than 1 and of achieving 3 rather than 2. We would expect these thresholds to be ordered such that along the scale of difficulty the threshold between scores of 0 and 1 comes first, then the threshold between scores of 1 and 2, and then that between scores of 2 and 3. The thresholds can, however, be disordered when one of the categories (scores) is under-represented in the data (as Section 8.4.1 explains, this might come about due to the chaining of responses).

The other indicator of disordered categories is somewhat simpler. This is based on the average ability level (i.e. ability on *LexCombi*) of participants who achieved each score category for an item. We would expect, for example, that participants who achieved a score of 2 on an item would on average have greater ability to produce collocates than participants who achieved a score of 1 on that item. The figure in Table 7.11 is the number of items for which this was not the case.

These two measures of problematic items give somewhat different impressions of the scoring procedures: the former flags dictionary-based and L1 norms scoring as having more problems; under the latter there are only small differences between the six approaches.

Rasch analysis produces two figures for the reliability of a test: person reliability and item reliability. These indicate the reproduceability of the measures, for the persons and items respectively: that is, the likelihood that if the same persons took the same test again, the same level of ability for the persons and the same level of difficulty for the items would be found. Person reliability is therefore equivalent to test reliability as measured by Cronbach's alpha in classical test statistics.

The figures (Table 7.12) show that item reliability was considerably higher than person reliability in all cases: that is, we can be more confident that the item measures would be reproduced than that the person measures would be reproduced. This is partly due to the structure of the data sets: for each person there are 30 observations, since each participant completed 30 *LexCombi* items, while for the items there are at least 73 observations. Nevertheless, the person reliability figures are somewhat low, indicating that there may be some uncertainty as to whether *LexCombi* under these scoring procedures provides reproduceable scores. As the

Table 7.12: Person and item reliability under each scoring procedure.

	Dictionary-based	Corpus-based	L1 norms	L2 norms	Multiply listed	ALC
Person reliability	.75	.63	.64	.74	.72	.81
Item reliability	.95	.96	.95	.94	.95	.95

figures show, the differences between the scoring approaches were relatively small, though ALC scoring does reach the .80 threshold often seen as desirable, while corpus-based and L1 norms scoring seem more problematic.

Reviewing the above, it would seem that the six scoring approaches were not widely differentiated in terms of their measurement properties. None of the approaches produced data that is clearly flawed, and none produced data that is ideal. Ultimately, the Rasch analysis suggests that any of the scoring procedures may provide reasonable data. Nevertheless, looking across the range of analyses, ALC scoring did perform somewhat more strongly than the other approaches. It produced the widest range of person measures, had relatively few misfitting persons and no misfitting items, had the fewest items with low point-measure correlations, had relatively few items with disordered categories and produced the highest person reliability. That being said, given that the six approaches were not widely differentiated, it may be more important to consider what the approaches actually imply about collocation. This is the issue considered in the next section.

7.6 Assumptions about the concept of collocation

The final way in which the scoring approaches were compared was to consider the assumptions that underlie each approach regarding the concept of collocation. As Chapter 2 described, collocation has been widely acknowledged as difficult to define, with no generally accepted definition. Wray (2009) has commented that "it is not a matter of us all agreeing on one definition or the other . . . but there is, I think, value in researchers reflecting on the implications of the definitions they use" (p. 239). Thus, in this section, the implicit definition of collocation inherent in each

scoring approach is examined and the implications of these definitions reflected on. Sections 7.6.1-7.6.7 look in turn at each area of debate in defining collocation reviewed in Sections 2.3.1-2.3.7, with each section beginning with a brief recap of the issue (grey box) before the position of each scoring approach with respect to that issue is considered. On occasion, this necessitates the repetition of details about the approaches given previously in Section 7.2.

7.6.1 Word forms and lemmas

There are different positions on the units that are involved in collocation. Some (e.g. Sinclair, 2008) argue that the collocation itself is a unit and so it is misguided to consider units within it. However, much work on collocation essentially treats collocation as a property of words (i.e. a given word has collocates, rather than being part of a collocation). If collocation is taken as a property of words, one issue is the nature of the co-occurring units. Thus, in considering the scoring lists, we can ask *What type of lexical unit is the entry/node/cue?* and *What type of lexical units are the collocates of this?* In the view of many scholars (e.g. Hoey, 2005; Kjellmer, 1994; Sinclair, 1991; Sinclair, et al., 1970; Taylor, 2012), each individual form of a word may have its own set of collocates, but there are suggestions that collocational relationships are often formed between lemmas (e.g. Michelbacher, et al., 2011), while in pedagogically oriented work, the focus is usually on base words, which seemingly stand for a lemma (e.g. Boers & Lindstromberg, 2009; Lewis, 1993, 1997, 2000; McCarthy & O'Dell, 2005).

The three collocations dictionaries, having a pedagogical focus and following the usual practice of dictionaries, have base forms as their entries. The FDCAE is clear in stating that these forms stand for lemmas. Both the BBI and the OCD2 make reference to "headwords". As for the collocates, the FDCAE is again explicit, explaining that "the most frequent form of a given collocate lemma may be an inflected form, not the head form as listed (e.g. *long* as a collocate of *no* almost always appears as *longer* in the corpus)" (Davies & Gardner, 2010, p. 8). The BBI and OCD2 do not provide this clarity. However, as their entries contain cases of a base form listed as a collocate while the example sentence features an inflected form, it seems that they too may work with a concept of "collocate lemmas".

For the corpus-based lists, as explained in Section 7.2.2, the node was the

specific word form, though limited to its noun uses. The collocates were also treated as individual word forms.

In developing the two norms-based lists, the cues were presented to informants without comment as to whether they were to be treated as individual forms or as representatives of lemmas or some other unit. Informants were told, however, that all of the cues were nouns. Looking at the responses of informants, there is an occasional indication that informants may have treated cues as lemmas (or more loosely as a base form standing for a set of closely related forms). For example, there are some plural determiners (e.g. *few*, *many*) in the norms lists, and in response to the cue GLASS both L1-user and L2-user informants gave the word *wear*, possibly indicating that these informants were thinking of the collocation *wear glasses*. Such responses were occasionally observed in the responses of learners, and so the norms lists have the advantage of allowing for this possibility. As for the collocates, as explained earlier, lemmatisation of the informants' responses was not implemented. There did seem to be a tendency, however, for informants to give base forms (albeit these could be present simple forms of the verbs). The norms-based lists may then lean towards a "collocate lemmas" approach, under the influence of dictionary and educational practice.

7.6.2 Lexical and grammatical words

Collocation, as outlined in Section 2.3.2, may at times be restricted to the co-occurrence of lexical words, or may include grammatical words also. In addition, there is the concept of colligation: the combination of a word with a particular class of items.

None of the scoring approaches take account of colligation, but they have different positions with regard to the inclusion of lexical and grammatical collocations. The dictionary-based lists include both lexical and grammatical collocations, but focus on the former. The BBI seeks to include both lexical and grammatical collocations; the OCD2's noun entries are restricted to seven types of collocation, only two of which involve grammatical words (preposition + noun, noun + preposition); and the FDCAE includes no grammatical words as collocates of its noun entries.

The corpus-based lists, being entirely statistically based, make no distinction between lexical and grammatical collocates and include both.

The two norms-based lists likewise do not distinguish between lexical and grammatical collocates and contain both. This reflects the instructions the informants were given. These instructions referred only to "words", making no mention of grammatical or lexical words, and the examples of collocations provided included a grammatical collocation.

7.6.3 Orthographic words and word parts

As explained in Section 2.3.3, collocation, by convention, is seen as involving orthographic words. However, the concept of "word" can be defined in various ways (Wray, 2015), which in turn affects what is included within the concept of collocation. In particular, the distinction between collocation on the one hand and multi-morphemic words and compounds on the other may be open to question. The issue is, therefore, whether collocation can involve word parts as well as separate orthographic words.

Both the dictionary-based and corpus-based approaches follow convention and focus on orthographically separate words. This is not the case for the norms lists, however. Both norms lists include quite a number of items which, it would seem, form single orthographic words in combination with the cue. Examples include *man* and *ful* for POWER (i.e. *manpower*, *powerful*), *news* for PAPER (i.e. *newspaper*), *less* and *non* for SENSE (i.e. *senseless* and *nonsense*), *sake* and *sur* for NAME (i.e. *namesake* and *surname*) and *able* for REASON (i.e. *reasonable*). That these responses are word parts rather than separate items is especially clear in the cases of *ful* with a single '*l*' for POWER and *sur* for NAME.

The fact that learners also provided such responses was noted in my trial of *LexCombi* (Section 4.5.1) and has been seen in each subsequent set of data collected. The initial assumption was that such responses reflected the learners' unfamiliarity with spelling conventions. This is a difficult area for learners given that compound nouns which seem similar to each other may be conventionally written as two separate words, be hyphenated or occur as a single orthographic word. However, since both sets of informants for the norms lists, the L1 users and L2 users, also provided such responses, the learners' behaviour does not seem to be abnormal in any way. It might thus be thought that this is a result of the elicitation method: that something about the format of *LexCombi* encourages such responses. This may be

part of the explanation. However, it also seems likely that psycholinguistically collocation and morphology are related (see Section 11.5 for further discussion).

7.6.4 Semantic opacity

Section 2.3.4 described how a defining characteristic of collocation in the phraseological tradition is a degree of semantic opacity (Cowie, 1998a; Gyllstad, 2007; Howarth, 1998a; Nesselhauf, 2003, 2005). The phraseological tradition describes a scale of opacity, with free combinations at one extreme, idioms at the other and collocations occupying some middle ground (see also Section 2.2.1). In most frequency-oriented work on collocation, semantic opacity is not considered.

The position of the dictionary-based lists on this issue is somewhat difficult to characterise because of the divergent nature of the three dictionaries. The BBI, rooted in the phraseological tradition, states that "Collocations should be included in dictionaries; free combinations, on the other hand, should generally not be included" (p. XIX), and adds specifically that it "does not include free lexical combinations" (p. XXXI). The OCD2 likewise states explicitly: "Totally free combinations are excluded" (p. vi). The FDCAE in contrast makes no comment on this issue, but, given its frequency basis, might be expected to overlook the distinction.

To confirm the approach of each dictionary, searches were made for examples of free combinations cited in the collocations literature. In each case, the question was whether the verb or preposition component of the free combination appears under the entry for the noun. Table 7.13 shows whether examples given by various scholars are included in the dictionaries; Table 7.14 shows how many out of 27 free combinations compiled by Gyllstad and Wolter (2016) the dictionaries include. It should be noted that lack of inclusion does not necessarily indicate that the dictionary excluded the combination on the grounds that it was a free combination; it may have not met other criteria for inclusion. As can be seen, the BBI and the OCD2, despite stating that free combinations are excluded, do in fact include these examples more often than not. This likely reflects the difficulty of distinguishing between free combinations and restricted collocations, as discussed in Section 2.2.1. It is, then, concluded that the dictionary-based lists do not strictly align with the notion of semantic opacity and include both transparent and opaque collocations.

Table 7.13: Inclusion of examples of free combinations in three collocations dictionaries.

	BBI	FDCAE	OCD2
read a newspaper[1]	not included	included	included
want a car[1]	not included	not included	not included
blow a trumpet[2]	included	no entry for TRUMPET	included
under the table[2]	included	not included	included
kick the ball[3]	included	included	included
throw a disk[4]	not included	not included	not included
pay money[4]	not included	included	included

[1]Cited as an example of a free combination by Nesselhauf (2003).

[2]Cited by Howarth (1998a).

[3]Cited by Gyllstad (2007).

[4]Cited by Laufer and Waldman (2011).

Table 7.14: Number of free combinations from Gyllstad and Wolter (2016) included in three collocations dictionaries.

	BBI	FDCAE	OCD2
Gyllstad and Wolter (2016) free combinations (K = 27)	23	24	22

The corpus-based lists are simpler to characterise: no distinction between free combinations and restricted collocations was made in their construction.

The two norms-based sets of lists likewise do not take account of semantic opacity in any formal way. The instructions for informants defined collocation as pairs of words frequently used together. However, the perception of language users of what is and is not frequent may not be determined by raw frequency alone: semantically opaque collocations may be more salient and thus more likely to be given by informants. Nevertheless, while restricted collocations do occur in the lists, free combinations do also. For instance, of the nine examples of free combinations from Table 7.13 and from Gyllstad and Wolter (2016) which feature *LexCombi* cues, four appear in the L1 norms lists, and six occur in the L2 norms lists.

7.6.5 Structural integrity

As seen in Section 2.3.5, structural integrity is the notion that there must be a grammatical relationship between a collocation's component words (e.g. verb + object noun), and is another notion central to the phraseological approach to collocation (Cowie, 1998a; Howarth, 1998a; Nesselhauf, 2005). Structural integrity is less central in frequency-oriented work, but can be a criterion (e.g. Kjellmer, 1994), and many frequency-oriented studies focus on particular types of collocations, such as verb + noun or adjective + noun collocations, thus also attending to structure.

The dictionary-based lists seem to be consistent with the view that structural integrity is part of the concept of collocation, though there are differences between the three dictionaries. The BBI is very much committed to this feature, only including collocations that manifest one of eight types of grammatical collocations (e.g. noun + preposition or noun + *to* + infinitive) and seven types of lexical collocations (e.g. adjective + noun or noun + verb). The OCD2 also uses a system of types of combination; its noun entries including seven types (e.g. verb + noun or preposition + noun). The FDCAE, reflecting its frequency orientation, is the least wedded to this feature, but nonetheless arranges its noun entries as verb collocates, adjective collocates and noun collocates, and also indicates whether individual collocates typically occur before or after the headword.

The corpus-based lists take no heed of structural integrity. The +/-4 span used in the searches means that the collocates may occur in any arrangement with the node, and indeed a collocate may have made the list due to it occurring in multiple slots. It should be noted, however, that nodes were specified as nouns when the searches were made.

In developing the two norms-based lists, no particular attention was given to structural integrity. As noted previously, however, the explanation for informants defined collocations as words "which are frequently used together", which may arguably be seen as implying that collocations have some sort of structural-unit status. Furthermore, the format of *LexCombi*, in which the cue appears with space either side of it in which a response can be made, was intended to prompt informants to think in terms of the linear arrangement of words, thus also making it likely that the responses given form structurally integral units with the cue. Indeed, looking at

both the L1 norms lists and the L2 norms lists, informants did often seem to give responses that form structural units with the cues, such as verb and object constructions. Thus, while the norms-based lists may in principle take no account of structural integrity, in practice the collocates they contain may reflect the notion to a certain extent.

7.6.6 Language variety

The next issue is which variety of English should serve as the basis for identifying collocations (as discussed in Section 2.3.6): that is, whose collocations the lists should include. In spite of debates over the "native-speaker model" and much research into ELF, research on L2 collocation mostly uses L1-user language production as the basis for comparison with little discussion (though Henriksen (2013) and Nesselhauf (2005) acknowledge the issue). Another aspect of this is the influence of culture on collocation. This has attracted some attention of a descriptive nature in the wider field of phraseology (e.g. Ooi, 2000; Skandera, 2007), but its importance to the evaluation of learners' production of collocations has not been recognised.

The dictionary-based lists combine US and UK sources. The corpus-based lists likewise. The L1 norms lists meanwhile are derived solely from the responses of British informants. In contrast, the L2 norms lists, drawn from the responses of Japanese users of English, were developed in acknowledgement of critiques of the notion that L1 users are or should be the model for language learners. It might, however, be questioned in some quarters whether the collocates provided by the L2-user informants count: whether they are indeed collocates. One answer to this query is provided by comparing the overlaps between the L2 norms lists and the L1 norms lists with the dictionary-based and corpus-based lists. As Tables 7.3 and 7.4 showed, these overlaps are very similar, with, if anything, a slightly greater degree of overlap for the L2 norms lists. That is, the L2 norms lists appear to have as much similarity with the dictionary-based and corpus-based lists as the L1 norms lists. Given these facts, the L2 norms lists would appear to have as great a claim to provide a representation of collocation as the L1 norms lists.

The L2 norms lists were also intended to recognise the influence of culture on collocation. As previously noted, Table 7.5 showed that for the cue NAME a single

collocate appears in the other three primary lists but not in the L2 norms list: *christian*. Likewise for the cue TABLE, *snooker* and *league* are in the other three primary lists but not in the L2 norms list. Conversely, and more likely to affect *LexCombi* scores, collocates appearing exclusively in the L2 norms lists are *golden* for the cue WEEK and *festival* for the cue SCHOOL, both clearly deriving from Japanese culture: *Golden Week* (Japanese: ゴールデンウィーク *Gōruden wīku*) being a holiday period in early May; *school festivals* being a major annual event for every school in the country and important community events. The influence of culture on the lists is clear.

7.6.7 Textual and mental views of collocation

In Section 2.3.7, it was seen that collocation can be viewed as a property of language or as a property of our linguistic knowledge. That is, an approach may imply that language is an artefact, in which collocation is one phenomenon that can be observed, or language may be treated as a form of knowledge in the brain, knowledge of collocation being one part of that. Most research on collocation is aligned with the former view. This is very much apparent in the frequency-based approach to collocation, with its focus on corpora of texts, but also somewhat in the phraseological approach which, while involving intuition and thus arguably psycholinguistic knowledge, comes very much from the tradition of descriptive linguistics.

The dictionary-based approach used here is somewhat awkward to define, since it involves the combination of three dictionaries each with their own approach. The FDCAE is very much frequency-oriented, the OCD2 likewise though with a seemingly stronger role for the compilers, while the BBI is more a product of intuition and is also rooted in the phraseological approach to collocation. Nonetheless, it seems reasonable to describe the dictionary-based approach as viewing collocation as a property of language. The corpus-based approach is very much textual and statistical and thus also views collocation as a property of language. The two norms-based approaches, in contrast, may be argued to view collocation in more psychological terms as a feature of linguistic knowledge (indeed, this was one motivation for adopting these approaches; see Section 7.2.3). Certainly, the norms-based approaches parallel word association studies, which aim to discover

links in the mental lexicon.

A related, but slightly different point is the contrast between language as idiolect (i.e. the linguistic knowledge of an individual) and language as a communal entity (i.e. the shared linguistic norms of a community). *LexCombi* essentially seeks to discover to what extent the idiolect of a learner conforms with the language of the community. Two points follow. First, it should be recognised that we can ask the same question of anyone, not just learners, since the idiolect of any individual will not conform precisely with the communal language. Second, the scoring approaches involve different ways of accessing the communal language. The dictionary-based approach involves the considered interpretation of corpus data by expert members of the community. The corpus-based approach looks at language produced by the community and counts occurrences. The norms-based approach asks members of the community to perform the task, and only includes responses given by at least two members of the community. In addition, the two norms-based approaches involve different conceptions of who forms the community of users. Thus, the four scoring approaches differ from each other along two distinct planes, as shown in Figure 7.2. In theory, it would also be possible to have two further scoring approaches to fill the two gaps in the figure: a Japanese-users-of-English dictionary-based approach and a Japanese-users-of-English corpus-based approach. There is, however, no such dictionary currently in existence, and while there are Japanese learner corpora, there is no sizeable Japanese-users-of-English corpus.

Figure 7.2: Two contrasts between the four primary scoring approaches.

	L1 users	**L2 users of English (specifically Japanese)**
Expert opinion of language produced by the community	dictionary-based scoring	
Language produced by the community	corpus-based scoring	
Task performance of members of the community	L1 norms scoring	L2 norms scoring

7.6.8 The scoring approaches' conceptualisations of collocation: A summary

The above sections have explored the implicit definitions of collocation provided by the four primary scoring approaches. This discussion can be summarised as follows:

The dictionary-based approach implies that:

- Collocation occurs between lemmas;
- Collocation involves lexical words primarily, but also functional words;
- Collocation involves separate orthographic words;
- Collocations are not necessarily semantically opaque;
- Collocations are word combinations with structural integrity;
- Collocation is about the usage and intuitions of L1 users;
- Collocation is a property of language.

The corpus-based approach implies that:

- Collocation occurs between word forms;
- Collocation involves lexical and functional words;
- Collocation involves separate orthographic words;
- Collocations are not necessarily semantically opaque;
- Collocations do not necessarily have structural integrity;
- Collocation is about the usage of L1 users;
- Collocation is a property of language.

The L1 norms approach implies that:

- Collocation occurs between lemmas;
- Collocation involves lexical and functional words;
- Collocation involves separate orthographic words and the co-occurrence of word parts;
- Collocations are not necessarily semantically opaque;
- Collocations are word combinations with structural integrity;
- Collocation is about the intuitions of L1 users;
- Collocation is about links in the mental lexicon.

The L2 norms approach implies that:

- Collocation occurs between lemmas;
- Collocation involves lexical and functional words;
- Collocation involves separate orthographic words and the co-occurrence of

word parts;

- Collocations are not necessarily semantically opaque;
- Collocations are word combinations with structural integrity;
- Collocation is about the intuitions of L2 users;
- Collocation is about links in the mental lexicon.

The multiply listed and ALC approaches combine the above features and therefore imply that:

- Collocation can occur between word forms or between lemmas;
- Collocation involves lexical and functional words;
- Collocation involves separate orthographic words and the co-occurrence of word parts;
- Collocations are not necessarily semantically opaque;
- Collocations may have but do not necessarily have structural integrity;
- Collocation is about the usage and intuitions of L1 users and the intuitions of L2 users;
- Collocation is both a property of language and about links in the mental lexicon.

It must now be determined which of the approaches is most appropriate for scoring *LexCombi* responses, the subject of the next section.

7.7 Discussion and conclusions

This chapter has explored a variety of approaches to determining the canonicity of learners' responses to *LexCombi*. After explaining the motivations for looking into this issue (Section 7.1) and introducing the six approaches (Section 7.2), this exploration covered three areas:

- how similar the approaches are in terms of words which are and are not considered to be collocates for each cue (Section 7.3);
- what type of scores data is produced by scoring *LexCombi* responses of learners using the six sets of lists (Sections 7.4 and 7.5);
- what views of collocation are implied by each approach (Section 7.6).

This third area is key as it revealed the consequences of selecting each approach for the definition of collocation it implies. It also made it plain that there are distinct choices to be made among the scoring approaches. Approaches to collocation may

vary in at least seven ways, as Sections 7.6.1-7.6.7 detailed. The scoring approaches considered here do not, however, make it possible to pick and choose freely from among these variables to build a measure. Rather, the choice of one approach involves a constellation of these variables. The two secondary approaches meanwhile, in which the lists developed by the primary approaches are combined, involve a mixture of these variables.

As explained in Section 7.1, the motivation for investigating the scoring of *LexCombi* was three concerns about the approach used by Barfield and in the earlier chapters of this thesis. These were the appropriateness of the sources used in creating the lists, a lack of clarity in the criteria used to determine what the lists contain and the breadth of the lists. The first of these issues has been approached by examining L2-user norms as one approach to *LexCombi* scoring, with the conclusion that this approach provides a reasonable representation of collocation. The second has been addressed for all of the potential scoring approaches through detailed descriptions of the approaches in the above sections. The extent to which each approach deals with the third issue can be judged by their coverage of learners' recurrent responses (Section 7.4.1), which showed that the four primary scoring approaches covered very different proportions of the recurrent responses, while the ALC lists covered almost all of them.

Furthermore, as Section 7.1 explained, this thesis has adopted Durrant's (2014) definition of collocation and so the scoring approach adopted should be coherent with this definition. Central to the definition is the idea that collocations are units that "may not be understood or appropriately produced without specific knowledge" or that "occur with sufficient frequency that their independent learning will facilitate fluency" (p. 448). The scoring approach selected should then permit both word combinations which feature specialised uses of words and combinations which are frequent.

Going forward, the method for scoring *LexCombi* will be as follows. First, the chief means of scoring *LexCombi* will be the ALC lists approach. This approach is consistent with Durrant's definition of collocation and incorporates the widest view of collocation, both conceptually and practically, thereby giving the greatest opportunity for learners to display their developing knowledge of English collocations. It is considered that ALC scores give a global perspective on learners'

productive collocation knowledge by maximally accommodating all the different potential sources of collocation knowledge that they might be drawing on. The ALC scores will therefore be used in Chapter 8 to evaluate the quality of the 70 cues under trial in order to make the final selection of cues for *LexCombi*.

Second, the four primary scoring approaches will be used alongside ALC to give a more nuanced view of learners' performance on *LexCombi*. The differences between these approaches may allow differences in the development of learners' productive knowledge of collocations to be explored, which would not be possible using the ALC scores alone. It may be, for example, that at different points in development, there is faster or slower growth in the types of collocations measured by one particular means of scoring (as examined in Section 10.6).

After identifying *LexCombi* as a potentially useful instrument for exploring learners' productive knowledge of collocations (Chapter 3), the experimental phase of this thesis began by trialling and exploring *LexCombi* (Chapters 4 and 5). The previous chapter and this chapter then considered possible improvements to *LexCombi*, looking into its format and scoring respectively. The next chapter continues this quest and considers another aspect of *LexCombi* that requires attention: its cues.

Chapter 8 Revising the test cues

8.1 Introduction

The previous two chapters reported on efforts to improve *LexCombi* with respect to its format and scoring. This chapter describes efforts to improve *LexCombi*'s cues, and reports on the selection and trialling process for a new set of cues for *LexCombi*. A full revision of *LexCombi*'s cues was considered necessary as a number of problems were identified with Barfield's original 30 cues in the course of carrying out the studies reported in Chapters 4-6. This work on cue selection was conducted concurrently with that reported in Chapter 7 on scoring *LexCombi* responses and made use of the same data set, which is described in Section 7.4.

Section 8.2 reports on four problems identified with Barfield's original *LexCombi* cues. Section 8.3 explains the processes undertaken to identify potential new cues. Section 8.4 describes the trialling of a number of those potential cues, in which Rasch analysis was used to explore their measurement properties and eliminate cues which performed poorly. Section 8.5 then reports on the final selection of cues for *LexCombi*.

8.2 Problems with the original *LexCombi* cues

Barfield's (2009a, p. 97) brief description of his selection of the original *LexCombi* cues is as follows: 50 nouns were taken from a list of the most frequent 500 lemmas in the BNC; these 50 nouns were piloted with 35 British L1 users of English and 35 highly proficient Japanese users of English; then the 30 nouns which best differentiated the responses of the two groups were selected as the *LexCombi* cues.

In the process of trialling *LexCombi*, as reported in Chapters 4-6, four problems became apparent with Barfield's original cues. These were: (1) a tendency for some participants to seemingly treat certain cues as verbs rather than as nouns (as discussed in Section 4.5.5); (2) a tendency for some cues to apparently be misread by participants; (3) clear relations between some of the cues, raising the possibility of interference between them; and (4) the presence of cues which performed poorly in statistical terms. Each of these problems is explained more fully below.

8.2.1 Noun cues treated as verb cues

Barfield's intention was to select 30 nouns as cues for *LexCombi*. However, since *LexCombi* presents cues to participants in isolation, and part of speech is a matter of word use rather than an inherent property of a word form, the cues may not be viewed as nouns by participants.

In my initial trial of *LexCombi*, as explained in Section 4.5.5, some responses seemed to invoke verbal uses of the cues rather than nominal uses, such as *him* and *somebody* for the cue SUPPORT, *for* and *with* for WORK, *something* for RESEARCH and *something* for EXPERIENCE. In subsequent rounds of data collection (reported in Chapters 5 and 6), the *LexCombi* instructions were altered to inform participants that all the cues were nouns, but this appeared to have little effect on the tendency for such responses to be given.

It seems likely that *LexCombi* participants view the cues simply as words, and consequently the noun status of the cues is dependent on the typical uses of each cue word (or an individual participant's knowledge or perception of its use). As was shown in Figure 4.3, verb occurrences of the cues SUPPORT and WORK are more frequent in the COCA than noun occurrences, while verb occurrences of the cue EXPERIENCE make up a quarter of its total occurrences. For all the other cues, in contrast, verb occurrences account for less than 10% of total occurrences. Thus, the prevalence of verb uses of these three cues may explain the way some learners responded to them. In the case of RESEARCH, in contrast, it was suggested that the issue may be that this word was simply not very well known by participants, and therefore participants were searching for any possible response that could be given.

Given the above, it would seem necessary for *LexCombi* cues to be words that are used exclusively or almost exclusively as nouns.

8.2.2 Misread cues

In the data sets reported on in Chapters 4-6, there were indications that some of the cues were on occasion misread or misunderstood by participants. It was at times difficult to judge whether this was the case or not. However, two of Barfield's cues in particular were problematic: ROLE seemed sometimes to be misread as its homophone *roll*, as indicated by responses such as *toilet*, *ball* and *paper*; and LAW seemed to be sometimes misread as *raw*, as shown by the responses *egg*, *fish* and

meat, or as *low*, as the responses *pay*, *salary* and *score* appeared to indicate.

It should be noted that this problem accounted for only some of the responses to these cues. The majority of responses to, for example, ROLE, did not suggest that it was misread as *roll*. Nevertheless, these cues were considered problematic in that on a number of occasions participants had been led to provide responses which had no possibility of being scored as canonical, and therefore had missed opportunities to display their knowledge of collocations. For this reason, should this problem be observed in cues that have been trialled, those cues will be rejected.

8.2.3 Relations between cues

The third issue with Barfield's *LexCombi* cues is that there are relations between some of them. First, in some cases there are semantic relations between two cues, as with the antonyms LIFE and DEATH. Second, there are cues which may be collocates of another cue. On the basis of the OCD2, 24 of the 30 cues are canonical collocates of at least one of the other cues, and in total there are 60 instances of this kind. For example, the cues HOUSE and LIFE are both given in the OCD2 as collocates of the cue COUNTRY. Third, there are cues which were recurrent responses of another cue. Based on the Chapter 6 data, consisting of *LexCombi* responses from 125 learners, there were three cases in which at least 10% of those learners gave one cue as a response to another. Specifically, *car* was given as a response to POLICE by 18 participants, *problem* as a response to HEALTH by 21, and *work* as a response to HOUSE by 33 (CAR, PROBLEM and WORK all being cues).

Strong relations between cues are problematic because the presence of one cue may have an impact on the responses to the other. That is, a cue may prime certain items and so increase the likelihood of those items being given as responses to another cue, while at the same time decreasing the likelihood of other potential responses which were not primed. This may mean that an accurate view of learners' knowledge of collocations is not gained, while cues affecting one another is problematic for statistical tests for which each item must be independent of others. In practice these problems are very difficult to eliminate entirely, but it would seem necessary to avoid them as far as possible.

8.2.4 Poorly performing cues

The final problem with Barfield's cues was the poor performance of some of them in statistical terms with regard to *LexCombi* scores. On the basis of the *LexCombi* scores for the Chapter 6 data from 125 learners, classical test statistics showed that 13 of the 30 cues had item-total correlations below .3, a benchmark often used in judging whether an item is operating in concert with a test as a whole (Field, 2009).

There were, then, a number of problems with Barfield's *LexCombi* cues, and a total of 20 of the 30 cues were affected by at least one of the above four issues. It was thus determined that a full revision of the *LexCombi* cues should be undertaken in order to produce a new selection of 30 cues. That is, the work reported in this chapter was in response to the question:

- Which words should be used as cues for *LexCombi*?

8.3 Identifying potential cues

In order to identify potential *LexCombi* cues, five criteria were employed. The following gives details on their operationalization and the exact processes used.

1. *Potential cues should be of high frequency*. High-frequency cues were sought since the intention is that *LexCombi* be a measure of collocation knowledge rather than of knowledge of the cues themselves. That is, it is intended that all the *LexCombi* cues will be known to participants to some extent and what is at stake is whether a participant is able to produce canonical collocates for these words. This is important since Section 3.3 highlighted how, in a number of previous studies, measures of knowledge of collocations have been confounded by the issue of whether learners have knowledge of the component words in those collocations.

   High-frequency cues will, in addition, make it possible for *LexCombi* to be used with learners of a wide range of proficiency levels. Cues of high frequency are necessary if *LexCombi* is to be used with learners of lower proficiency. Yet at the same time, given that Nesselhauf (2005) found that advanced learners' problems with collocation were primarily with the uses of common words, it is believed that highly frequent cues will also enable advanced learners to display their collocation knowledge.

   The first step in selecting potential cues was therefore to identify nouns

of high frequency. This was achieved by taking all the nouns from the JACET8000 (Ishikawa et al., 2003) level 1 vocabulary list: a frequency-based list of lemmas based in part on the BNC which also takes into account the uses of English in Japan. This resulted in an initial list of 610 candidate words. Among these words were all 30 of Barfield's original cues. To further verify the high-frequency status of these candidates, words were eliminated if they did not also appear in both the first 1,000 items in the General Service List (West, 1953) and the first 1,000 items in Nation's BNC-based word lists (Nation, 2006a). Also at this stage, the word LAW was removed from the list of candidates since it had previously been found to be misread by learners (the other word previously found to be misread, ROLE, had already been removed). This reduced the list to 394 candidates.

2. *Cues should typically occur as nouns*. Section 4.5.5 reported some problems with cues which regularly occur both as nouns and as verbs. Therefore, cues were sought which are principally used as nouns.

   The focus on nouns has two bases. First, restricting the cues to a single word class makes the construct of *LexCombi* simpler. Second, there have been several suggestions regarding the primacy of nouns with reference to collocation, as Section 4.5.5 described. In addition, verb + noun collocations and adjective + noun collocations have attracted the most interest in studies of collocation, and thus the use of noun cues can be seen as following this research tradition, enabling comparisons with previous work.

   The next step was thus to ensure that each candidate cue is primarily used as a noun. The COCA was used, and the total number of occurrences of the word form and the number of noun occurrences checked. Any words for which less than 90% of the occurrences were noun uses were deleted from the list. This reduced the list to 196 candidates.

3. *Cues should not have a tendency to elicit paradigmatic responses in word association studies*. Word association studies have found certain words are more likely to elicit paradigmatic associations, while others tend to elicit syntagmatic (i.e. collocational) associations. Furthermore, as Section 4.5.3 explained, one concern with *LexCombi* was that it was eliciting associations

in general from learners rather than collocations in particular. While changes to the *LexCombi* format were made in an attempt to mitigate this issue (see Chapter 6), it was also thought sensible to avoid cues which may tend to elicit paradigmatic responses.

The third step in selecting potential cues was therefore to eliminate words of this sort. This was done by checking each candidate in the *Edinburgh Associative Thesaurus* (EAT) (Kiss, et al., 1973). Words were eliminated if any of the following conditions were met: (a) the top five EAT responses contain paradigmatic associations accounting for 30% or more of the total responses; (b) the top five responses are all paradigmatic; or (c) a single paradigmatic response accounts for 20% or more of the total responses. In addition, the word UNIVERSITY was removed from the list at this stage since, when previously used as a practice item in *LexCombi* data collection, it had been observed to frequently elicit the names of universities, and the same tendency was also seen in the EAT data. At this stage, the list of candidates stood at 108 items.

4. *Any potential cues for which data has previously been collected should not show poor performance*. Poorly performing cues reduce the reliability and thus also validity of an instrument. Of the 108 remaining items, 12 were among Barfield's 30 *LexCombi* cues, and since data had already been collected using these cues, the performance of these words was checked. Based on the Chapter 6 data from 125 learners, item-total correlations below .3 were found for 3 of the 12. However, in only one case was the reliability of *LexCombi* improved if the item were deleted. This potential cue was thus eliminated from the list. This left 107 candidates, of which 11 were among Barfield's *LexCombi* cues.
5. *Cues should not have a strong relation with any potential cues for which data has previously been collected*. Relations between cues may mean that one cue affects the responses to a subsequent cue. As explained above, this may have consequences for the insights gained and for statistical testing.

The final step was thus to consider potential relations among the cues. The 11 remaining original *LexCombi* cues were used as the basis for this, since, as step four explained, they had already been found to perform

adequately. Three criteria were applied: (1) Words occurring as collocates of these 11 words in the OCD2 were excluded; (2) Words occurring as strong associations of these 11 words in the EAT (defined as a word accounting for 5% or more of EAT responses) were excluded; (3) Words which were recurrent responses to the 11 words in either the Chapter 4 data or in the Chapter 6 data (defined as words which 10% or more of participants provided as a response) were excluded. These three criteria were first applied to the 11 words with respect to each other. This led to the elimination of one word. The three criteria were then applied to all the remaining possible cues with respect to the 10 remaining original *LexCombi* cues.

Note that a full application of this procedure to all the cues under consideration with respect to each other was not implemented at this stage. This was because doing so could have led to good cues being excluded as a result of links to cues that were later discarded. Therefore, the remainder of this procedure was postponed until the larger set of cues had been trialled and whittled down further. That shortlist was then tested on the three criteria (see Section 8.5).

After applying the criteria above, 88 candidates remained on the list. Unfortunately, two words which should have remained on the list were removed in error at this stage, while two words which should have been excluded remained mistakenly on the list.

The above process thus produced a list of 88 candidates for *LexCombi*, of which 10 were Barfield's original cues. In summary, these 88 words have the following characteristics. They:

- are highly frequent, being drawn from the 1K frequency band;
- are used principally as nouns, with over 90% of their occurrences as nouns;
- and do not show a tendency in word association studies to elicit primarily paradigmatic responses.

In addition, the 10 remaining original cues:

- have been found to perform adequately in statistical terms in previous *LexCombi* data;
- do not have strong relations with each other, in terms of being collocates of

each other, associations of each other, or recurrent responses to each other;

- and appear to be rarely misread by learners.

Furthermore, the other 78 candidates (barring the two mistakenly included):

- do not have a strong relation with any of the 10 remaining original cues, in that they are not collocates of these 10 cues, nor associations of them, nor recurrent responses to them.

8.4 Trialling of potential cues

In order to trial as many potential cues as possible, the 10 remaining original cues were used as a foundation and three sets of 20 further candidates assembled to create three versions of *LexCombi* each with 30 cues. In this way 70 candidates could be trialled in total. The three sets of cues are shown in Figure 8.1. The appearance of 10 cues in all three versions was designed to allow the similarity of participants completing each version to be checked, and therefore enable comparisons of all 70 cues. The 60 new cues to be trialled were distributed across the three sets following their frequency order in the JACET8000 list. That is, the first new cue in the first set,

Figure 8.1: The three sets of cues trialled.

Cue set 1		**Cue set 2**		**Cue set 3**	
country	war	country	war	country	war
family	power	family	power	family	power
problem	reason	problem	reason	problem	reason
question	paper	question	paper	question	paper
car	decision	car	decision	car	decision
school	situation	name	fish	eye	action
water	machine	book	mile	room	rate
money	stage	week	condition	fact	player
student	station	food	product	age	trouble
company	glass	letter	floor	class	size
sense	leader	picture	church	teacher	page
member	box	society	relationship	party	worker
table	science	music	blood	nature	bank
bed	character	field	trade	piece	scientist
space	development	amount	training	wall	ball

Note. The 10 remaining original cues are shown first, then the 20 new candidates for each set below.

SCHOOL, had the highest frequency rank among these words, the first in the second set, NAME, was the second highest ranked, and so on. The 10 remaining original cues all had frequency ranks higher than the least highly ranked of these new cues (BALL). Unfortunately, the two words mistakenly included in the list of candidates were included here: MEMBER and PIECE. In analysing the cue trialling data, these two items were thus excluded from consideration.

The three sets of cues were trialled with students from two universities in Japan. The data set consisted of the responses of 223 Japanese L1 university students to two instruments: a Yes/No vocabulary test and *LexCombi*. This data set was also used in exploring different scoring approaches for *LexCombi*, as reported in Chapter 7, and full details on the data set are given in Section 7.4.

For the current purpose of examining the performance of the cues, the responses were scored using the ALC lists (see Chapter 7). These lists are a combination of four different approaches to defining collocation, and thus are a broad, generous approach to defining canonicity.

The potential cues were assessed both quantitatively and qualitatively. The following sections detail the analyses and findings.

8.4.1 Quantitative assessment of the potential cues

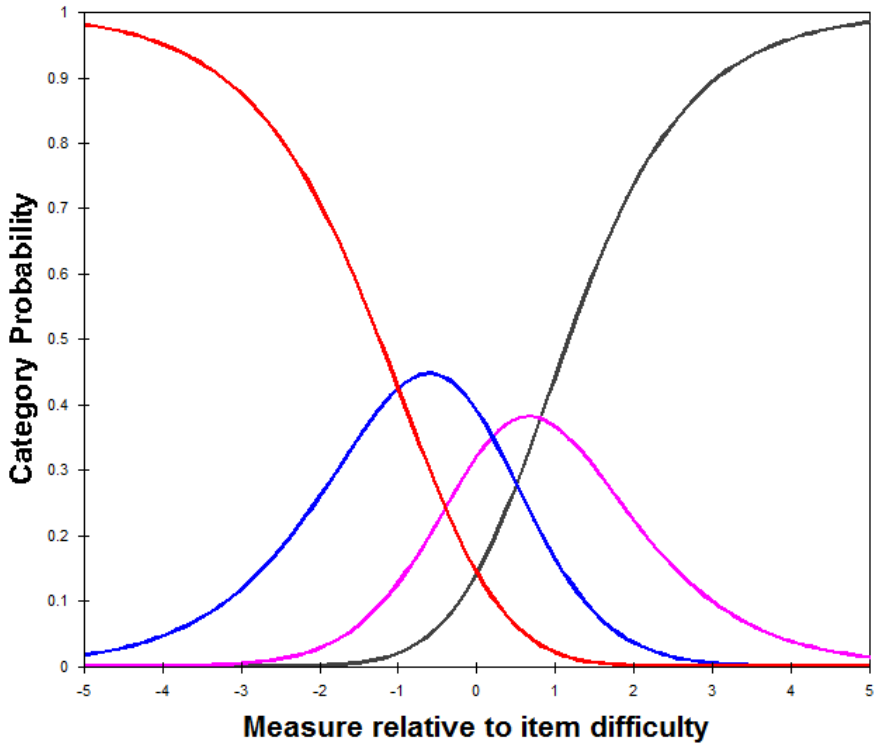
Rasch analysis was employed to analyse the performance of the cues in statistical terms, using Winsteps (Linacre, 2015). Rasch analysis, rather than classical test statistics, was used for two reasons. First, the structure of the data set, in that it was not possible to have each participant respond to all 70 cues under trial, would pose considerable problems for classical test statistics. With Rasch analysis, by contrast, there is no such issue, since the 10 cues in common across each version of *LexCombi* allowed the measures to be calibrated against one another. Second, Rasch allows "person-distribution-free" item estimates (Bond & Fox, 2007; McNamara, 1996). This means that, provided that the data fulfil the requirements of the model, the estimates of item difficulty are not dependent on the sample of participants that supplied the data. That is to say, a different sample of participants from the same population would be expected to produce the same results (within a margin of error). Rasch is therefore particularly suited to prospective work, such as cue selection, as opposed to the retrospective orientation of classical test statistics.

In assessing the quality of the cues, two features of Rasch analysis were used: item fit statistics and Rasch-Andrich thresholds. As already mentioned in Section 7.5, item fit statistics indicate whether an individual item "fits" with all the items as a set. The Rasch model requires that a data set involves a single construct and the fit statistics indicate "whether each item contributes to the measurement of only one construct" (Bond & Fox, 2007, p. 35). Items that do not fit are termed misfitting. There are two forms of misfit: underfit and overfit. Underfit refers to items which do not meet the expectations of the model. Such items therefore degrade the quality of the measurement. Overfit refers to items that match the expectations of the model too closely. This may seem counter-intuitive, but the Rasch model, being probabilistic, expects some degree of variability in the data. Overfitting items do not display this. Consequently, overfitting items do not degrade the quality of the measurement, but may lead to an overly positive assessment of the instrument.

Following the recommendations of Wright and Linacre (1994), a two-stage approach was taken to evaluating the fit of items. In the first step, underfitting items, defined as those with an infit mean squares > 1.2, were removed from the data set. These are items for which some participants achieved either surprisingly high scores or surprisingly low scores given the overall ability level of the participant and the overall difficulty level of the item. This resulted in the removal of two items: BED (Infit mean squares = 1.24) and TRAINING (Infit mean squares = 1.21). The second step was to re-analyse the data, the removal of items at step one causing a re-calibration of the statistics, and then remove any newly underfitting items, again those with an infit mean squares > 1.2, as well as any overfitting items, defined as those with an infit mean squares < 0.8. This second step found no items requiring removal. It should be noted that these fit criteria are relatively strict and thus items that met the criteria can be seen with some confidence to be performing well. It may also be noted that the fit criteria used here were different from those used in Section 7.5. This is because the purpose of the analysis, in this case cue selection, required more stringent criteria than the global view of performance that was sought in Chapter 7.

The second feature of Rasch used in assessing the cues was Rasch-Andrich thresholds. These thresholds are the points at which the probability of obtaining one score match that of obtaining a different score. To illustrate, Figure 8.2 displays

Figure 8.2: Probability category curves for the item COUNTRY (see text for details).



Note. Red = probability of a score of 0; blue = probability of a score of 1; purple = probability of a score of 2; black = probability of a score of 3.

probability category curves for the item COUNTRY. The graph's x-axis shows ability (i.e. ability on *LexCombi*) in comparison to the difficulty of the item in logits, and the y-axis shows the probability of a response. The curves indicate the probability of each score being achieved by a participant of a particular ability level. Since participants were asked to give three responses to each cue, there were four possible scores, or categories, for COUNTRY (i.e. a participant could achieve a score of 0, 1, 2 or 3), and so there are four probability category curves. Thus, for a participant with an ability two logits below the difficulty of the item, there is a probability of approximately .7 that they will achieve a score of 0 (red curve) on this item, a probability of slightly under .3 that they will achieve a score of 1 (blue curve), a probability of perhaps .03 that they will achieve a score of 2 (purple curve), and a very small probability that they will achieve a score of 3 (black curve). It can therefore be seen that whichever curve is highest at any point along the x-axis is the most probable score at that point. The Rasch-Andrich thresholds are the points at which the probability category curves meet. Thus, for this cue, the Rasch-Andrich

threshold between the category (score) of 0 and the category (score) of 1 was at -1.00 logits, the threshold between 1 and 2 was at 0.20 logits, and the threshold between 2 and 3 was at 0.81 logits.

The Rasch-Andrich thresholds for the item shown, COUNTRY, were ordered as we would expect. That is, the 1-2 threshold (the threshold between scores of 1 and 2) was at a greater level of difficulty than the 0-1 threshold, and the 2-3 threshold was at a greater level of difficulty than the 1-2 threshold. It is possible, however, for Rasch-Andrich thresholds to be disordered. This occurs when one of the categories (scores) is relatively under-represented in the data, such that at no point on the x-axis does it become the most probable score.

Items with disordered Rasch-Andrich thresholds are considered to deviate from expectation in that the probability of achieving a higher score does not correlate straightforwardly with ability (Tennant, 2004). Such items were therefore eliminated from the list of potential cues. Two cues were eliminated on this basis: WEEK and PAGE.

The disordered Rasch-Andrich thresholds for these two cues may have been caused by the chaining of responses (a phenomenon also observed in the think-alouds study; Section 5.6.2). In the case of WEEK, the under-represented category was a score of 1: that is, relatively few participants achieved a score of 1 compared with the number that achieved scores of 0, 2 and 3. The recurrent responses (i.e. responses given by at least 10% of the participants) give some insight into how participants responded to a cue. For WEEK, the recurrent responses, which were all canonical, were *end* (given by 34 participants), *last* (30), *next* (18), *this* (12), *holiday* (11), *one* (9) and *every* (8). Three of these responses, *last*, *next* and *this*, form a lexical set, and it is possible that this could have prompted chains of responses. That is, a participant may have given one of these responses, and this led them to give one or both of the other two members of the set, meaning relatively few participants achieved a score of 1. In the case of PAGE, the under-represented category was a score of 2. The recurrent responses for PAGE (all canonical collocates) were: *next* (25), *turn* (21), *first* (10), *last* (9) and *one* (8). Here, there was also a lexical set, albeit smaller and recurring less frequently, involving *first* and *last*. Since the under-represented category was a score of 2, it is possible that a number of participants gave one of the other recurrent responses or another canonical response, as well as

one member of the lexical set which led them on to the other member of the set.

Thus the quantitative, Rasch-based analysis of cue performance resulted in the elimination of four candidates out of the 70 that were trialled.

8.4.2 Qualitative assessment of the potential cues

The qualitative assessment of the cues involved looking for instances of the types of problems identified in Sections 8.2.1 (cues invoking verbal uses of a cue) and 8.2.2 (cues being misread). This was done informally at the time when the *LexCombi* responses were typed up, and later by examining a list of responses for each cue.

Three candidates which had been trialled were eliminated as a result. First, the cue TRADE attracted a number of responses which appeared to invoke its verbal uses rather than its nominal uses, such as *money* and *something*. Second, the cue GLASS appeared to have been misread by a number of participants as *grass*, as shown by the responses *field*, *green*, *hopper* and *seed*, in keeping with *l*/*r* confusion among Japanese learners of English. Finally, the cue STATION was observed to elicit a number of proper names, with one local station name appearing as a recurrent response for this cue (as mentioned in Section 7.4.1).

8.5 Cue selection

The quantitative and qualitative evaluations of 70 candidate words which had been trialled therefore resulted in the elimination of seven candidates. In addition, the two mistakenly included words had already been removed from consideration. Thus 61 candidates remained.

To select 30 cues from the 61, three factors were considered:

- Cues should be as evenly spaced as possible in terms of their location on the scale of difficulty (as determined by the Rasch analysis). This is because *LexCombi* is intended to allow the measurement of a wide range of abilities and having cues at evenly spaced difficulty intervals would allow learners anywhere along the range to be measured equally well.
- Cues must not be related to one another. This is to avoid the presence of one cue having an impact on the responses to another.
- Cues with larger Rasch-Andrich threshold ranges should be preferred. That is, preferred cues are those for which there is a greater difference between the difficulty of getting a score of 0 and a score of 3. This is because such

items assist in measurement across the range of abilities.

Based on the above, a three-step process was followed.

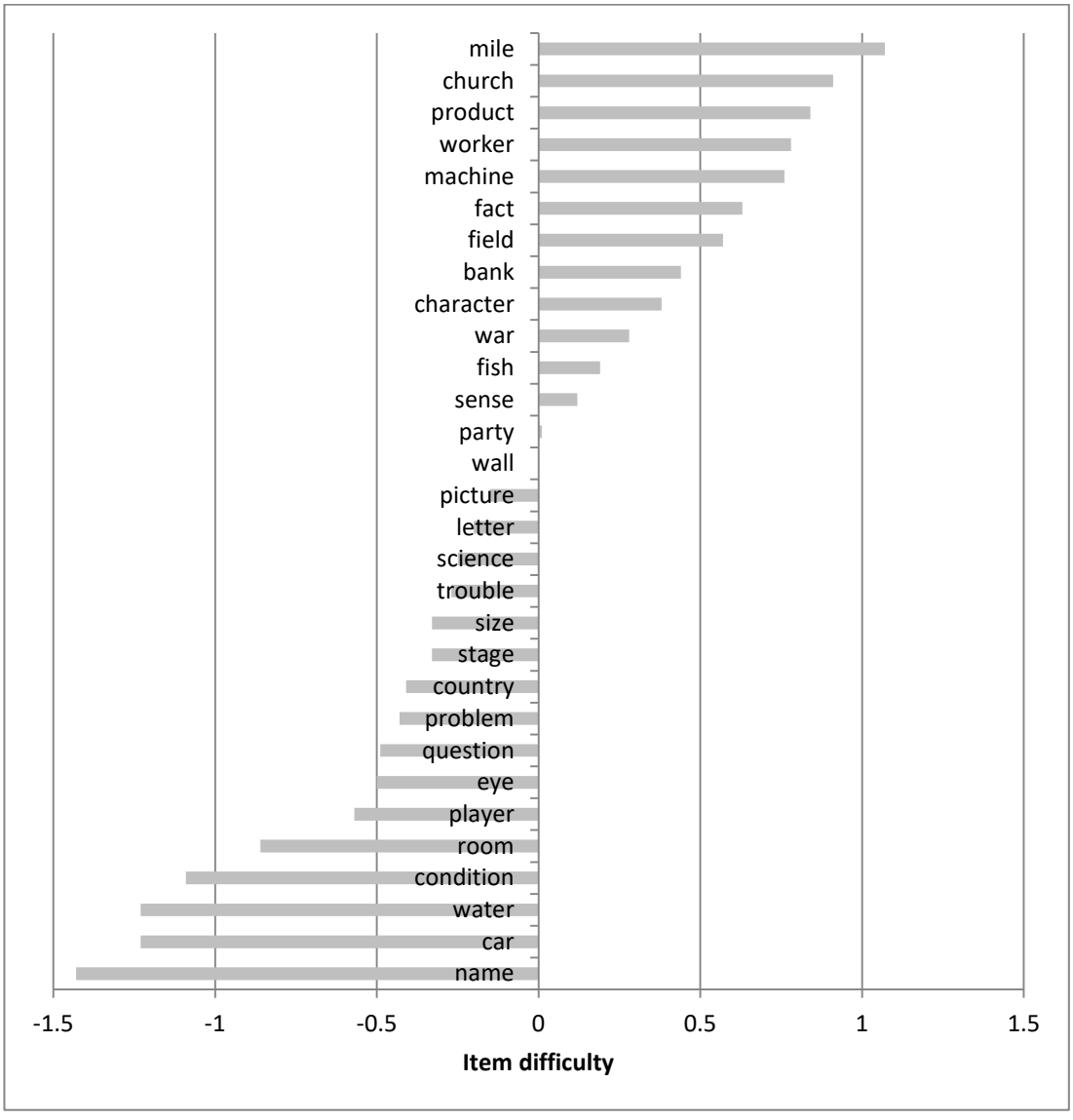
1. A new Rasch analysis was conducted including only the data for the 61 candidates under consideration. The difficulty of each item was found and the range between the most and least difficult items calculated. This range was 2.50 logits. Thirty evenly spaced points along that range were then identified as ideal locations for cues. With the range of 2.50 logits, the 30 locations were spaced at approximately 0.09 logit intervals. The 61 candidates were then fitted against the ideal locations (in terms of their item difficulties) such that each was as close to the nearest ideal location as possible. As may be expected, the candidates were clustered towards the centre of the range of difficulty, with relatively few at the extreme locations.
2. Potential relations between the 61 candidates were checked for. Three types of relation were investigated, using the same criteria as adopted when identifying potential cues (Section 8.3), this being the full application of the procedure that was postponed at that point. The first check was for whether a candidate was a collocate of another candidate. Candidates were considered collocates if listed in the OCD2. Second, it was checked whether a candidate was a strong association of another candidate, defined as a word accounting for 5% or more of responses in the *Edinburgh Associative Thesaurus*. The third check was for candidates which were a recurrent response (a word which 10% or more of participants provided as a response) to another candidate (see Section 7.4.1). In each case where there were relations between two candidates, one was eliminated. When determining which to eliminate, candidates located at the extremes of the range of difficulty were retained, since there was a surplus of candidates in the middle of the range and fewer at the extremes.
3. The above two steps resulted in there being only a single remaining candidate at some of the locations along the range of difficulty, and so those candidates were selected as cues. Particularly in the middle of the range of difficulty, however, there were still selections to be made from among two or three candidates at various locations. Where this was the case, the candidate with a larger Rasch-Andrich threshold range was selected.

The above processes led to the selection of the 30 words in Figure 8.3 as the final *LexCombi* cues. Figure 8.4 gives the cues in difficulty order, and shows their

Figure 8.3: The final selection of *LexCombi* cues.

bank	eye	mile	product	stage
car	fact	name	question	trouble
character	field	party	room	wall
church	fish	picture	science	war
condition	letter	player	sense	water
country	machine	problem	size	worker

Figure 8.4: Item difficulty of the cues selected (most difficult at top).



location on the scale of difficulty. As can be seen, the items are not perfectly evenly spaced, but there is good coverage across the entire range of difficulty.

8.6 Conclusion

Barfield's original 30 *LexCombi* cues suffered from a number of problems which may have prevented learners from fully displaying their productive knowledge of collocations and raised doubts about the quality of the data collected using them. As a result, a full revision of *LexCombi*'s cues was undertaken in order to determine which words should be used as cues for *LexCombi*. This involved: (1) a principled selection of potential words to act as *LexCombi* cues; (2) the trialling of 70 of these candidates with 226 learners, followed by both a Rasch-based quantitative analysis of their performance as well as qualitative analysis; and (3) a principled final selection of 30 words to serve as *LexCombi* cues. These 30 words:

- are highly frequent, appearing among the first 1,000 items in three separate frequency lists;
- are used overwhelmingly as nouns, with over 90% of their occurrences in the COCA being noun uses;
- do not primarily elicit paradigmatic responses in word association studies;
- have been found, through trialling and Rasch analysis of the results, to produce data with strong measurement properties;
- appear to rarely invoke verbal uses, be misread by learners, or elicit proper names;
- are not strong collocates or associations of each other;
- and are well-spaced along the range of difficulty.

Consequently, these cues are believed to offer the potential to elicit data of high quality from learners, enabling more reliable insights into the development of their productive knowledge of collocations. Chapter 9 outlines a new set of data collected with this cue set with the aim of verifying the quality of *LexCombi* in its revised form.

Chapter 9 Evaluating *LexCombi 2*

9.1 Introduction

Chapters 6-8 reported on the development of *LexCombi* in three areas. In Chapter 6, a revision to its format was made in an attempt to more directly elicit collocational responses. It was suggested that the adapted *LexCombi* format has some advantages in guiding participants towards giving collocational responses and in eliciting more responses from learners. Chapter 7 reviewed the scoring of responses in order to more explicitly address the issue of collocational definition, and with a desire to take a more inclusive approach towards learners' responses. It was concluded that ALC scoring should be the primary approach to scoring since it has the broadest view of collocation and so gives learners the greatest opportunity to demonstrate their productive knowledge of collocations. In Chapter 8, a new process of cue selection was undertaken in response to a number of problems identified with Barfield's (2009a) original set of cues. It was reasoned that the new selection of cues may provide data of higher quality from learners, and so enable more reliable insights into the development of their productive collocation knowledge. Given these substantial changes to the original *LexCombi* in three areas, the instrument in its final form will be referred to hereafter as *LexCombi 2*. For clarity, I will refer to the original version of *LexCombi* as *LexCombi 1* when it is directly contrasted with *LexCombi 2*. The term *LexCombi* will be used for generic reference to the instrument (i.e. where it would be misleading to single out a particular version).

This chapter reports a study in which a further set of data was collected in order to evaluate whether *LexCombi* 2, after the cumulative changes made to it in Chapters 6-8, is indeed a more reliable and valid instrument for the elicitation of learners' productive knowledge of collocations. That is, the study sought to put to the test the adapted *LexCombi* format (after Chapter 6), with the new set of cues (after Chapter 8), and with responses scored using the ALC lists (after Chapter 7). Specifically, the study addressed three questions:

1. *What was the psychometric quality of the data provided by* LexCombi 2*?* This involved evaluating the quality of the new data set. The analysis used the various tools of Rasch analysis and considered the difficulty of

LexCombi 2 items with respect to current and future possible participants, the dimensionality of the construct measured by *LexCombi 2*, the presence or absence or problematic items and the reliability of the measures. This analysis differs from that reported in Chapters 7 and 8 in that it uses a new set of data collected with *LexCombi 2* in its current form.

2. *How did the difficulty of the items in the current data set compare with their difficulty in the cue trialling data (Chapter 8)?* Part of the cue trialling process involved selecting items on the basis of their difficulty, as estimated by the Rasch analysis, with the aim of producing an instrument which would enable learners across a range of ability levels to be measured with similar levels of accuracy. The current data set offered an opportunity to examine the extent to which our expectations about the cues based on the cue trialling data were borne out. That is, confirmation was sought as to whether individual cues were in a similar position on the scale of difficulty in the current data set as compared with their position as seen through the Chapter 8 data set.
3. *Can the* LexCombi 2 *scores data be treated as an interval scale?* In the human sciences, raw scores from measures rarely constitute true interval scales, and one of the chief benefits of the Rasch model is that it allows true interval scales to be constructed from ordinal-level data (Bond & Fox, 2007; McNamara, 1996). However, it may be that a particular data set can be treated as an interval scale, which would allow the results to be reasonably presented as raw scores, with clear benefits for clarity and interpretability. Whether this was the case with the current data set was judged by comparing the raw *LexCombi 2* scores for each person with the Rasch person measures.

9.2 Method

Following the studies reported in earlier chapters, data was collected using two instruments: *LexCombi 2* and a Yes/No vocabulary test. As mentioned above, *LexCombi 2* has an adapted format, with a new set of cues, and the ALC lists are used to identify canonical collocates. The cues were arranged in six orders, distributed randomly to participants. The *LexCombi 2* responses were typed up,

cleaned up as explained in Table 4.1 and run through the CollCheck (Imao & Brown, 2014) program to produce the scores. Rasch analyses were conducted using Winsteps (Linacre, 2015).

The Yes/No test provided an estimate of vocabulary size and was intended to be an approximate measure of proficiency. The test was the same as described in Section 7.4 and is provided in Appendix E. Following Milton (2009), an estimate of vocabulary size was calculated by multiplying the number of real words checked by 50, and then subtracting 250 for each pseudoword checked.

The participants were Japanese university students from four universities. Data was initially collected from 193 respondents, with respondents excluded as follows: 9 had an L1 other than Japanese; 9 did not fully complete one or other of the instruments or indicated that they were unwilling for their data to be included in the study; and 29 displayed a high level of guessing in the Yes/No test. The set of data to be analysed consisted then of responses from 146 participants. The Yes/No vocabulary test showed these participants to have an estimated mean vocabulary size of 5,230 words (*SD* = 915) and the scores ranged from 3,050 to 7,800. Thus, by this measure, the participants varied quite substantially in proficiency, though even the lowest proficiency participants were by no means beginners.

9.3 Results

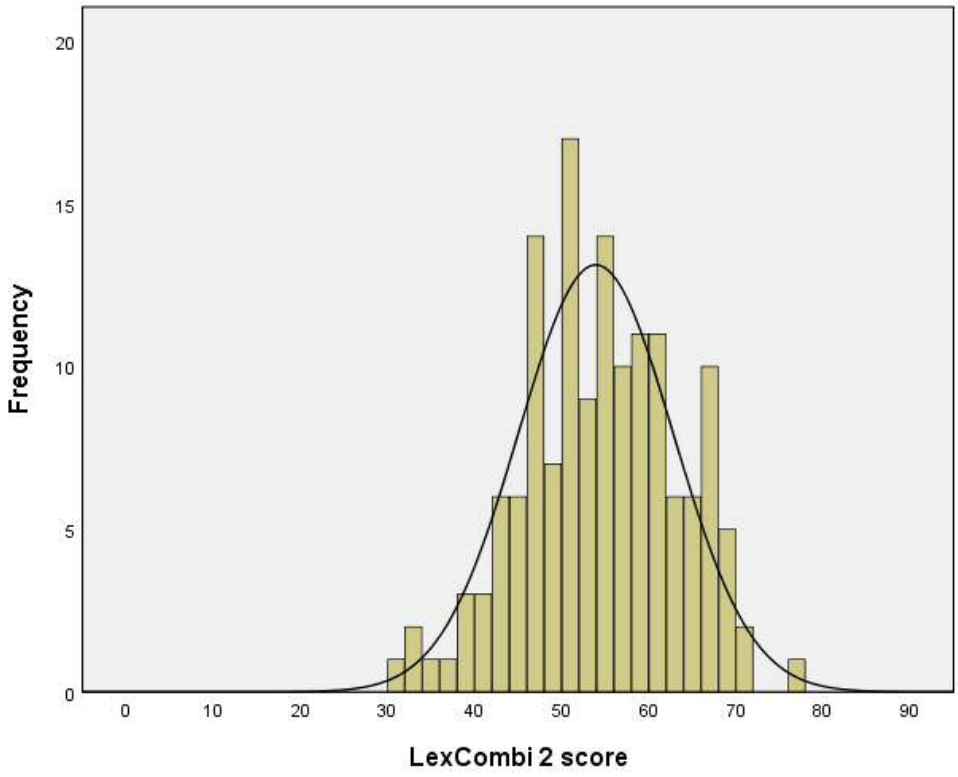
9.3.1 Descriptive statistics

Table 9.1 provides the descriptive statistics for the *LexCombi 2* scores, while Figure 9.1 shows a histogram of the scores. *LexCombi 2* scores can lie between 0 and 90 and the scores in this data set covered a reasonable degree of this distribution, albeit

Table 9.1: Descriptive statistics for the raw *LexCombi 2* scores.

	LexCombi 2 scores
N	146
Mean	53.92
SD	8.875
Minimum	31
Maximum	76

Figure 9.1: Distribution of the *LexCombi 2* scores (N = 146).



concentrated somewhat in the upper half of the range. A Kolmogorov-Smirnov test confirmed the scores data to be normally distributed. Also of note is that no ceiling or floor effects were evident.

9.3.2 Rasch analysis of the current data

As explained in Sections 7.5 and 8.4.1, Rasch analysis offers a range of insights into the statistical properties of a data set and allows the quality of the data to be evaluated in various ways. The analysis below followed the procedures used in the analyses of scoring approaches in Section 7.5. There was one possible scoring category not present in the data: for the item CAR, no participant achieved a score of 0. Consequently, a dummy record was added to the data set to force Winsteps to include this category, with this dummy record then excluded when conducting analyses.

Table 9.2 gives descriptive statistics for the Rasch measures. The mean person measure shows the performance of the participants relative to the *LexCombi 2* items in logits. By convention, the mean item measure is placed at zero. Therefore, the mean person measure of 0.43 indicates that the items were somewhat easy for these

Table 9.2: Descriptive statistics for the persons and items in logits.

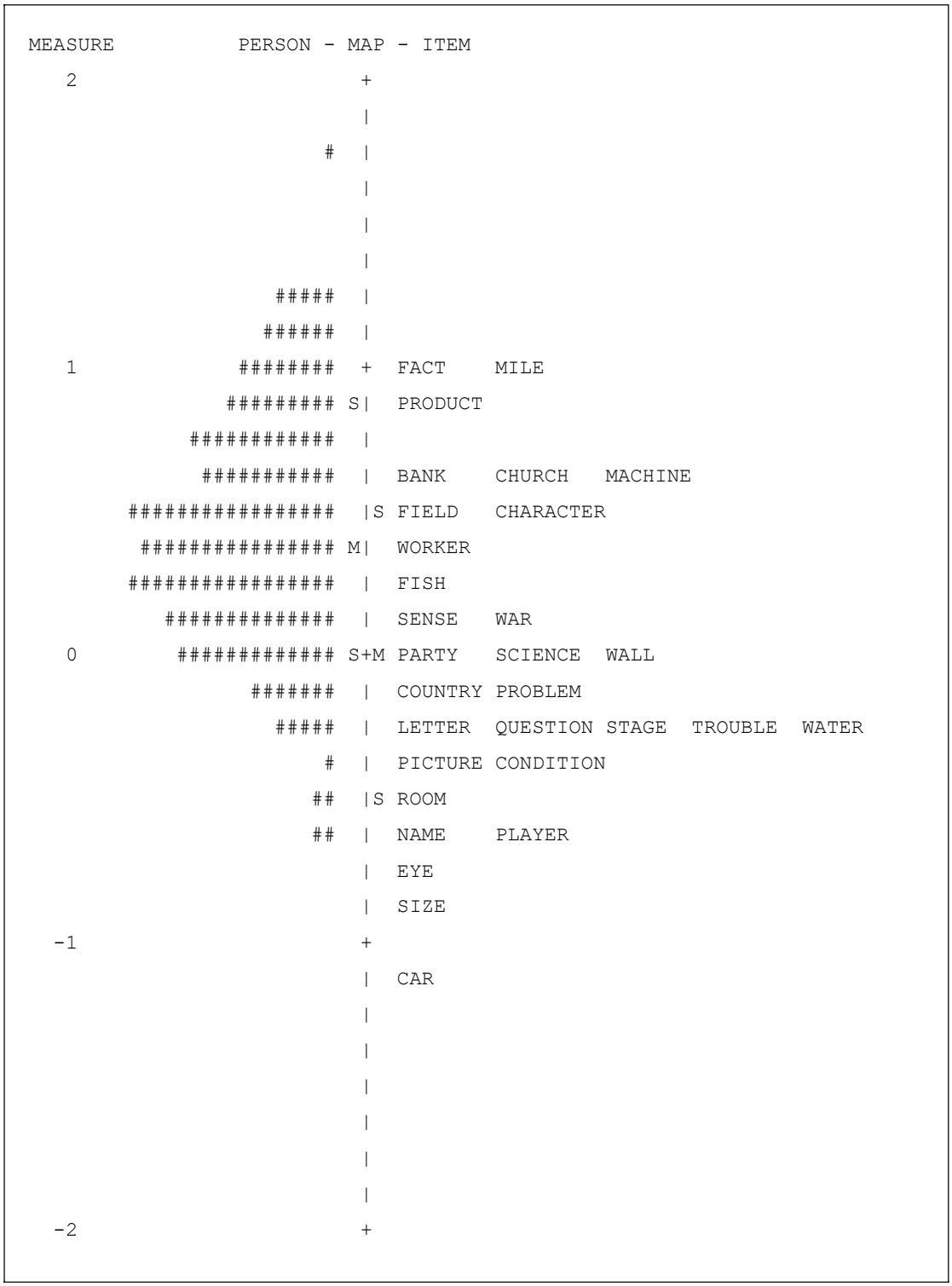
	LexCombi 2
Mean person measure	0.43
Person *SD*	0.44
Range of persons	2.40
Mean item measure	0
Item *SD*	0.55
Range of items	2.11

participants, in that it is above the average difficulty level of the cues. In Rasch analysis, a person at the same point on the scale as a particular item has a 50% probability of success on that item. With the mean person measure at 0.43, the probability of success on the mean item (with a measure of 0) was therefore above 50%, and in fact in this case the mean person had an approximately 60% chance of success on the mean item (Linacre, 2014).

A fuller picture of the alignment of the *LexCombi 2* items with the participants is provided by examining a Wright map of the data. Figure 9.2 shows the placement of each person and each item on the logits scale. The left side of the figure shows the ability of persons, with persons near the top being those that did best on *LexCombi 2*. The right side of the figure shows the difficulty of the items, items near the top being the most difficult.

Generally, the Wright map shows a good deal of overlap between the persons and the items, which aids the accuracy of the measures. It may appear, however, that there was a lack of items of sufficient difficulty for the most able participants. This is not in fact an issue since for each item there are four possible scores (0, 1, 2 and 3) and thus three difficulty thresholds (between 0 and 1, 1 and 2, and 2 and 3). That is, in addition to working out the overall difficulty of an item, we can also calculate the difficulty of achieving a score of 1 rather than 0, of achieving 2 rather than 1, and of achieving 3 rather than 2 (as explained previously in Section 7.5). The Wright map shows the overall difficulty threshold for each item, but in fact the upper difficulty thresholds (between a score of 2 and 3) for the most difficult items were some way above the ability level of even the most able participant and thus there was

Figure 9.2: Wright map of the 146 participants and 30 items.



Note. Participants appear on the left, with each # symbol representing one participant. Items appear on the right. The most able participants and most difficult items appear at the top of the figure, the least able participants and easiest items at the bottom. *M* shows the mean; *S* one standard deviation from the mean.

reasonable coverage even at this point on the scale. Similarly, at the lower end of the scale, the lower difficulty thresholds (between a score of 0 and 1) for the easiest items were some way below what is shown on the Wright map. This means that were *LexCombi 2* used with learners of higher or lower proficiency than the participants in this study, it should still be possible to obtain reasonable measures of their ability to produce collocations.

Table 9.3 presents several statistics which give indications of the extent to which the data matched the Rasch model. The Rasch model expects a certain degree of variability in data, and thus while the model predicts that a participant will be more likely to achieve a good score on an easier item than on a more difficult item, it is expected that this will not always be the case. The fit statistics indicate the extent of the variability in the data for individual persons and items. That is, the fit statistics show to what extent an individual participant "fits" with the participants as a whole and to what extent an individual item "fits" with the items as a whole. As can be

Table 9.3: Fit and dimensionality.

Fit	
Number of underfitting persons (Infit mean squares > 1.3)	12
Number of overfitting persons (Infit mean squares < 0.75)	17
Number of underfitting items (Infit mean squares > 1.3)	0
Number of overfitting items (Infit mean squares < 0.75)	0
Point-measure correlations	
Number of items with a negative point-measure correlation	0
Number of other items with a low point-measure correlation (< .2)	3
Dimensionality	
Number of possible additional dimensions	1

seen, there were a number of misfitting persons among the 146 participants. Possible causes of misfit were discussed in Section 7.5, as was the fact that recommendations for fit criteria vary somewhat among Rasch scholars, which can lead to quite different assessments of the number of misfitting items and persons. The criteria used here are relatively conservative and may overestimate the number of misfitting persons somewhat. Importantly, however, despite the relatively conservative criteria, there were no misfitting items. This is perhaps unsurprising given that in the cue trialling process reported in Chapter 8 all the cues selected for use in *LexCombi 2* showed good fit. However, the data in the current study serve as confirmation of this.

Also in Table 9.3, the point-measure correlations indicate the correlation between the score on an item and scores on the instrument as a whole (item-total correlations in traditional test statistics). No items had a negative point-measure correlation or a very low (< .10) correlation, though there were three items with a point-measure correlation below .20.

Finally, the table shows the number of additional dimensions in the data. The Rasch model constructs a unidimensional scale from the data. That is, the assumption is that a single trait is being measured (in this case, that single trait is presumed to be the ability to produce collocations). In reality, data is never perfectly unidimensional and the fit and point-measure correlations provide some indication of problems. Dimensionality statistics address the issue more directly using factor analysis. This analysis indicated the possible presence of one further dimension in the data. However, Linacre (2014) states that when further dimensions are found, the content of the items involved (which the Winsteps program identifies) should be examined. If no pattern in the items' content can be found, the additional dimension may simply be the result of the randomness found in any real data. In this case, the additional dimension involved the items CAR, SCIENCE, MILE and PRODUCT. There was, then, a mixture of concrete and abstract words, loanwords in Japanese and non-loanwords and words of extremely high frequency and slightly lower frequency (albeit still frequent). Furthermore, it is hard to see anything that distinguishes these words from the other *LexCombi 2* cues. Thus, it was considered that there were likely no substantial issues of dimensionality.

Table 9.4 shows the number of items which had disordered categories: that is, cues for which the expectation that the four possible scores that can be achieved (i.e.

Table 9.4: Number of items with disordered categories.

	Number of items
Number of items with disordered Rasch-Andrich thresholds	2
Number of items with disordered ability levels of persons in each category	9

0, 1, 2 and 3) are progressively more difficult to achieve was not fulfilled. The first indicator, Rasch-Andrich thresholds, is about the ordering of the difficulty thresholds for achieving each score for an item. This was explained more fully in Section 8.4.1. Two items displayed disordered Rasch-Andrich thresholds: NAME and PLAYER. In both cases, this was due to a relatively small number of participants achieving a score of 2, given the numbers that achieved scores of 0, 1 and 3. For these two items, it seems that it was in some way easier to give three canonical responses than it was to give two. Section 8.4.1 suggested that such disordered thresholds may be caused by the chaining of responses, and this may have been the case with these cues also.

The other indicator of disordered items is based on the average ability level (i.e. ability on *LexCombi 2*) of participants who achieved each score for an item. That is, it may be expected that the average ability of participants who achieved a score of three on an item, for example, was higher than that of participants who achieved a score of two on that item. For nine items, the ability level of persons who achieved each score were not perfectly ordered.

Rasch analysis provides separate figures for the reliability of the person measures and item measures. As Table 9.5 shows, the item reliability figure was very high, but the person reliability figure was somewhat lower. This means we can have some confidence in the reproduceability of the person measures and a high degree of confidence in the reproduceability of the item measures.

Table 9.5: Reliability of the person and item measures.

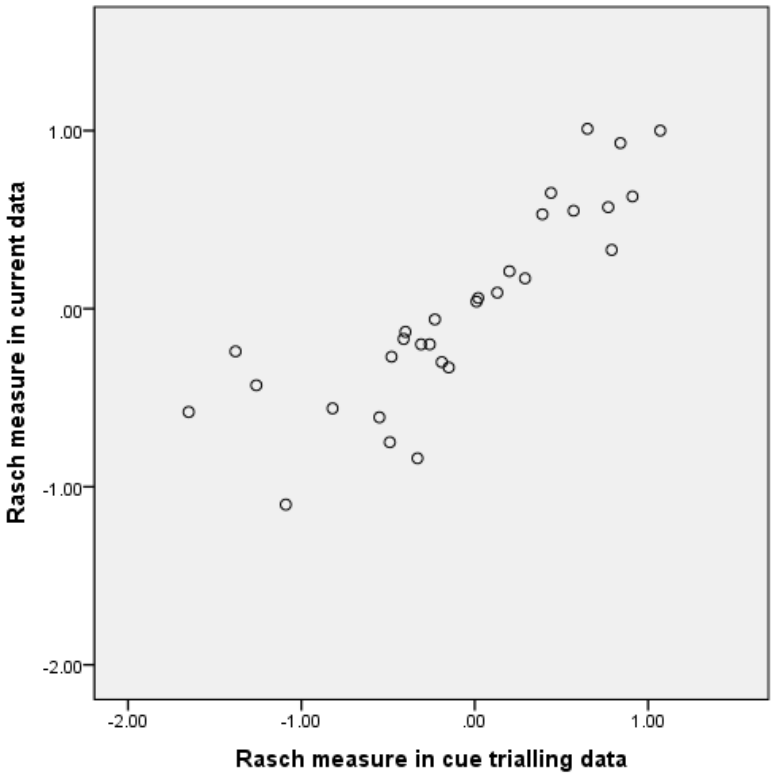
	LexCombi 2
Person reliability	.71
Item reliability	.96

9.3.3 Current item difficulties and previous item difficulties compared

The cue trialling data reported in Chapter 8 provided an estimate of the difficulty of each cue and this section examines the extent to which those estimates were matched in the present data set.

First, a correlation (Pearson's) of .85, $p < .001$, was found between the Rasch item difficulties for the 30 items in the current data set and their item difficulties in the cue trialling data set. As the scatterplot (Figure 9.3) shows, the correlation was very strong in the middle of the distribution, where the accuracy of the measures was greatest because there were more points of measurement in this part of the scale, but somewhat looser at either extreme. Nevertheless, it seems reasonable to suggest that overall the items behaved similarly in each set of data.

Figure 9.3: Scatterplot of the Rasch item difficulties in two data sets.



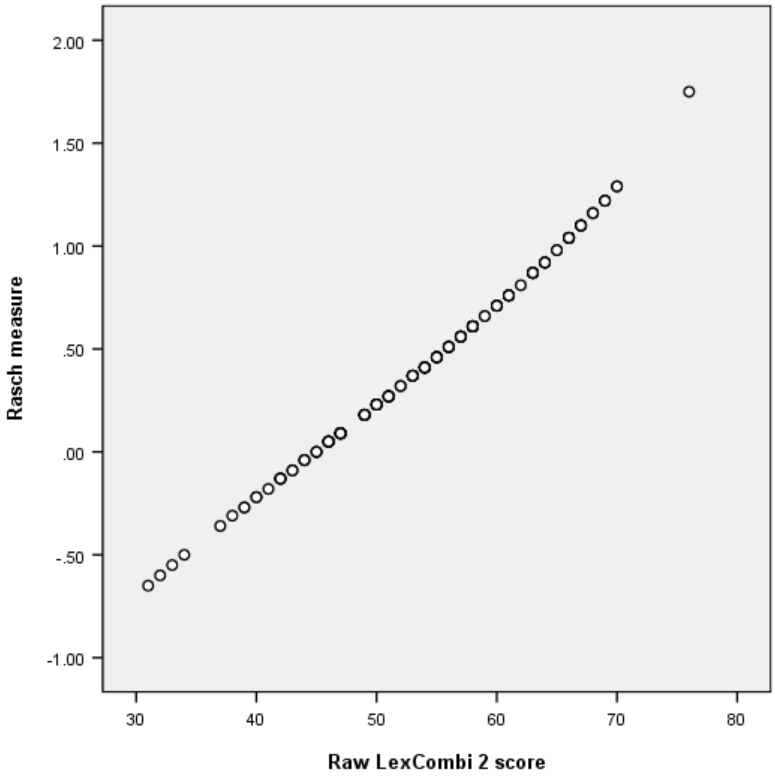
Looking at the individual cues, as the scatterplot makes clear, the majority were placed at a very similar point along the scale of difficulty in each data set. Nevertheless, some cues did appear at somewhat different points along the scale, and this seemed to be particularly the case for some of the less difficult cues. Linacre

(personal communication) suggests that changes in item difficulty of more than one logit are cause for concern, but smaller changes are due to natural variation in data sets. The largest change in the present data was 0.85 logits, exhibited by two cues (WATER and NAME), each of which had a higher difficulty measure in the current data (though still of low difficulty in comparison with the cues overall).

9.3.4 Raw *LexCombi 2* scores as an interval scale

The Rasch model constructs an interval scale from raw data. To assess the extent to which the current set of *LexCombi 2* scores can itself be treated as an interval scale, the correlation between the raw scores for each participant and the Rasch person measures was calculated. This correlation (Pearson's) was extremely strong at .998, $p < .001$. The scatterplot for the two variables is shown in Figure 9.4. This means that this particular set of raw *LexCombi 2* scores can be treated as interval data.

Figure 9.4: Scatterplot of the participants' raw scores and Rasch person measures.



9.4 Discussion and conclusions

This chapter sought to evaluate whether the changes made to *LexCombi* in previous chapters have resulted in a useful instrument for the elicitation of learners' productive knowledge of collocations. Three specific questions were asked. The first

was about the psychometric quality of the data *LexCombi 2* provides. This was assessed through a Rasch analysis, with four areas of interest examined.

First, it was found that the current set of *LexCombi 2* cues are able to provide information on the productive collocation knowledge of learners of a wide range of abilities. This should allow *LexCombi 2* to be used with learners of greater ability than the learners in the sample considered here, and with learners of lesser ability (albeit that *LexCombi 2* would not be appropriately used with learners of such low proficiency that the cue words are entirely unknown). The ability to use a single instrument with a single set of cues with learners with quite considerable differences in ability has clear advantages over approaches, common among vocabulary instruments, in which learners at different levels of ability are targeted with different tests or sub-sets of items, in that directly comparable scores may be produced without the need for statistical interventions.

Next, the dimensionality of *LexCombi 2* was considered. It was seen that there were no misfitting items, while two items had low, but not very low, point-measure correlations. It was, though, found that there may be one further dimension in the data. Consideration of the specific items involved, however, as recommended by Linacre (2014), could identify no characteristics that could account for this. At this stage then, it will be assumed that this finding stems from normal degrees of randomness found in data. *LexCombi 2* does then appear to be a measure of a single construct.

Thirdly, the results showed that some items displayed disordered scoring categories. One means of judging disorder, looking at the ability levels of participants achieving different scores, is rather broad, and it was not altogether surprising that a number of items were found to be disordered by this method. Under the other method, Rasch-Andrich thresholds, two cues were found to be disordered, and it was suggested that the chaining of responses may be responsible. In the cue trialling process in contrast (reported in Chapter 8), these two cues were not found to be disordered under this method. At present this is seen as a potentially problematic issue, and the data on these items should be carefully examined under future uses of *LexCombi 2*. It is not, however, seen as necessary to take immediate action with regard to these items.

Finally, the reliability figures indicated that, for this particular sample of

participants on this administration of *LexCombi 2*, the reproduceability of the item measures is estimated to be high, but is more moderate for the person measures. This was likely in part due to the structure of the data set, as there were 146 observations for each item compared with 30 observations for each person. Another contributing factor was the somewhat limited spread of scores in this data set with few participants achieving low *LexCombi 2* scores, this limited spread also being seen in the Rasch measures in the relatively small person standard deviation of 0.44 logits.

A higher person reliability figure could potentially be achieved by increasing the number of items in *LexCombi 2*. However, one advantage of *LexCombi 2* in its current form is its brevity, and thus there is a balance to be struck between seeking a higher level of reliability and having an instrument which is practical. In addition, Linacre (2014) recommends a minimum reliability of .50 for meaningful measurement, and .80 for serious decision-making. Person reliability of .71 is arguably reasonable insofar as *LexCombi 2* is not intended for use as a high-stakes test, but rather for use as a research instrument.

The second question asked in this chapter was how the item difficulties in the current data set compared with those in the cue trialling data examined in Chapter 8. One indication of this was provided by the Rasch item reliability mentioned above, a figure of .96 indicating that the item difficulties should be replicable with a similar group of participants. In addition, the data set from the cue trialling process provided an opportunity to see in practice how the items performed with two different groups of participants from the same broad population. The results indicated that overall the cues did indeed perform similarly in the two data sets. It was seen, however, that some cues appeared at somewhat different points along the scale of measurement in the two data sets. This variation was within normal bounds and was likely the result of natural variation, though with future uses of *LexCombi 2* it is advisable to look again at this issue and in particular at the two items that differed most in the two data sets.

The final issue of interest was whether the raw *LexCombi 2* scores can be treated as interval data. An extremely strong correlation was found between the raw scores for participants and the Rasch measures, showing that, in the case of this particular set of *LexCombi 2* scores, the data can be treated as interval data. It should be noted, however, that this finding is limited to the current data set. The very strong

correlation merely indicates that the raw scores in this particular data set may be treated as interval data; it does not show that raw scores data produced by *LexCombi 2* can be assumed to be interval-level data since raw scores are non-linear (Bond & Fox, 2007; McNamara, 1996).

To what extent, then, could a researcher interested in the collocation knowledge of English learners rely on *LexCombi 2* for data collection? The wider issue of *LexCombi 2*'s validity as a research instrument will be discussed in Section 11.2, but given the conclusions above, it seems that he or she could do so with regard to the reliability of the items, and the spread of the abilities that can be tested. On the other hand, caution would be needed with regard to person reliability and the performance of some particular items. Furthermore, it must be noted that *LexCombi 2* was developed with Japanese learners of English, with the scoring in part based on norms from Japanese users of English, and thus might not be as sensitive to the performance of other learner groups due to factors such as false friends, cognates and local cultural associations in the non-Japanese context. Overall, however, the data provided by *LexCombi 2* does seem of sufficient quality for its intended use as an elicitation instrument for research purposes.

Furthermore, the very strong correlation between the raw scores and the Rasch measures means that, in the case of this particular data set, the raw scores data may be treated as interval-level data, making interpretation of the scores far simpler. Consequently, in Chapter 10, in which some initial explorations of this *LexCombi 2* data will be undertaken, the raw scores rather than the Rasch measures will be used.

Chapter 10 Exploring *LexCombi 2* data

10.1 Introduction

After the revisions made to *LexCombi* in terms of its format (Chapter 6), scoring (Chapter 7) and cues (Chapter 8), Chapter 9 provided an evaluation of the final form of the instrument, *LexCombi 2*. It was concluded that *LexCombi 2* provides data of good quality, making it suitable for its intended use as an elicitation instrument for research purposes. This chapter considers, after all of this development and evaluation, what *LexCombi 2* data can actually tell us about what learners know with regard to collocations. The chapter describes initial explorations of four areas.

The relationship between LexCombi *scores and proficiency*. That there should be some kind of relationship between the ability to produce collocations and language proficiency in general seems intuitively plausible. Yet previous studies of learners' productive knowledge of collocations have found varied results in this regard (see Section 3.2.3). With respect to *LexCombi* specifically, Barfield (2009a) found a moderate correlation between *LexCombi 1* scores and a measure of proficiency, while a mixed picture is provided by the data collected with *LexCombi* reported in this thesis. In Section 10.3, an explanation for these contrasting findings will be sought.

The range and frequency level of participants' responses to LexCombi 2. The range (i.e. the number of different words used) and the frequency level of *LexCombi 2* responses potentially offer insights into the differences between learners in terms of their *LexCombi 2* scores. Did participants who achieved higher scores tend to use a wider range of lexical items as responses? Did they make use of words of lower frequency as responses? Alternatively, did they achieve their scores using largely the same pool of lexical items as participants who achieved lower scores only deploying those items more successfully? Resolving such questions would have implications for our understanding of the developing L2 lexicon. That is, the first two possibilities above would suggest that overall lexical development is more important to the development of collocation knowledge; the third possibility that the quality or depth of learners' knowledge of words is of greater importance. Section 10.4 describes an investigation of these issues.

The word class of participants' responses to LexCombi 2. Previous studies of L2 learners' knowledge of collocations have mostly focused on a particular type, such as verb + noun collocations (e.g. Eyckmans, 2009; Nesselhauf, 2005; Revier, 2009). *LexCombi* has the advantage of allowing the elicitation of collocates of any word class. Consequently, as attempted by Barfield in his initial work with *LexCombi 1*, examining the word class of the responses given by participants may make it possible to pinpoint the nature of learners' developing knowledge of collocations, in that there may be different rates of development for different types of collocates. Barfield's claims and the possibilities of this type of analysis were thus scrutinised, as Section 10.5 describes.

Differences between overall LexCombi 2 *scores and scores achieved under alternative lists of canonical responses*. *LexCombi 2* is scored by comparing learners' responses with a list of canonical collocates for each cue. The set of lists used presently (the ALC lists) were drawn from four sources: (1) dictionaries of collocations; (2) corpus searches; (3) L1-user norming; and (4) L2-user norming. Chapter 7 demonstrated that, in terms of their content, the four sets of source lists are quite different from one another, and detailed how they imply differing conceptions of collocation. Accordingly, Section 10.6 reports on an investigation into whether there were differences among participants in the scores achieved against each set of source lists, since any such differences could be seen as indicating divergent development for different types of collocation. That is, scoring better under one set of lists than under another would suggest relatively greater progress towards the characterisation of collocation provided by those lists.

10.2 Method

The analyses reported in this chapter at times made use of all the data sets assembled in the exploration and development of *LexCombi*, as reported in Chapters 4-9. The primary focus, however, was on the Chapter 9 data, since this was collected with a revised and improved version, *LexCombi 2*. As Section 9.2 explained, this data set consisted of responses from 146 Japanese university students to two instruments: (1) *LexCombi 2*; and (2) a Yes/No vocabulary test. Section 9.3.4 showed that the Rasch measures for each participant and their raw *LexCombi 2* scores correlated very strongly, and thus this particular set of raw scores can be treated as interval-level

data. Consequently, the analyses reported in this chapter made use of the raw scores, since they have the advantage of being more easily interpretable.

Many of the analyses made use of the full set of data from 146 participants. However, it was also expedient at times to look at the data through the lens of certain groups of participants. As detailed below, two groupings were made, based on collocation knowledge and proficiency respectively, to provide insights into the data from different perspectives. In both cases, groups were formed containing 25 participants each, a number intended to ensure that individual variation would not have an undue influence, while also allowing the establishment of groups that were clearly distinguished from one another.

Collocation Knowledge Groups: Three groups of participants were formed on the basis of their *LexCombi 2* scores: low, mid and high *LexCombi 2* scorers. These groups consisted of the 25 lowest-scoring participants, 25 participants clustered around the mean *LexCombi 2* score for the entire sample of participants and the 25 highest-scoring participants. When there were multiple participants at the cut-off point for the formation of a group, random selection was performed. Table 10.1 shows the descriptive statistics for the three groups, along with the overall figures for reference. The three groups were sharply distinguished in terms of their *LexCombi 2* scores, and each was statistically different from the others, as seen through a Kruskal-Wallis test ($H(2) = 66.011$, $p < .001$) and post hoc Mann-Whitney tests (non-parametric tests used due to issues with normality and homogeneity of variances). Also notable is that there were no significant differences between these groups in their Yes/No scores: ($F(2, 72) = 0.704$, $p = .50$): Low *LexCombi 2* scorers ($N = 25$), $M = 5{,}262$ ($SD = 992$); Mid *LexCombi 2* scorers ($N = 25$), $M = 5{,}066$ ($SD = 1{,}055$); High *LexCombi 2* scorers ($N = 25$), $M = 5{,}388$ ($SD = 841$).

Proficiency-based groups: Three groups of participants were formed on the basis of their Yes/No vocabulary test scores: low, mid and high proficiency groups. These groups contained the 25 lowest-scoring participants, 25 clustered around the mean Yes/No score for the entire sample and the 25 highest-scoring participants. Random selection was performed when more than one participant was at the cut-off point for the formation of a group. Table 10.2 gives the descriptive statistics for the groups. Each group was statistically different from the others in terms of their Yes/No scores, as seen through a Kruskal-Wallis test ($H(2) = 65.869$, $p < .001$) and post

Table 10.1: Descriptive statistics for the *LexCombi 2* scores of low, mid and high *LexCombi 2* scorers.

	Mean	*SD*	Minimum	Maximum
Low *LexCombi 2* scorers (*N* = 25)	40.68	4.394	31	46
Mid *LexCombi 2* scorers (*N* = 25)	53.44	1.294	51	55
High *LexCombi 2* scorers (*N* = 25)	66.88	2.728	63	76
All participants (*N* = 146)	53.92	8.875	31	76

Note. Maximum score = 90.

Table 10.2: Descriptive statistics for the Yes/No scores of low, mid and high proficiency groups.

	Mean	*SD*	Minimum	Maximum
Low proficiency (*N* = 25)	3,770	389	3,050	4,350
Mid proficiency (*N* = 25)	5,220	103	5,050	5,400
High proficiency (*N* = 25)	6,510	480	6,000	7,800
All participants (*N* = 146)	5,230	915	3,050	7,800

Note. Maximum score = 10,000.

hoc Mann-Whitney tests (non-parametric tests used as there were issues with normality and homogeneity of variances). There were no significant differences between the groups in terms of their *LexCombi 2* scores: ($F(2, 72) = 0.472$, $p = .63$): Low proficiency ($N = 25$), $M = 52.84$ ($SD = 10.197$); Mid proficiency ($N = 25$), $M = 52.96$ ($SD = 9.270$); High proficiency ($N = 25$), $M = 55.08$ ($SD = 7.947$).

10.3 *LexCombi* scores and proficiency

The first area of interest concerned the relationship between *LexCombi* scores and proficiency. It might be anticipated that as learners develop in their overall proficiency in a language, their ability to produce canonical collocates should also develop. This could occur since the learner has more words available with which to respond and/or because the learner has a deeper knowledge of words and so greater knowledge of conventional forms of language (issues pursued in Section 10.4). There have, however, been mixed findings regarding the relationship between knowledge of collocations and general proficiency, with Barfield (2009a), González

Fernández and Schmitt (2015) and Revier (2009) finding just such a relationship, while Howarth (1998a) and Nesselhauf (2005) did not (see Section 3.2.3 for discussion). This section seeks to answer one question:

1. What is the relationship between *LexCombi* scores and general language proficiency?

10.3.1 *LexCombi*-proficiency correlations

As noted above, Barfield's *LexCombi 1* study found a moderate positive relationship between *LexCombi 1* scores and a measure of proficiency, while in the various data sets assembled for this thesis, there was a mixed picture. As shown in Table 10.3, Studies 2 and 4 in the present project found weak correlations, Study 3 a moderate correlation and Study 5 no correlation.

Table 10.3: Correlations between *LexCombi* scores and a proficiency measure in five data sets.

Study	*N*	Correlation (Pearson's)	Significance
1. Barfield (2009a)	89	.57	$p < .001$
2. Chapter 5 (full-phrase responses) data	35	.28	$p > .05$
3. Chapter 6 data	125	.51	$p < .001$
4. Chapters 7 and 8 data	223	.13	$p > .05$
5. Chapter 9 data (*LexCombi 2*)	146	.05	$p > .05$

What could account for these mixed results? There were a number of factors that may provide an explanation. As Table 10.4 shows, there were differing measures of proficiency across the studies, differences between the characteristics of the learners involved, and different versions of *LexCombi* were used. Each difference was examined and an assessment made as to whether it could explain the contrasting correlation results.

Beginning with the measures of proficiency employed, there was a clear difference between Barfield's study, which used TOEIC scores, and my studies, which all used vocabulary size as estimated by Yes/No tests as a proxy for proficiency. It may then be asked whether vocabulary size is a good proxy for

proficiency and whether these Yes/No tests are good measures of vocabulary size.

A number of instruments are available for estimating vocabulary size, and in most cases the test developers make some claim regarding vocabulary knowledge and aspects of proficiency. For example, Schmitt, Schmitt and Clapham (2001) report that the rationale for the *Vocabulary Levels Test* (Nation, 1983, 1990) "stems from research which has shown that vocabulary size is directly related to the ability to use English in various ways" (p. 55); Beglar (2010), regarding the *Vocabulary Size Test*, suggests that "because of the key role that lexical knowledge plays in reading and listening, it is important that estimates of receptive vocabulary size be available" (p. 101); and Meara and Fitzpatrick (2000), explaining the development of the *Lex30* test, state that "In most practical contexts it is clear that communicative effectiveness is achieved more successfully by learners with a larger vocabulary than by learners with a more detailed command of a smaller one" (p. 20).

Empirical evidence backs these claims: Meara and Buxton (1987) reported that a Yes/No test was able to successfully predict the results of learners on the *Cambridge First Certificate* examination; Meara and Jones (1988) found a strong correlation between a Yes/No test and a general proficiency test; Qian (2002) reported a strong correlation between the *Vocabulary Levels Test* and a TOEFL reading test; Meara and Milton (2003), cited by Milton (2009), found strong correlations between vocabulary size and scores in writing, reading comprehension and grammatical accuracy, and a moderate correlation with oral fluency; Stæhr (2008) reported strong correlations between the *Vocabulary Levels Test* and tests of listening, writing and especially reading; and Milton, Wade and Hopkins (2010) found strong correlations both between the *X_Lex* test and IELTS reading and writing scores and between an aural version of *X_Lex* and IELTS listening and speaking scores, with both correlating strongly with the overall IELTS scores. Reviewing much of the evidence cited above, Milton (2009) concluded that "it appears that it [vocabulary size] can be a quick and useful way to assess the overall level of knowledge and proficiency of a foreign language learner" (p. 191).

Support for the Yes/No format can also be found in a number of evaluations of it (Beeckmans, et al., 2001; Huibregtse, et al., 2002; Mochida & Harrington, 2006), with Mochida and Harrington finding that a Yes/No test produced similar results to the *Vocabulary Levels Test*, while having clear practical advantages. It does

Table 10.4: Differences between studies providing *LexCombi*-proficiency correlations.

	Learners	*LexCombi* format	*LexCombi* scoring	*LexCombi* cues	Proficiency measure
Barfield (2009a) *N* = 89, *r* = .57	"The students' L1 was Japanese, and they belonged to different first-, second- and third-year undergraduate Faculty of Law English classes, ranging in proficiency from low-intermediate to advanced" (p. 98); TOEIC scores ranged from 325-900 (Mean and *SD* not given)	Barfield's original format (cue presented along with three boxes to the right for responses)	canonical collocates lists based on the *Oxford Collocations Dictionary (1st edition)* and collocates identified using *Collins Wordbanks Online*	Barfield's original 30 cues	TOEIC scores
Study 2 *N* = 35, *r* = .28	Japanese L1 university students, not majoring in English; volunteers from two classes; placed in these classes by an internal placement test; *X_Lex* scores from 2,950-4,500 (*M* = 3,876, *SD* = 373)	Barfield's original format	canonical collocates lists based on the *Oxford Collocations Dictionary (2nd edition)*	Barfield's original 30 cues	vocabulary size scores based on an *X_Lex* Yes/No vocabulary test
Study 3 *N* = 125, *r* = .51	Japanese L1 university students; volunteers from six classes of different faculties, one class being English majors; *X_Lex* scores from 2,400-4,950 (*M* = 3,836, *SD* = 479)	adapted format (cue presented within three boxes with space on either side of the cue for a response)	canonical collocates lists based on the *Oxford Collocations Dictionary (2nd edition)*	Barfield's original 30 cues	vocabulary size scores based on an *X_Lex* Yes/No vocabulary test
Study 4 *N* = 223, *r* = .13	Japanese L1 university students from two universities and a range of faculties, including some English majors; combined *X_Lex/Y_Lex* scores from 2,950-7,850 (*M* = 5,438, *SD* = 847)	adapted format	canonical collocates lists based on combining four source lists: dictionaries of collocations, corpus searches, L1-user norms, L2-user norms	a total of 70 cues which were under trial, though individual participants responded to one of three sets of 30 cues	vocabulary size scores based on a combined *X_Lex/Y_Lex* Yes/No vocabulary test
Study 5 *N* = 146, *r* = .05	Japanese L1 university students from four universities and a range of faculties, including some English majors; combined *X_Lex/Y_Lex* scores from 3,050-7,800 (*M* = 5,230, *SD* = 915)	adapted format	canonical collocates lists based on combining four source lists, as above	a new selection of 30 cues from the 70 trialled in Study 4	vocabulary size scores based on a combined *X_Lex/Y_Lex* Yes/No vocabulary test

therefore seem reasonable to have used vocabulary size as a proxy for proficiency and to have employed Yes/No tests in particular to estimate vocabulary size.

Furthermore, for Study 5 there is additional corroborating evidence for the weak correlation between the Yes/No scores as a measure of proficiency and the *LexCombi 2* scores. The participants in Study 5 were asked for any language proficiency test results obtained within the previous 12 months, and 29 participants provided a TOEIC score and 35 a TOEFL score. The correlations between these scores and the *LexCombi 2* scores were also very weak: .14, $p > .05$, for the TOEIC scores and -.02, $p > .05$, for the TOEFL scores. In Study 5, then, proficiency as judged in three different ways (Yes/No scores, TOEIC scores and TOEFL scores) had only a weak correlation with *LexCombi 2* scores. It was concluded therefore that differences in the measures of proficiency between Barfield's study and the studies in the current project were likely not the primary cause of the contrasts seen in the correlations between *LexCombi* scores and proficiency.

Regarding the learners, the chief difference between the studies was that some featured learners majoring in subjects other than English, while others featured both learners majoring in other subjects and learners majoring in English. Broadly, however, the participants in the five studies were similar: they were all undergraduate university students in Japan, with Japanese as their L1. Four out of the five studies also had learners from a wide range of proficiency levels. The one exception was Study 2, in which participants were drawn from two classes of a similar proficiency level as judged by their university's internal placement test, and thus were from a narrower proficiency range. Study 2's weak correlation may therefore have been due to a lack of variance in the proficiency of the participants. Excluding Study 2, however, there was no clear difference between the participants in the studies which showed a moderate correlation (Barfield's and Study 3) and the participants in the studies that found a weak or no correlation (Studies 4 and 5). It seems unlikely that the learners were the primary source of the contrast in correlations.

With respect to the *LexCombi* format, there was a clear difference between the first two studies in Table 10.4, which used Barfield's original format, and the latter three, which used the adapted *LexCombi* format. However, this difference did not match up with the correlation findings: that is, the studies finding a moderate

correlation and those finding a weaker correlation. This factor then was also discounted.

Looking at the *LexCombi* scoring, the main difference was between the first three studies in Table 10.4, which used canonical collocates lists based primarily on the *Oxford Collocations Dictionary*, and the latter two which used lists based on multiple sources. Furthermore, this difference tallied with the contrast in correlations (after discounting Study 2). To eliminate this difference between the studies, the scoring program, CollCheck (Imao & Brown, 2014), was used to re-score the *LexCombi* responses in Studies 4 and 5 using the OCD2 lists (i.e. the same lists as used in Studies 2 and 3). Calculated anew, the correlations between *LexCombi* scores under the OCD2 lists and the Yes/No scores were .28 for Study 4 and .16 for Study 5. In both cases a stronger correlation was therefore found. However, the correlations remained weak and the contrast between these two studies and Barfield's study and Study 3 persisted. Differences in *LexCombi* scoring were also likely not the main factor.

Finally, a key difference between the first three studies and the latter two was the *LexCombi* cues. Discounting Study 2 due to the narrow proficiency range among its participants, there were two studies (Barfield's and Study 3) which used Barfield's original set of cues and that showed a moderate correlation between *LexCombi* scores and proficiency, and two (Studies 4 and 5) which used different cues and that showed a weak or indeed no correlation. Thus, the differences between the studies and the contrast in the correlations coincided.

The explanation for this apparent effect of the cues may lie in the procedures used by Barfield for selecting cues. As reported in Section 8.2, in selecting cues, Barfield piloted 50 frequent nouns with 35 British English L1 users and 35 highly proficient Japanese users of English and then chose 30 cues that "differentiated well between the responses" (p. 97) of the two groups. It seems plausible to suggest that cues which differentiated these two groups might also differentiate learners at different proficiency levels. In other words, there was some circularity in selecting cues on the basis of their ability to discriminate between respondents, using those cues to calculate scores and then finding a correlation between those scores and proficiency. In contrast, the cue selection procedure which produced the 70 cues trialled in Study 4, and then the final selection of 30 cues for *LexCombi 2* used in

Study 5 (see Chapter 8) did not make use of any similar process. The current *LexCombi 2* cues are then free from any bias towards cues that distinguish between participants of differing proficiency levels.

10.3.2 Discussion

The preceding paragraphs have reviewed five studies of the correlation between *LexCombi* scores and proficiency. It has been seen that differences in the proficiency measures used in these studies, between the learners involved in the studies, between the *LexCombi* format used and between the lists of canonical responses employed do not seem to account for the contrasting correlations found in the studies. The key difference between the two groups of studies appears, instead, to be the *LexCombi* cues. In particular, it may be that the cue selection procedure used by Barfield was biased towards selecting cues that differentiated learners of different proficiencies, thus explaining the correlations between *LexCombi* scores and proficiency in the studies that used those cues. Consequently, the very small correlations found between *LexCombi* scores and proficiency in Studies 4 and 5 may more accurately reflect the relationship between these variables. It can, then, be concluded that for the learners concerned here, Japanese university students between low intermediate and high intermediate levels of proficiency, there was no discernible relationship between their general proficiency in English and their ability to produce collocations.

It seems, therefore, that whatever increases proficiency is not the same thing that develops collocation knowledge. It should be emphasised, however, that this finding is limited at present to the type of learners involved in the study. Is there any particular characteristic of these learners that could account for this finding? It is possible that the lack of correlation is a result of the education system in Japan in which English is mostly studied for the purposes of high-stakes examinations, which are overwhelmingly focused on receptive skills. This leads learners to focus on building receptive skills with little attention given to productive skills. Given this background, a lack of correlation between a receptive test (the Yes/No test) and a productive test (*LexCombi*) may not be surprising. The apparent independence of the two measures is, nonetheless, intriguing and is explored further in Chapter 11.

10.4 Range and frequency level of participants' responses

The range of words a learner uses in responding to *LexCombi 2* and the frequency

level of those words potentially give useful information on the developing lexicon in that both may be indicators of the breadth of vocabulary a learner is able to deploy. These two characteristics were explored with a view to considering what lay behind differences in participants' performance on *LexCombi 2*. As seen in Section 10.3, there was no relationship in the Chapter 9 data between general proficiency as judged by Yes/No scores and *LexCombi 2* performance. Consequently, the analyses this section reports focused on the three collocation knowledge groups (see Section 10.2) (i.e. the low, mid and high *LexCombi 2* scorers), and sought a better understanding of what enabled higher *LexCombi 2* scores. Higher scores may have been achieved by deploying a wider range of words in giving responses and/or by having knowledge of collocates which are words of lower frequency. However, the high scorers may also have achieved their scores simply by making better use of approximately the same pool of words as used by the low scorers (i.e. using the same overall pool of words but giving different responses to particular items). The former would suggest that overall lexical development drives the development of collocation knowledge; the latter that it is the quality or depth of the learners' knowledge of words that is of importance.

This section deals with two questions:

1. Were there differences in the range of words used as responses by participants?
2. How many responses of different word frequency levels were given by participants?

10.4.1 Range of words used as responses

On the first question, Table 10.5 gives the number of responses (response tokens), the number of different words used as responses (response types) and the ratio between these two (type-token ratio) for the three collocation knowledge groups. A Kruskal-Wallis test, used since the data violated the assumption of homogeneity of variances and included non-normal distributions, found a significant difference between the groups for response tokens ($H(2) = 27.213$, $p < .001$) and post hoc Mann-Whitney tests showed each group to be significantly different from the others. There was also a significant difference between the groups for response types ($F(2, 72) = 13.392$, $p < .001$), though in this case post hoc Tukey tests showed the high

Table 10.5: Response tokens, response types and type-token ratio for the low, mid and high *LexCombi 2* scorers.

		Mean	*SD*
Response tokens	Low *LexCombi 2* scorers (*N* = 25)	63.40	14.874
	Mid *LexCombi 2* scorers (*N* = 25)	74.36	7.905
	High *LexCombi 2* scorers (*N* = 25)	84.44	4.565
	All participants (*N* = 146)	75.12	11.967
Response types	Low *LexCombi 2* scorers (*N* = 25)	53.60	11.864
	Mid *LexCombi 2* scorers (*N* = 25)	55.60	9.434
	High *LexCombi 2* scorers (*N* = 25)	66.88	7.574
	All participants (*N* = 146)	60.10	10.964
Type-token ratio	Low *LexCombi 2* scorers (*N* = 25)	.85	.087
	Mid *LexCombi 2* scorers (*N* = 25)	.75	.111
	High *LexCombi 2* scorers (*N* = 25)	.79	.083
	All participants (*N* = 146)	.80	.101

scorers to be significantly different from the other two groups, but no difference between the low and mid scorers. For the type-token ratio, meanwhile, there was a significant difference between the groups ($H(2) = 10.598$, $p < .01$), but post hoc tests showed the low scorers to be significantly different from both the mid and high scorers (non-parametric tests used since these data are ratios).

In interpreting these results, it should first be noted that type-token ratios are known to be affected by text length. A number of attempts have been made to provide measures of lexical diversity unaffected by text length (Malvern, Richards, Chipere, & Durán, 2004; Meara & Bell, 2001), but these measures were not used here as a set of *LexCombi 2* responses do not constitute a text in any normal sense of the word. Nevertheless, the decrease in the type-token ratio across the three groups may well have resulted from the increase in the number of response tokens.

The data, then, showed that participants achieving higher *LexCombi 2* scores gave more responses (both ones that were and were not canonical), and to some extent used a wider range of words as responses. Certainly, the high scorers produced both more response tokens and more types. What was less clear was

whether giving more response tokens led to the production of more types (i.e. participants carried on looking for words to respond with and that led them to come up with new ones) or vice versa (i.e. because they came up with more new words, the net result was a higher total number of responses). The results for the mid scorers were intriguing in this regard: they gave significantly more response tokens than the low scorers, but used a similar number of types (while, as noted earlier, the Yes/No test indicated that there was no difference between these two groups in terms of passive recognition of vocabulary). Giving more response tokens did not then necessarily result in more types being used, while these findings also suggest that the ability to produce varied types was not necessary in order to produce more response tokens.

10.4.2 Frequency of words used as responses

The frequency level of the words used as responses may be an indication of the breadth of vocabulary a learner is able to make use of. The frequency level of each response was checked in the JACET8000 word list (Ishikawa, et al., 2003). As mentioned previously, this is a lemma-based word list primarily based on the BNC, though the ranking of words in the list in part reflects English usage in Japan. Each response was classified as either JACET 1 (within the first 1,000 words in the list), JACET 2 (the second 1,000) or Beyond JACET 2. Table 10.6 gives the figures.

Since there were issues with normality and homogeneity of variances, it was not possible to perform a MANOVA to identify differences between the three groups. Consequently, separate analyses were conducted at each frequency level. Kruskal-Wallis tests were used to compare the three groups, with a Bonferroni correction

Table 10.6: Number of responses at different JACET8000 levels.

	JACET 1		JACET 2		Beyond JACET 2	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Low *LexCombi 2* scorers (*N* = 25)	46.80	11.776	7.64	3.988	8.96	4.979
Mid *LexCombi 2* scorers (*N* = 25)	59.52	10.215	7.84	3.682	7.00	3.731
High *LexCombi 2* scorers (*N* = 25)	67.32	6.053	8.84	3.091	8.28	3.348

applied, setting the significance level at .0167 (i.e. .05/3). At JACET 1, a significant difference was found ($H(2) = 33.129$, $p < .0167$) and post hoc Mann-Whitney tests with a Bonferroni correction (significance level = .0167 (i.e. .05/3)) showed significant differences between all three groups. For JACET 2, there was no significant difference between the groups ($H(2) = 3.373$, $p = .185$), and for Beyond JACET 2 likewise ($H(2) = 2.962$, $p = .227$). It was then the ability to give more JACET 1 responses that distinguished the groups.

Examining instead only the canonical responses, in order to see if there were differences from the above, broadly similar results were found (Table 10.7). Again, a MANOVA could not be performed due to problems with normality. Hence, separate analyses were conducted at each frequency level using one-way ANOVAs, with a Bonferroni correction setting the significance level at .0167 (i.e. .05/3). At JACET 1, a significant difference was found ($F(2, 72) = 159.007$, $p < .0167$), and post hoc Tukey tests showed a difference between all three groups. For JACET 2, the result was also significant ($F(2, 72) = 4.576$, $p < .0167$), and post hoc tests showed a difference between the low scorers and the high scorers. For Beyond JACET 2, there was no significant difference between the groups: $F(2, 72) = 2.650$, $p = .078$. Thus, with respect to canonical responses, the use of JACET 1 responses distinguished all three groups from each other, with the low scorers and high scorers further distinguished by JACET 2 responses.

The differences between the number of responses and the number of canonical responses at different frequency levels were brought into sharper focus by examining the percentage of responses that were canonical at each level of frequency. This gave

Table 10.7: Number of canonical responses at different JACET8000 levels.

	JACET 1		JACET 2		Beyond JACET 2	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Low *LexCombi 2* scorers ($N = 25$)	32.24	4.693	4.32	2.462	4.12	2.128
Mid *LexCombi 2* scorers ($N = 25$)	44.00	4.062	5.40	2.754	4.04	2.263
High *LexCombi 2* scorers ($N = 25$)	54.92	4.707	6.56	2.631	5.40	2.614

an insight into the control participants had over words at different frequency levels. That is, if all the responses provided by a participant at a given frequency level were canonical, it would seem that the participant had a high degree of knowledge of the uses of words at that frequency level. If, however, only 50% were canonical, it would imply that there was a lesser degree of knowledge of the uses of words at that level. Table 10.8 and Figure 10.1 show the results. Since all the figures were percentages, non-parametric tests were used throughout this analysis. For the overall figures, six statistical tests were conducted (i.e. comparisons between JACET 1, JACET 2 and Beyond JACET 2 within the three groups, as well as comparisons across the three groups at each frequency level), so a Bonferroni correction set the significance level at .0083 (i.e. .05/6 = .0083).

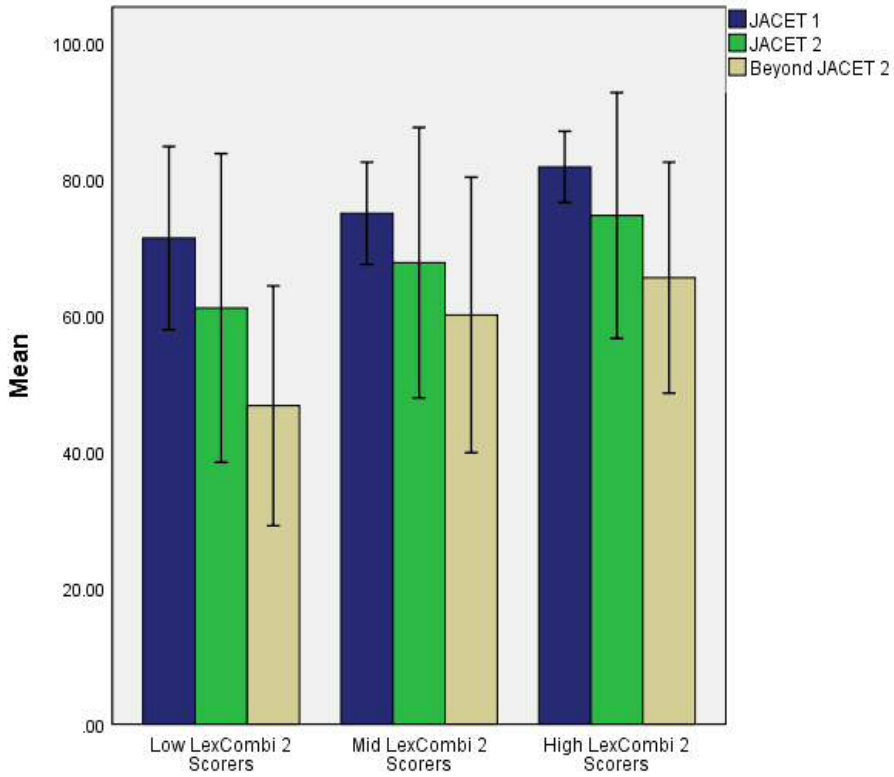
First, looking within each group, a Friedman's ANOVA found a significant difference between the percentage of responses which were canonical at JACET 1, JACET 2 and Beyond JACET 2 for all three groups: low *LexCombi 2* scorers, $\chi^2(2) = 23.592$, $p < .0083$; mid *LexCombi 2* scorers, $\chi^2(2) = 11.120$, $p < .0083$; high *LexCombi 2* scorers, $\chi^2(2) = 13.273$, $p < .0083$.

Then, comparing the groups across each frequency level, for JACET 1, a Kruskal-Wallis test found a significant difference between the groups ($H(2) = 11.471$, $p < .0083$) and post hoc Mann-Whitney tests with a Bonferroni correction applied (significance level = .0167 (i.e. .05/3)) showed a significant difference between the high scorers and both the low and mid scorers, but no difference between these latter two. For JACET 2, there was no significant difference between the groups: $H(2) = 4.869$, $p = .088$. Finally, for Beyond JACET 2, the result was

Table 10.8: Percentage of canonical responses at different JACET8000 levels.

	JACET 1		JACET 2		Beyond JACET 2	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Low *LexCombi 2* scorers ($N = 25$)	71.34	13.448	61.09	22.651	46.74	17.601
Mid *LexCombi 2* scorers ($N = 25$)	74.94	7.501	67.73	19.848	60.06	20.218
High *LexCombi 2* scorers ($N = 25$)	81.78	5.232	74.65	18.033	65.54	16.939

Figure 10.1: Percentage of canonical responses at different JACET8000 levels.



Note. Error bars represent 1*SD*.

significant ($H(2) = 12.083$, $p < .0083$) and post hoc tests showed a significant difference between the high scorers and the low scorers.

10.4.3 Discussion

This section has explored differences between the three collocation knowledge groups on two characteristics of the words used as responses by participants. The intention in looking at these characteristics was to seek insights into the developing lexicon of learners, in particular into whether the development of collocation knowledge is related to overall lexical development or whether the quality, depth or organisation of their lexical knowledge is the key.

It should be remembered that the three collocation knowledge groups were not distinguished in terms of general English proficiency, as judged by their Yes/No scores. The Yes/No test is, however, a test of passive recognition of vocabulary, so there is a possibility that the groups differ in terms of their productive vocabulary (e.g. in their ability to retrieve items on demand) and this could explain some of the results discussed below.

The first characteristic explored was the range of words used as responses. It was seen that participants achieving higher *LexCombi 2* scores gave more response tokens and made use of more types in their responses. Most intriguing, however, was the finding that the mid scorers used a similar number of types to the low scorers, and yet gave significantly more responses overall (as well as, by definition, achieving significantly higher *LexCombi 2* scores). These results suggest that learners' progressive capacity to produce collocations derives from better knowledge of the uses of words or an improved sense of the connections between words, rather than from simply having a broader productive vocabulary.

The second characteristic investigated was the number of responses and the number of canonical responses at different levels of word frequency. One interesting finding here was that participants in all three groups showed less control over words at lower frequency levels, as seen in the fact that the percentage of responses that were canonical decreased as the frequency level of responses decreased. It is not then the case that any word that a participant had sufficient knowledge of to produce as a response was equally likely to be canonical. It seems that, starting from a cue, learners may link to a word which they have acquired sufficient knowledge of, regarding its form and meaning, to use as a response (presumably on a semantic basis), before having acquired knowledge of whether the cue and that word form a collocation. This may be seen as suggestive of the incremental nature of vocabulary acquisition (Churchill, 2008; Schmitt, 1998, 2010) (see Section 11.4 for further discussion).

Also interesting were the comparisons between the three collocation knowledge groups in the percentage of responses that were canonical at each frequency level. No significant differences were found between the low scorers and mid scorers at any of the JACET levels, the only significant difference between the mid scorers and high scorers was at JACET 1, while comparing low scorers and high scorers, there were significant differences at both JACET 1 and Beyond JACET 2. This suggests that the main area of development was with respect to JACET 1.

Combining the findings summarised above, the three groups of participants were distinguished as follows. First, contrasting the mid scorers and the low scorers, the mid scorers: (1) gave significantly more response tokens; and (2) gave significantly more JACET 1 responses. They therefore appear to have achieved higher *LexCombi*

2 scores by making better use of JACET 1 words. We may assume (due to their Yes/No scores) that both groups knew virtually all these items to some extent at least, so it seems the difference was in the quality or depth of their knowledge of these words. Second, comparing the high scorers and the mid scorers, the high scorers: (1) gave more responses; (2) used more types; (3) gave significantly more JACET 1 responses; and (4) a significantly higher percentage of their JACET 1 responses were canonical. It seems therefore that there were differences in both the quality/depth of their knowledge of words and differences in the breadth of their lexical knowledge. It should be noted again, however, that there were no differences between the groups in vocabulary size as estimated by the Yes/No test. Nonetheless, as mentioned earlier, it is possible that while the receptive vocabulary size of the groups did not differ, there could have been differences in their productive vocabulary. This could account for the apparent differences in the breadth of their vocabulary knowledge noted above.

These findings suggest that there is more than one path towards increased knowledge of collocations: it can be achieved by way of improving the quality or depth of knowledge about words or it can be achieved by extending the range of words known. If this is the case, a number of questions arise: What is the relative importance of each path? Are there changes in their relative importance at different stages in learners' development? Are there differences in their relative importance among individual learners? The above findings, what exactly is meant by vocabulary breadth and depth and some possibilities regarding these questions are discussed in Section 11.3.

10.5 Word class of participants' responses

Barfield (2009a) saw the word class of participants' canonical responses to *LexCombi 1* as providing insights into the development of collocation knowledge. Barfield divided the canonical responses to his 30 noun cues into five categories: adjectival, verbal, nominal, nominal/verbal and other. On the basis of this categorisation and comparisons between higher proficiency and lower proficiency groups, Barfield made two suggestions for how knowledge of collocations develops.

The first was that "adjective + noun collocations are the foundation of L2 collocation knowledge" (p. 108). This was based on the finding that both higher

proficiency and lower proficiency participants gave a large number of adjectival responses. However, Barfield did not consider the possibility that there may be differences among individual cues with respect to the word class of the responses they elicit. The preponderance of adjectival responses found by Barfield may therefore be in part a function of the particular set of cues he used.

Barfield's second claim, based on his finding of a large difference in the number of verbal responses between his higher and lower proficiency groups, was that "the marked increase in verbal collocates for the high group points to a major area of development in productive L2 collocation knowledge" (p. 108). Barfield's actual results, however, showed the higher group gave significantly more canonical responses in three of the five categories: nominal, verbal and nominal/verbal. In addition, there is a possibility that differences among individual cues may have influenced these results also.

This section thus examines two questions:

1. How much variety was there among the *LexCombi 2* cues in terms of the word class of the canonical responses they elicited?
2. Were there differences among the proficiency-based groups or among the collocation knowledge groups in the number of canonical responses of different word classes? If so, were these differences influenced by the results for individual cues?

In common with Barfield's study, each canonical response to each cue in the Chapter 9 data set was categorised as either adjectival, verbal, nominal, nominal/verbal or other.

The nominal/verbal category was required for responses such as *drive* for the cue CAR. This category was only used, however, when both verbal and nominal uses were probable for the cue in question. Thus, for example, the response *account* for the cue BANK was categorised as nominal, since it was judged unlikely that verbal uses of *account* were applicable in this case. For the majority of cues, only a small number of responses were assigned to the nominal/verbal category. There were just three cues, CAR, QUESTION and WATER, for which 10% or more of responses were nominal/verbal, and in each case this was due to a single recurrent response, (*drive* for the cue CAR, given by 34 of the 146 participants; *answer* for QUESTION, given by 43 participants; and *drink* for WATER, given by 60 participants).

It must also be acknowledged that the categorisation of responses into word classes involved some assumptions. One such was that participants giving a particular response had a particular use of that response in mind. In fact, it may have been that different participants had different uses of the response in mind, or indeed may have had no particular use in mind.

10.5.1 Differences between cues in the word class of canonical responses elicited

Differences between cues in the word class of the canonical responses they elicited were explored in two ways: in absolute terms and relative terms. In absolute terms, cues were identified for which responses of a single word class were dominant (i.e. responses of a single word class accounted for at least 50% of the total canonical responses). Table 10.9 shows that 14 cues elicited a majority of adjectival responses, while there were a number of cues for which at least half of the canonical responses were of another word class.

To look at this issue in relative terms, a 30 (cues) × 5 (categories) contingency table was constructed, the row and column totals of which allowed the calculation of an "expected" number of responses in each category for each cue. A Pearson's chi-

Table 10.9: Cues for which canonical responses of a single word class were dominant.

Adjectival responses dominant	Verbal responses dominant	Nominal responses dominant	Nominal/Verbal responses dominant	Other responses dominant
character condition country field machine name player sense size stage wall war water worker	letter picture	science		fact mile

square was then used to examine the extent to which the actual figures differed from these expected values. This found an overall significant difference between the 30 cues, $\chi^2(116) = 4531.244$, $p < .001$, indicating a different pattern of results across the variables. The standardised residuals for each cell of the contingency table were then examined to see which word classes were relatively underrepresented or overrepresented for each cue given the overall number of responses of each word class. This revealed a great deal of variety among the cues, as Table 10.10 shows. Indeed, for each and every of the 30 cues at least one of the five categories was either under- or overrepresented in comparison with the cues as a whole.

Table 10.10: Number of cues for which there were significantly ($p < .05$) more or fewer canonical responses of a particular word class as compared with the overall figures.

	Adjectival responses	Verbal responses	Nominal responses	Nominal/ Verbal responses	Other responses
Number of cues with significantly more responses of a particular word class	7	11	8	5	8
Number of cues with significantly fewer responses of a particular word class	13	15	13	15	16

10.5.2 Differences between participants in the number of canonical responses of different word classes

The second question was concerned with the number of canonical responses of different word classes given by different groups of learners. To re-cap, Barfield's claims were based on the number of responses of different word classes provided by two groups of participants of different proficiency levels as defined by TOEIC scores. In Barfield's study, there was a correlation between the participants' *LexCombi 1* scores and TOEIC scores. Barfield, therefore, was looking at groups that achieved different *LexCombi 1* scores and so his claims were, in effect, about the

locus of growth in participants' *LexCombi 1* scores. In the Chapter 9 data, in contrast, there was no correlation between *LexCombi 2* scores and proficiency (Section 10.3). Thus, in examining three proficiency-based groups in the Chapter 9 data set, the issue was not about the locus of growth in *LexCombi 2* scores since there was no such growth. Rather, differences between these groups in the number of canonical responses of different word classes would have indicated a qualitative difference between these participants, even though their quantitative *LexCombi 2* scores were equivalent.

Table 10.11 and Figure 10.2 present the data for the three proficiency-based groups and show there was very little difference between the groups for any of the word classes. Because of issues with normality and homogeneity of variances, a MANOVA could not be conducted and so a series of Kruskal-Wallis tests were used, with a Bonferroni correction applied setting the significance level at .01 (i.e. .05/5). These tests confirmed there were no significant differences:

Adjectival: $H(2) = 1.348$, $p = .510$.

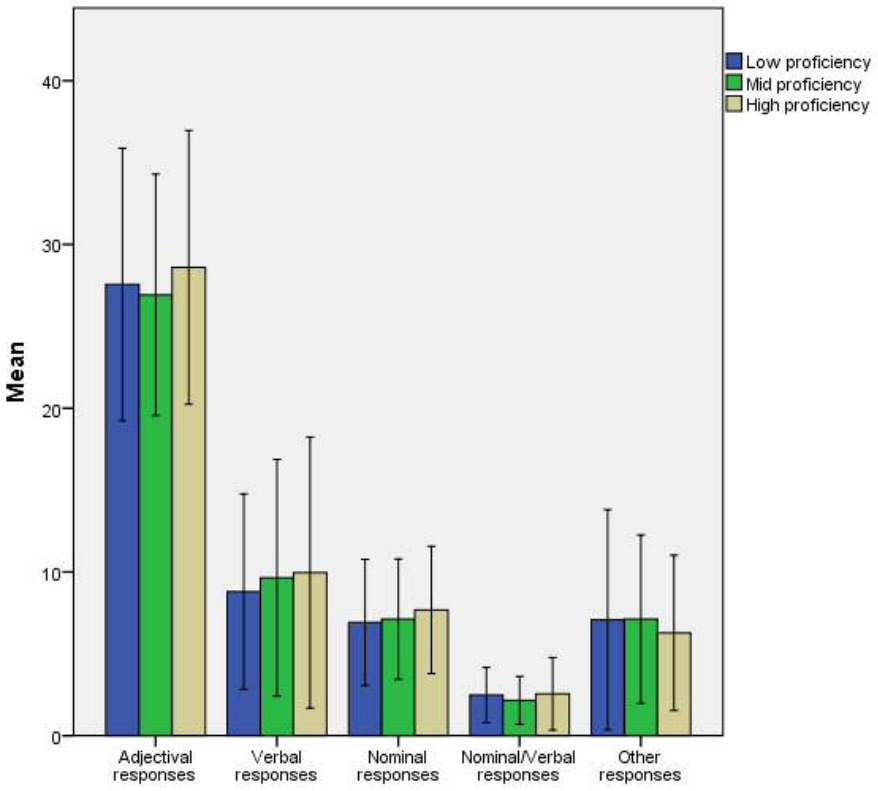
Verbal: $H(2) = 0.174$, $p = .917$.

Nominal: $H(2) = 0.417$, $p = .812$.

Table 10.11: Number of canonical responses of different word classes across three proficiency-based groups.

		Low proficiency (*N* = 25)	Mid proficiency (*N* = 25)	High proficiency (*N* = 25)
Adjectival responses	Mean	27.56	26.92	28.60
	SD	8.317	7.371	8.357
Verbal responses	Mean	8.80	9.64	9.96
	SD	5.965	7.233	8.279
Nominal responses	Mean	6.92	7.12	7.68
	SD	3.851	3.678	3.891
Nominal / Verbal responses	Mean	2.48	2.16	2.56
	SD	1.686	1.463	2.219
Other responses	Mean	7.08	7.12	6.28
	SD	6.733	5.142	4.739

Figure 10.2: Number of canonical responses of different word classes across three proficiency-based groups.



Note. Error bars represent 1*SD*.

Nominal/Verbal: $H(2) = 0.462$, $p = .794$.

Other: $H(2) = 0.906$, $p = .636$.

Given that there were no differences between the three proficiency-based groups in the number of canonical responses of different word classes, the three collocation knowledge groups were next analysed in the same way. Since there were stark differences in the number of canonical responses provided by these three groups, this analysis allowed the locus of growth in collocation knowledge to be identified in a manner akin to Barfield's work. Table 10.12 and Figure 10.3 present the data. To compare the groups, a MANOVA could not be conducted due to issues with normality and homogeneity of variances, and thus a series of Kruskal-Wallis tests were used with a Bonferroni correction setting the significance level at .01 (i.e. .05/5). A Bonferroni correction was also applied for post hoc Mann-Whitney tests, giving a significance level of .0167 (i.e. .05/3). The results showed significant differences between the groups in three of the five categories:

Adjectival: $H(2) = 15.058$, $p < .01$. Post hoc: a significant difference between the low scorers and the high scorers.

Verbal: $H(2) = 28.843$, $p < .01$. Post hoc: significant differences for all three comparisons.

Nominal: $H(2) = 3.435$, $p = .180$.

Nominal/Verbal: $H(2) = 5.462$, $p = .065$

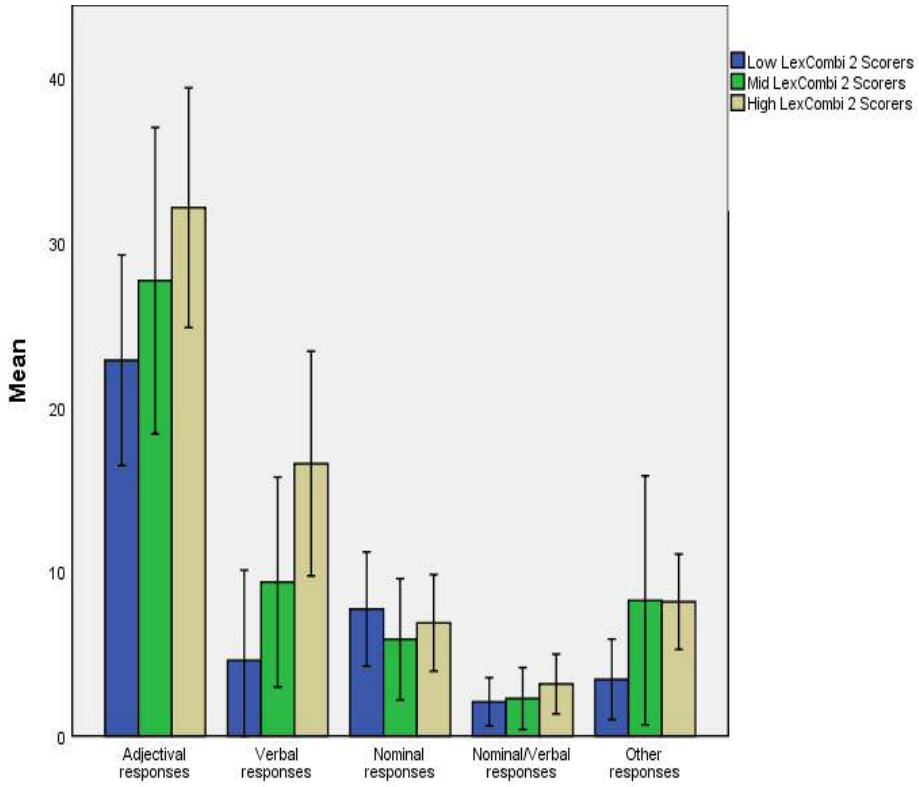
Other: $H(2) = 23.321$, $p < .05$. Post hoc: significant differences between the low scorers and both the mid scorers and the high scorers.

There were then differences among the collocation knowledge groups in the number of canonical responses of different word classes. However, the remaining question was whether these differences derived from generalised differences between the groups across the 30 cues or from differences among them in the responses to particular cues. If the latter, then the differences seen between the groups were at least partly due to the particular cues that *LexCombi 2* happens to include. In this case, making generalisations about collocational development would be questionable.

Table 10.12: Number of canonical responses of different word classes across three collocation knowledge groups.

		Low *LexCombi 2* scorers ($N = 25$)	Mid *LexCombi 2* scorers ($N = 25$)	High *LexCombi 2* scorers ($N = 25$)
Adjectival responses	Mean	22.84	27.68	32.12
	SD	6.408	9.317	7.293
Verbal responses	Mean	4.60	9.36	16.56
	SD	5.485	6.383	6.832
Nominal responses	Mean	7.72	5.88	6.88
	SD	3.470	3.689	2.934
Nominal / Verbal responses	Mean	2.08	2.28	3.16
	SD	1.470	1.882	1.818
Other responses	Mean	3.44	8.24	8.16
	SD	2.451	7.573	2.897

Figure 10.3: Number of canonical responses of different word classes across three collocation knowledge groups.



Note. Error bars represent 1*SD*.

This issue was examined by looking at differences in the canonical responses to individual cues among the three collocation knowledge groups. A 3 (groups) × 5 (categories) contingency table was constructed for each cue. Statistical tests were then run to see if there were significant differences among the groups in the number of responses of each word class. Fisher's exact test was used in each case since for a number of the cues there were expected frequencies below five for some of the cells in the contingency table. This analysis found a significant difference for 10 out of the 30 cues. An examination of the standardised residuals showed that in almost every case, the difference lay in the responses of the low scorers. The low scorers, in many cases, provided a greater than expected number of nominal responses, and in some cases a lesser than expected number of verbal responses to some individual cues.

10.5.3 Discussion

This section has examined to what extent cues vary in the word class of the canonical responses they elicit and whether there are differences in the word class of

the canonical responses elicited in accordance with the general (language proficiency) or specific (*LexCombi 2*) performance of the respondent. On the degree of variety among the cues, it was found that, in absolute terms, almost half of the cues elicited a majority of adjectival responses. However, there were also cues that primarily elicited verbal, nominal and other responses. In addition, in comparison with the overall numbers across the 30 cues, there was a great deal of variety among individual cues in the number of responses of different word classes, so much so that each and every cue differed in some way from the overall tendencies.

The fact that different cues can elicit such different patterns of responses may mean that describing any one type of response as the basis of collocation knowledge, as Barfield said of adjectival responses, is not possible. Rather, it may simply be that individual cues elicit individual patterns of responses. The suggestion that cues are idiosyncratic in terms of the types of responses they elicit fits with an increasing awareness, largely resulting from corpus linguistic approaches to linguistic description, of the idiosyncratic behaviour of words in general and the difficulties of describing word classes (see, for example, Smith, 2015; Taylor, 2012).

The second issue examined in this section was whether there were differences between groups of learners with respect to the word class of the canonical responses. It was found that among three proficiency-based groups, between which there were no differences in overall *LexCombi 2* scores, there were also no differences in the number of responses in any of the word class categories. In the case of three collocation knowledge groups, however, differences were found, specifically in the number of adjectival, verbal and other responses. The pairwise comparisons of the three groups provided a further picture of development. The mid scorers gave significantly more verbal and other responses than the low scorers; while the high scorers gave significantly more verbal responses than the mid scorers. Verbal collocates, therefore, appear to be the primary locus of growth as learners' knowledge of collocations develops. This accords with Barfield's claim that verbal collocates are the principal area of development, though his data actually showed significant growth in both verbal and nominal collocates.

Following on from the above finding, an analysis was carried out of whether groups of participants who demonstrated different degrees of collocation knowledge responded differently to particular cues. It was found that in most cases the three

collocation knowledge groups gave a comparable number of responses of each word class. However, for 10 cues there was a significant difference between the groups, and in most cases this was due to the responses given by the low *LexCombi 2* scorers, in particular the number of nominal responses given. There does seem to be some evidence therefore that certain cues elicit different types of responses from learners with different levels of collocation knowledge.

These findings might appear to give us a detailed view of collocational development. However, the fact that this development appears to differ cue-by-cue means that it may be difficult to make generalisations about development as a whole. That is, results dependent on the particular set of cues that *LexCombi 2* happens to include make any generalisation rather questionable. This may explain the discrepancy between Barfield's findings and those of the current study: that is, while both studies found significant differences between learners in the number of verbal collocates, Barfield also found a difference in the number of nominal collocates, while the current study found differences in the number of adjectival collocates and other collocates. This discrepancy may arise because Barfield's findings and the findings of this study come from investigations involving largely different sets of cues.

It should also be emphasised that, as with many other linguistic features, the set of responses to any particular cue exhibits a highly skewed distribution, as demonstrated by the recurrence of responses (see Sections 4.5.1 and 7.4.1). This means that what this section has described as the tendency of a cue to elicit responses of a certain class is in most cases the result of the classification to a certain word class of a very small number of items which occur as responses with high frequency. Further, this classification itself may be questioned due to the difficulty of identifying responses as belonging to a particular word class. Examining the number of response types of different word classes, rather than the number of response tokens, could overcome this problem to an extent. However, it is likely that the key finding to come out of this consideration of the word class of the canonical responses provided by *LexCombi 2* participants would still stand: that the individual nature of cues makes it hard to gain any overall picture of collocational development. Individual cues elicit individual patterns of responses of different word classes. Indeed, it is likely more accurate to say that cues do not elicit responses of a certain

word class at all. A cue elicits words, which we as researchers may see as belonging to a certain word class, but the word class may have no bearing on how learners respond. Ultimately, in examining the word class of responses, it might be that we are not really learning anything about the learners and the development of their collocation knowledge; all we are really doing is discovering something about the collocational behaviour of the cues. Thus, the whole enterprise of examining the word class of responses in order to gain insights into collocational development rests on a rather fragile base.

10.6 Participants' scores by alternative scoring lists

Responses to *LexCombi 2* are scored against a list of canonical responses for each cue (see Chapter 7). Under the current scoring approach, ALC, the list for each cue was made by combining four independently compiled sub-lists (explained in Sections 7.2.1-7.2.4) based on different sources: a compilation of collocations dictionaries, searches of corpora, L1-user norming and L2-user norming. These source lists were shown in Section 7.3 to be rather different from one another in terms of the collocates they contain. Further, each source list implies a somewhat different interpretation of the concept of collocation (see Section 7.6).

One question then was whether groups of participants differed in terms of the scores achieved against these source lists. Such differences could be seen as indicating that there are qualitative differences in learners' knowledge of collocation as their ability to produce collocations develops. For example, if participants achieving high *LexCombi 2* scores achieved even greater scores in comparison with other participants under the corpus-based source lists, it might be argued that this was an indication of a growing sensitivity to actual usage as encountered in input as collocation knowledge develops.

The question pursued was therefore:

1. To what extent were there differences in the scores of the collocation knowledge groups under alternative scoring lists?

To address this question, each participant's score was re-calculated under each of the source lists using CollCheck (Imao & Brown, 2014). Since the collocation knowledge groups were formed on the basis of their *LexCombi 2* scores, and since the *LexCombi 2* scores were derived from the combination of the four source lists, it

was thought that the groups might well also be distinguished under each source list. On the other hand, Section 7.3 showed that the degree of overlap between the source lists was relatively slight. It was thus thought possible that a participant's score under one of the source lists could be rather different relative to their overall *LexCombi 2* score.

10.6.1 Scores under alternative lists

Table 10.13 provides the scores under each source list for the three collocation knowledge groups. As can be seen, the three groups were in the same order under each source list and comparisons of the scores using one-way ANOVAs found a significant difference between the three groups in each case: dictionary-based scoring $F(2, 72) = 109.782$, $p < .001$; corpus-based scoring $F(2, 72) = 71.372$, $p <$

Table 10.13: Scores for the low, mid and high *LexCombi 2* scorers under different scoring source lists.

		Mean	*SD*
Dictionary-based scoring	Low *LexCombi 2* scorers ($N = 25$)	28.44	5.276
	Mid *LexCombi 2* scorers ($N = 25$)	39.40	4.907
	High *LexCombi 2* scorers ($N = 25$)	50.68	5.706
	All participants ($N = 146$)	39.46	8.698
Corpus-based scoring	Low *LexCombi 2* scorers ($N = 25$)	20.56	5.067
	Mid *LexCombi 2* scorers ($N = 25$)	27.24	3.586
	High *LexCombi 2* scorers ($N = 25$)	36.36	5.251
	All participants ($N = 146$)	28.21	6.637
L1 norms	Low *LexCombi 2* scorers ($N = 25$)	26.04	4.748
	Mid *LexCombi 2* scorers ($N = 25$)	32.68	5.647
	High *LexCombi 2* scorers ($N = 25$)	38.88	6.679
	All participants ($N = 146$)	32.88	6.656
L2 norms	Low *LexCombi 2* scorers ($N = 25$)	31.52	4.700
	Mid *LexCombi 2* scorers ($N = 25$)	42.28	4.468
	High *LexCombi 2* scorers ($N = 25$)	54.08	4.222
	All participants ($N = 146$)	42.49	8.128

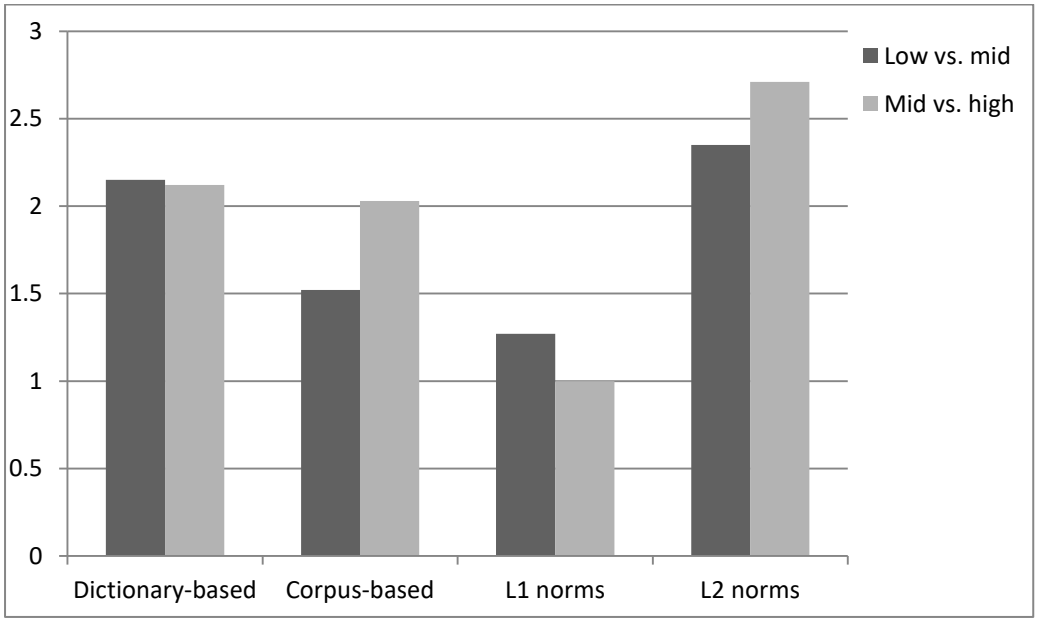
.001; L1 norms scoring $F(2, 72) = 31.223$, $p < .001$; L2 norms scoring $F(2, 72) = 159.48$, $p < .001$. Post hoc Tukey tests showed each group to be significantly different from the other two in all four cases.

Effect sizes (eta squared), that is the statistical magnitude of the difference, for each overall effect were: dictionary-based scoring $r = .87$; corpus-based scoring $r = .82$; L1 norms $r = .68$; L2 norms $r = .90$. Under all four scoring approaches, there was, then, a strong effect for the overall result. In comparing these effects, small differences in the figures should not be overemphasised. Nevertheless, the effect under L1 norms scoring, albeit strong, did appear to be comparatively weaker, while the effect was strongest under L2 norms scoring.

The effect sizes (Cohen's *d*) for the pairwise contrasts are presented graphically in Figure 10.4. These contrasts represent shifts in the relative position of the three groups. The greatest difference between the two pairwise comparisons was under corpus-based scoring, for which there was a weaker effect for the low versus mid comparison as compared with the mid versus high comparison.

Interestingly, it was possible to investigate the results for corpus-based scoring further since the corpus-based lists themselves were compiled from two sets of

Figure 10.4: Effect sizes (Cohen's *d*) for the pairwise comparisons between low and mid *LexCombi 2* scorers and between mid and high *LexCombi 2* scorers under different source lists.



component lists: one based on frequency-based searches of the corpora and the other on MI-based searches (frequency showing simply how often a given word combination occurs, MI indicating how much more than expected by chance a combination occurs; see Section 7.2.2).

Having re-scored the responses with these component lists (see Table 10.14) and conducted one-way ANOVAs once more, there were significant differences between the groups in each case: corpus-frequency-based scoring $F(2, 72) = 87.592$, $p < .001$; corpus-MI-based scoring $F(2, 72) = 35.876$, $p < .001$. The effect sizes (eta-squared) were $r = .84$ under corpus-frequency-based scoring and $r = .71$ for corpus-MI-based scoring. The effect size for corpus-frequency-based scoring was therefore very similar to the overall corpus-based scoring effect ($r = .82$), while the effect size for corpus-MI-based scoring was somewhat weaker. Looking at the pairwise contrasts, post hoc Tukey tests showed each group to be significantly different from the other two, with the exception of the low versus mid contrast under corpus-MI-based scoring. Figure 10.5 presents the effect sizes (Cohen's *d*) for these contrasts.

10.6.2 Discussion

This section has considered what an examination of the scores achieved by participants under different scoring criteria may reveal about collocational development. Two principal findings were made.

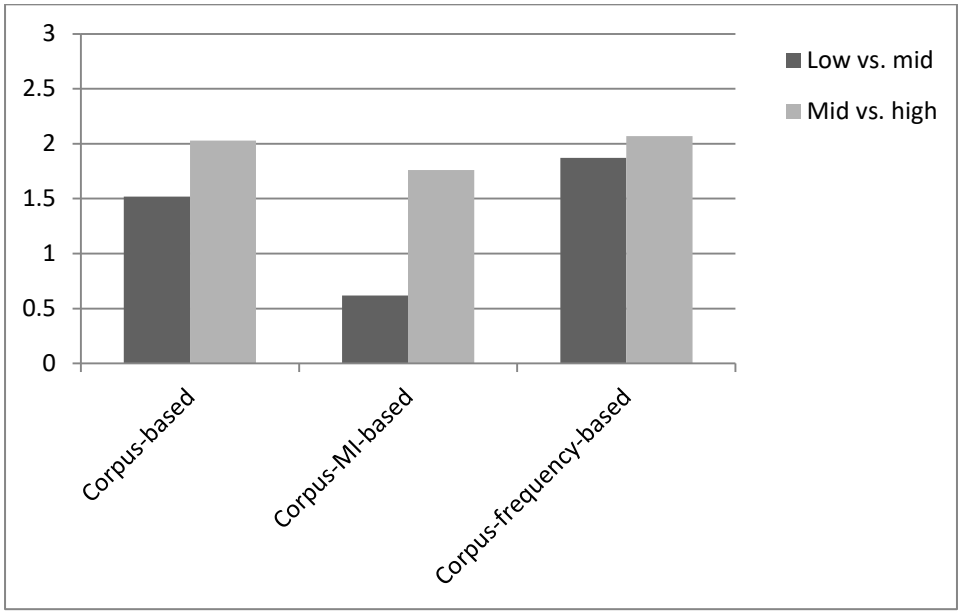
First, it was seen that, just as with the overall *LexCombi 2* results, the collocation knowledge groups were sharply distinguished when scores were re-calculated under each of the four source lists instead. It was also seen, however, that there were differences in the degree to which groups of participants were distinguished by different source lists. Looking at the effect sizes for the differences between the groups under each source list, the effect was somewhat weaker under L1 norms scoring and was strongest under L2 norms scoring.

These results are interpreted as showing that it was towards L2-user norms that the highest scoring participants had progressed. Likewise, they seem to have progressed less strongly towards the L1-user norms. The L2-user norms were compiled from responses to the *LexCombi 2* cues by Japanese L1 speakers who are highly proficient in English and use English regularly in everyday life. The notion that Japanese learners' collocational development may be progressing towards

Table 10.14: Scores for the low, mid and high *LexCombi 2* scorers under two corpus-based sets of lists.

		Mean	*SD*
Corpus-frequency-based scoring	Low *LexCombi 2* scorers (*N* = 25)	17.24	4.055
	Mid *LexCombi 2* scorers (*N* = 25)	23.96	3.048
	High *LexCombi 2* scorers (*N* = 25)	32.04	4.614
	All participants (*N* = 146)	24.59	6.075
Corpus-MI-based scoring	Low *LexCombi 2* scorers (*N* = 25)	14.52	4.806
	Mid *LexCombi 2* scorers (*N* = 25)	17.32	4.210
	High *LexCombi 2* scorers (*N* = 25)	24.84	4.327
	All participants (*N* = 146)	18.85	5.293

Figure 10.5: Effect sizes (Cohen's *d*) for the pairwise comparisons between low and mid *LexCombi 2* scorers and between mid and high *LexCombi 2* scorers under corpus-based scoring and the two component sets of lists.



such norms seems therefore logical. There has been much discussion in applied linguistics of the commonplace assumption that command of the language akin to L1 users is an appropriate goal for language learning. It has been argued that this goal, being unrealistic, unwarranted and unfair, should be replaced by a goal of successful use of the language (V. Cook, 1999), while it has also been noted that, as bilinguals,

learners can never become monolinguals in the target language (Grosjean, 1989). In these results, there may arguably be an indication of the reality of this argument. These participants appear to have been progressing towards becoming Japanese users of English, not L1 users of English.

The second principal finding concerned the effect sizes for the pairwise comparisons between the three groups. It was seen that although under three of the four source lists the two pairwise effect sizes were broadly similar, there was a greater contrast between the two effect sizes under corpus-based scoring. Examining the scores of learners under the two components of the corpus-based lists, it was found that under corpus-frequency-based scoring the effect size was somewhat stronger, and the two pairwise effects were also very strong. Under corpus-MI-based scoring, in contrast, there was no significant difference between the low *LexCombi 2* scorers and the mid *LexCombi 2* scorers, while between the mid *LexCombi 2* scorers and the high *LexCombi 2* scorers a significant difference and a strong effect was seen. This may be interpreted as indicating sustained sensitivity to the representation of English provided by frequency-based searches of the corpora, and a change from no effect to a strong effect for the representation of English provided by MI-based searches of the corpora as collocational proficiency develops.

This contrast between the corpus-frequency and corpus-MI results is interesting given research (highlighted in Section 3.2.3) suggesting that learners are relatively insensitive to MI-based collocations as compared with frequency-based collocations (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis, et al., 2008; González Fernández & Schmitt, 2015). It may be that learners can develop knowledge of MI-based collocations, but this development takes place at a later point in comparison with knowledge of frequency-based collocations. These issues are discussed further in Section 11.4.

10.7 Conclusions

This chapter has sought to explore the potential of *LexCombi 2* data to provide us with insights into the development of learners' knowledge of collocations. The intention was to highlight areas that future work may wish to investigate in a more rigorous and targeted fashion. The chapter has looked into four areas:

- the relationship between *LexCombi* scores and general proficiency, it being

concluded that, with respect to learners of the type considered, there is essentially no such relationship;

- the range and frequency of participants' *LexCombi 2* responses, with differences found between groups of participants distinguished by their *LexCombi 2* scores in both the range of words deployed and their use of highly frequent words;
- the word class of participants' *LexCombi 2* responses, the conclusion being that this reveals more about the collocational behaviour of the individual cues than about the learners, and thus further investigation in this area will not be pursued;
- the scores of groups of participants distinguished by their *LexCombi 2* scores under the four scoring source lists, which revealed that, while the groups were still sharply distinguished in each case, there were differences in the size of the effect under different scoring source lists.

This exploration has then raised a number of questions about the nature and development of learners' collocation knowledge:

- Why is it that there was no correlation between participants' Yes/No test scores and their *LexCombi 2* scores?
- If knowledge of collocations can be achieved by way of improving the quality or depth of knowledge about words or by extending the range of words known, what is the relative importance of each path?
- Are there changes in the relative importance of each path at different stages in learners' development?
- Are there differences in the relative importance of each path among individual learners?
- Why do learners seem to make more progress in acquiring some types of collocation than others?

The next chapter reviews the strengths and weaknesses of *LexCombi 2* as a research instrument, and discusses wider issues raised across this thesis about collocations and how learners acquire them, including the questions listed above.

Chapter 11 General discussion

11.1 Introduction

In Section 3.4, three questions were set out which this thesis has sought to answer:

- What strengths and weaknesses does *LexCombi* show as a tool for measuring L2 learners' productive knowledge of collocations and how can it be improved?
- What insights into the development of productive collocation knowledge may *LexCombi* provide?
- What wider issues for our understanding of collocation are raised by the development of *LexCombi* and an initial exploration of *LexCombi* data?

Chapters 9 and 10 have provided some answers to the first two of these questions. In Chapter 9 *LexCombi 2*, the final form of the instrument as developed in this thesis, was judged capable of providing data of good quality, making it suitable for its intended purpose as an elicitation instrument to enable research into learners' productive knowledge of collocations. In Chapter 10 *LexCombi 2* data was explored with regard to four areas and it was found that: (1) there appears to be no correlation between *LexCombi 2* scores and general English proficiency; (2) groups of participants formed on the basis of their *LexCombi 2* scores primarily differed in how well they made use of the most frequent English words as responses; (3) in terms of the word class of the responses elicited, individual *LexCombi 2* cues elicited their own patterns of responses; and (4) groups of participants distinguished by their *LexCombi 2* scores were also clearly distinguished under other scoring approaches, but there were differences in the strength of the effect under different approaches.

This chapter considers the third question and looks at wider issues raised across the thesis as a whole. Building on Chapter 9's discussion of *LexCombi 2*'s suitability for exploring learners' productive knowledge of collocations, Section 11.2 considers the validity of *LexCombi 2* as a research instrument. Sections 11.3-11.5 then explore vocabulary breadth and depth, the learning of collocations and the representation of collocations in the mental lexicon.

11.2 The validity of *LexCombi 2*

To determine whether *LexCombi* is suitable for probing learners' productive

knowledge of collocations, the empirical phase of this thesis began by trialling *LexCombi* and exploring how learners responded to it. This led to a series of changes being made to the instrument: its format was changed, a new set of cues were selected and a new scoring approach was adopted. This section makes use of a framework for assessing the validity of a test and discusses to what extent *LexCombi 2*, the current form after this series of changes, is a valid instrument for exploring learners' productive knowledge of collocations.

There are a number of approaches to establishing validity, and no single method is dominant. However, in recent years the argument-based approach to validity has attracted interest (Chapelle & Voss, 2013; Purpura, Brown, & Schoonen, 2015) and this approach will be applied to help assess *LexCombi 2*. The following analysis is based on the descriptions of the argument-based approach by Chapelle and Voss (2013) and Purpura, Brown and Schoonen (2010), along with applications of the approach to vocabulary research by Koizumi (2015) and Voss (2012).

In the argument-based approach, validity is considered to be a matter of degree. Tests (or more properly test scores and interpretations) do not either have or lack validity; they have a certain degree of validity. The degree of validity is demonstrated by building a validity argument. This involves a series of steps, for each of which supporting evidence should be collected. Since the validity argument is viewed as incremental and additive, a lack of evidence for one step makes it difficult to continue to the next.

The six steps are: domain definition, evaluation, generalisation, explanation, extrapolation and utilization. Koizumi (2015) suggests that for tests intended for research purposes, the first four of these steps suffice: the fifth step being about extrapolating the test results to a learning context, and the sixth step about using test results to make decisions. Since *LexCombi 2* is intended not as a test of collocation knowledge, but as an elicitation instrument for research purposes, neither of these steps applies, and evidence is presented below only for the first four steps.

As validity is about test scores rather than an instrument itself, the following is concerned with the Chapter 9 data. Portions of earlier chapters in the thesis are also of relevance, however, in that they explain the series of changes that were made to *LexCombi 1* which resulted in the current form of the instrument, *LexCombi 2*, as used in collecting the Chapter 9 data.

- Domain definition. This concerns whether test items cover or elicit knowledge relevant to the target domain. *LexCombi 2* was developed with Japanese learners of English and its scoring is based in part on norms for Japanese users of English (as Section 9.4 highlighted), but it is not focused on any particular domain of language use. The question, then, is whether the *LexCombi 2* task is akin to and relevant to the production of collocations in general settings of L2 use. First, the *LexCombi 2* cues are all words of high frequency. High-frequency words are used in a wide variety of contexts and thus knowledge of their collocates is relevant to all types of productive language use. Second, the full-phrases responses study (Sections 5.2-5.4) considered the relevance of the *LexCombi* task; in particular whether the ability to give a response to a noun cue is indicative of the ability to use that response in combination with the cue. It was concluded, with some caution since for practical reasons is was necessary to investigate this issue by making comparisons across groups, that *LexCombi* participants likely did have knowledge of the use of cue-response combinations since approximately 90% of one group of participants' canonical responses exhibited "regular" use.
- Evaluation. This is about the extent to which a test provides accurate information. First, before scoring the *LexCombi 2* responses, obvious spelling errors were corrected (Table 4.1), since *LexCombi 2* is a measure of collocation knowledge, not of spelling ability. Second, the think-alouds study reported in Sections 5.5-5.7 found that many *LexCombi* responses were seemingly given spontaneously. There was therefore a concern, backed by a number of observed instances, that participants may have responded with associations generally rather than collocations in particular. This could cause participants to miss out on opportunities to give scoring responses. This concern drove the development of a new format for *LexCombi*, trialled in Chapter 6, which was intended to guide participants towards providing collocational responses. As Chapter 6 showed, the adapted format did not eliminate non-collocational responses, but did have a number of advantages over the original format and was therefore adopted. Third, out of concern that the original approach to scoring *LexCombi* was too narrow and did not

properly reflect learners' knowledge of collocations, the scoring of *LexCombi* was reviewed and new scoring lists developed (Chapter 7). Fourth, the *LexCombi 2* scores were normally distributed, with no ceiling or floor effects (Section 9.3.1). Finally, a Rasch analysis of the scores data found no misfitting items (Table 9.3). There were, however, two items with disordered Rasch-Andrich thresholds (Table 9.4), and it is possible that these are the result of chained responses. This is a threat to the validity of *LexCombi 2* and thus the two items concerned, for which the thresholds were not disordered in the cue trialling data (Section 8.4.1), should be carefully examined in future uses of *LexCombi 2* and may need replacing with other items.

- Generalisation. This is about the consistency of test items. First, as detailed in Chapter 8, in order to avoid a number of potentially problematic issues, several criteria were adopted to identify possible cues. These cues were subsequently trialled and a principled selection made from among those that were found to perform well. Second, item and person reliability figures for *LexCombi 2* are of relevance. In the Chapter 9 data, item reliability was seen to be very high at .96, while person reliability was somewhat lower at .71 (Table 9.5). This level of person reliability means *LexCombi 2* may not be suitable for decision-making (i.e. for determining, for example, whether learners may take a course or not), but this is not its intended use in any case.
- Explanation. This is the extent to which test scores reflect a construct relevant to the context. In this case, this is about how well *LexCombi 2* scores reveal a construct of productive collocation knowledge. First of relevance is the Rasch analysis of the Chapter 9 data, in particular the fit statistics and dimensionality (Table 9.3). This showed no issues with the fit of the *LexCombi 2* items, but did reveal the presence of a possible additional dimension in the data (i.e. the possibility that another trait, besides productive knowledge of collocations, may have influenced the *LexCombi 2* scores; see Section 9.3.2). An examination of the items involved, however, led to the conclusion that this was likely due to random variation and it was concluded that *LexCombi 2* does measure a single construct. Also of

relevance once more is the full-phrase responses study (Sections 5.2-5.4), which investigated whether participants are able to use their *LexCombi* responses in combination with the cue. The positive conclusion of this investigation goes partway towards linking *LexCombi* with the construct of productive collocation knowledge. Third, this step can be supported by differences in test scores over time or differences among groups of participants. While the former could not be explored in this thesis, the latter was investigated with respect to the relationship between *LexCombi* scores and general proficiency (Section 10.3). It was found that *LexCombi 2* scores had very low correlations with proficiency as judged by Yes/No test scores. This indicates that the construct that *LexCombi 2* measures is different from that of the Yes/No test. While on one level, the low correlation between the two measures may seem unexpected (i.e. one might anticipate a correlation between collocation knowledge and general proficiency), it is arguably not surprising that a test of productive knowledge of collocations and a test of receptive word recognition seemingly involve different constructs. Furthermore, *LexCombi 2* does produce a good range of scores, with learners spaced out along a performance continuum.

To summarise, the above has highlighted three areas of concern for *LexCombi 2*'s validity: (1) Disordered Rasch-Andrich thresholds for two items. This may indicate the chaining of responses and so the two items may need to be replaced if this problem is found to persist. Fortunately, the cue trialling process described in Chapter 8 means there is a ready pool of items that have already been successfully trialled which could be used instead. (2) A somewhat low person reliability figure. This means *LexCombi 2* should not be used for decision-making. For its intended use, however, as an elicitation instrument for research purposes, this level of reliability is acceptable. (3) The possible presence of another dimension (i.e. another trait, in addition to knowledge of collocations, potentially influencing the *LexCombi 2* scores). In future uses of *LexCombi 2*, dimensionality should be checked again. Should an additional dimension be found once more and the same items be flagged up in the analysis, the current assumption, that the indication of another dimension in the data stems from random variation, would clearly be quite questionable.

While these three areas of concern require attention, and further evidence for

LexCombi 2's validity would be useful, the above account does overall constitute an initial validity argument for *LexCombi 2* as an elicitation instrument for research purposes. *LexCombi 2* does then seem suitable for exploring learners' productive knowledge of collocations, and indeed has a number of merits as a research instrument: (1) it elicits a large amount of data in a short period of time; (2) the data is in a form (single-word responses) that is easily analysable; (3) it is open and allows the elicitation of collocates of any word class; (4) it is a single instrument with a single set of cues, but can be used with learners with quite considerable differences in ability, enabling cross-sectional research designs; and (5) since there is a pool of already trialled cues available, all calibrated on a single scale by a Rasch analysis, it can be administered multiple times with cues varying between administrations, making longitudinal research designs possible.

The above features mean that *LexCombi 2* could be used quite flexibly to elicit data from learners for research purposes. As to the contexts in which it could be so used, there are some inevitable limitations as well as opportunities, due to its design. With regards to proficiency, provided that it could be reasonably assumed that the learners would have a minimum of at least some familiarity with the cue words, *LexCombi 2* could be used with learners of almost any level of proficiency. In respect of age, the abstract nature of some of the words used as cues means that it would not be suitable to use *LexCombi 2* with younger children, but from the teenage years upwards there should be no such issues and so *LexCombi* 2 could be used with learners across a wide age range. Finally, regarding linguistic/cultural background, the ALC scoring lists, one component of which is norms from Japanese users of English, are clearly tailored towards the learners of interest in this thesis (i.e. Japanese university learners of English). However, it would be possible to adapt the scoring of *LexCombi 2* so that it could be used with learners of other linguistic/cultural backgrounds. This could be done by replacing the norms derived from Japanese users of English with the norms of another relevant group, whether a linguistic group (e.g. the norms of Korean users of English) or a professional/vocational group (e.g. multinational university students studying successfully in an English-speaking country). *LexCombi 2* should therefore be of potential interest to researchers in a variety of contexts.

Having considered the validity of *LexCombi 2* as a research instrument, the next

sections explore wider issues that have arisen in the course of this thesis, beginning with vocabulary depth and vocabulary breadth.

11.3 Vocabulary depth and vocabulary breadth

Section 10.4 suggested that learners' *LexCombi 2* scores may depend in part on the depth, quality or organisation of their vocabulary knowledge. This was based on two observations. First, among three groups of participants distinguished by their *LexCombi 2* scores, the mid scorers used a similar number of word types to the low scorers but gave significantly more response tokens. This suggests learners from these two groups were drawing on a productive lexicon of a similar size, but differed in their ability to deploy their vocabulary. Second, comparing the three groups in terms of the percentage of responses at different frequency levels which were canonical, the main difference was in the use of JACET 1 words, the most frequent words. That is, the groups were distinguished by how successfully they made use of highly frequent vocabulary items. This is especially interesting when we consider that, on the basis of the Yes/No test results, participants in all three groups were most likely familiar with almost all the JACET 1 words. Is *LexCombi 2*, then, a test of vocabulary depth?

The notion of "vocabulary depth" is widely discussed among vocabulary researchers. Read (2004) notes, however, that in contrast to the term "vocabulary breadth" which is generally accepted as being about size, there is no such agreement regarding vocabulary depth, with researchers using the term to talk about three things:

1. precision of meaning, "the difference between having a limited, vague idea of what a word means and having much more elaborated and specific knowledge of its meaning" (p. 211);
2. comprehensive word knowledge, "knowledge of a word which includes not only its semantic features but also its orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics" (p. 211);
3. network knowledge, "the incorporation of the word into a lexical network in the mental lexicon, together with the ability to link it to – and distinguish it from – related words" (p. 212).

This third sense is perhaps the most prominent and is closely linked with Meara's description of the lexicon as featuring two dimensions: size and organisation (Meara, 1996; Meara & Wolter, 2004). Milton (2009) notes that many researchers are interested in this aspect of depth because of the possibility that it may be independent of breadth. There may be learners with a similar breadth of vocabulary, but different degrees of depth, which might explain "how learners with the same volumes of vocabulary knowledge can sometimes perform so differently in academic examinations and in practical communication" (p. 150).

It is not clear, however, that vocabulary depth can be separated from breadth. Read (2004) notes that although depth and breadth are often contrasted and spoken of as if they were quite distinct, they may in fact be closely related. Vermeer (2001) argues that there is no real difference between depth and breadth: depth is dependent on breadth since as breadth increases the number of connections between items inevitably increases also. Qian (2002), however, used multiple regression to show that, despite strong correlations (.70) between the two, depth made a significant contribution, independently of breadth, to explaining the variance in scores on a reading test, a finding backed by several other studies (see Schmitt, 2014).

Milton (2009) shares the view that vocabulary depth and breadth may be separate, but are very much related. He reviews several tests that purport to tap vocabulary depth and suggests that they are unsuccessful in doing so due to the interrelation of depth and breadth. In particular, Milton discusses Wolter's (2005) *V_Links* test, in which participants are presented with 10 randomly selected words and are asked to indicate pairs of words that are linked in some way. Milton argues that, because many links in the lexicon are not language specific, this is actually a breadth test. That is, teenage/adult L2 learners already have knowledge of most of the links since they have a mature L1 lexicon, but may lack L2 labels to allow the links to be demonstrated. *V_Links* is not then testing whether learners know a link (for they probably do), but whether learners know the words that are involved in that link. Continuing this line of thought, Milton argues that, from an L2 perspective, vocabulary depth is a narrower concept than often thought: it is not about all links in the lexicon but is restricted only to links that are not predictable from the L1.

Links not predictable from the L1 are very often collocations. This is not to say that unpredictable links and collocations are equivalent: there may be unpredictable

links that have a cultural basis (e.g. foods associated with particular meals) and are not collocations, and between most pairs of languages there are certainly many L2 collocations that are predictable from the L1. Nevertheless, many unpredictable links are collocations, and even in the case of predictable L2 collocations learners may not always be willing to trust the link that they suspect exists when producing language, depending on factors such as the perceived distance between the two languages (Yamashita & Jiang, 2010). Thus, *LexCombi 2*, in eliciting learners' productive knowledge of L2 collocations, may be in effect a test of vocabulary depth (as conceptualised by Milton as links between words not predictable from the L1).

This could provide an answer to one question raised in Section 10.7: Why is it that there was no correlation between participants' Yes/No test scores and their *LexCombi 2* scores (as seen in Section 10.3)? Might it be that these two tests represent independent measures of breadth and depth? The Yes/No test was of course intended as a proxy measure of general proficiency, but it is a measure of vocabulary size. *LexCombi 2*, it seems, unlike the measures of depth reviewed by Milton, may be able to tap learners' depth of knowledge largely without dependence on the size of their vocabulary. If this is the case, it is in part because *LexCombi 2* uses only highly frequent words as cues. This means that, provided that the learners are not beginners, what is at issue when learners respond to *LexCombi 2* is not whether a cue is known, but whether collocations involving that cue are known. If *LexCombi 2* were used with beginners, such that some of the cues were unknown to participants, scores of zero would be highly likely for those cues and a correlation with Yes/No test scores would likely be found.

In the Chapter 9 data, all the participants very probably had sufficient vocabulary size to be familiar, to some extent, with all of the *LexCombi 2* cues. Thus, *LexCombi 2* distinguished the participants by the depth of their knowledge of the cues. Section 10.3.2 suggested that the Japanese education system may have had an influence on the correlation results in that it encourages receptive skills required for success on tests rather than productive skills needed for genuine language use. Accordingly, while all learners are encouraged to develop the receptive skills necessary for test success, those learners with little interest in English (or aptitude) beyond passing tests may do relatively poorly on *LexCombi 2* relative to their vocabulary size. In contrast, those with some interest in using English may do

relatively well on *LexCombi 2*. It is therefore possible that while *LexCombi 2* scores do not correlate with general proficiency, there is a correlation with productive skills. A relationship between the use of multi-word units, including collocations, and language production has been suggested by many scholars (e.g. Lewis, 1993; Pawley & Syder, 1983) and, as mentioned in Section 1.3, has been found empirically (Boers, et al., 2006; Hsu & Chiu, 2008; Kyle & Crossley, 2015; Stengers, et al., 2011; Wood, 2010). It may be useful therefore in future work to investigate whether there is a correlation between *LexCombi 2* scores and measures of language production.

Returning to the findings of Section 10.4, it was also found that those participants who achieved the highest *LexCombi 2* scores, in addition to making better use of the most frequent English words, also made use of more types in their responses. Despite the Yes/No results showing the high scorers to have a similar receptive vocabulary size to other participants, their use of more types perhaps indicates that they have a larger productive vocabulary. This result, combined with the findings regarding how successfully different participants made use of the most frequent words, led to the suggestion in Section 10.4.3 that there may be two ways in which collocation knowledge can be developed: by increasing vocabulary breadth or by increasing vocabulary depth. Putting this in terms of the network view of the lexicon, as pursued by Meara (Meara, 2004, 2006; Wilks & Meara, 2002), which envisions the lexicon as a series of nodes (words) which are interconnected, the development of collocation knowledge can be seen as adding new words to the lexicon or adding new links to the lexicon.

If it is the case that there are two pathways along which collocation knowledge develops, this raises three questions, as Section 10.4.3 noted: (1) What is the relative importance of each pathway? (2) Are there changes in their relative importance over the course of learners' development? (3) Are there differences among individual learners in their relative importance? For the first two of these questions, there are hints in the data that allow some speculations.

On the first question, it may be noted that with regard to the three collocation knowledge groups, what has been interpreted as depth of knowledge of words distinguished both the low scorers from the mid scorers and the mid scorers from the high scorers. Breadth of knowledge, in contrast, only distinguished the mid scorers and the high scorers. This would suggest that depth is the more important pathway in

that it seems to have an impact over a wider range of development.

This pattern of results also seems of consequence for the second question. To state the results differently, the low scorers and the mid scorers seemed to be distinguished by the depth of their knowledge of high-frequency words, while the mid scorers and high scorers were distinguished both by the depth of their knowledge of high-frequency words and by the breadth of their vocabulary. It might then be that productive knowledge of collocations develops first in terms of depth and that breadth becomes important only at a later stage.

What could account for such a pattern of development? In terms of the network view of the lexicon, the answer may be simply that adding links between existing items in the lexicon is easier than adding new items to the lexicon. This is because adding a new item necessarily also involves adding a new link: presumably, an item cannot be free of any links at all and so must be linked to at least one other item. Thus, while adding a link to existing nodes entails only one operation, adding a new node entails two or more.

Regarding the third question, individual differences among learners, the data collected for this thesis does not allow any speculations. Milton (2007) and Booth (2013) have found differences among learners with regard to vocabulary acquisition and learning styles, and perhaps such differences pertain in this area as well. It is also possible that individual differences could arise due to variations in the nature of learners' exposure to English or variations in instructional practices which affect how learners attend to the language they are exposed to.

All the above is predicated on an assumption that learners first acquire individual words and then acquire links between those words. However, this is not necessarily the case. Wolter (2009) suggests that in the L1 words can be acquired incidentally as part of collocations. How much knowledge of such words is ultimately acquired depends on their collocational productivity (i.e. how many collocations each is part of). For a word with high collocational productivity, we may first learn to recognise the word, then gain some understanding of the patterns it occurs in and its prosody, and later gain definitional meaning. For words with limited collocational productivity, definitional meaning may never be acquired. Wolter suggests that for L2 learners this meaning-last process of acquisition may occur less often since L1 knowledge makes them more inclined to analysis and to seeking meaning from the

beginning. Nevertheless, meaning-last acquisition may occur for L2 learners in some cases and there are occasional signs of this in the data collected for this thesis. Some rather infrequent words occurred as responses in the Chapter 9 *LexCombi 2* data, such as *vending* for the cue MACHINE and *slumber* for PARTY. Neither of these words is a loanword in Japanese, neither appears in the JACET8000 word list and in both cases the collocation is very strong in terms of frequency (i.e. each has a high MI score in COCA). On occasion, therefore, learners may acquire knowledge of word forms as component parts of collocations without necessarily acquiring knowledge of the words as independent items. It does seem likely, however, that this occurs only in a minority of cases. Such cases are unlikely therefore to substantially impact the pattern of results reported in Section 10.4 or the interpretation of those results presented above.

11.4 Learners and collocations

As reported in Section 9.3.1, the mean *LexCombi 2* score for the 146 low-intermediate to high-intermediate proficiency participants was 53.92 out of 90. Since we do not have data on what sort of scores competent users of English (L1 or L2) would achieve on *LexCombi 2*, it is not possible to make any statements on the degree of collocation knowledge these learners possess, but it is clear they have some knowledge of collocations. Yet, as noted in Chapter 1, collocations have long been seen as an area of particular difficulty for learners, with even highly advanced learners reportedly struggling with collocations. What can account for the fact that learners at low-intermediate to high-intermediate levels have knowledge of quite a number of collocations, but learners at advanced levels apparently find collocations challenging?

A possible explanation is offered by Jiang's (2000; 2002; 2004) model of the process of vocabulary acquisition. This model sees vocabulary acquisition as an incremental process (see also Churchill, 2008; Schmitt, 1998, 2010) and specifies three stages in the learning of an L2 word. In the first stage, the L2 word form is learnt along with a pointer to the L1 equivalent. At this stage, there is no link between the L2 word form and the concept. In the second stage, L1 information on syntax and semantics is copied across and becomes part of the L2 lexical entry. At this stage, when the L2 word is used, the L1 form may not be activated, but the

transferred L1 information on syntax and semantics mediates L2 use of the word. In the third stage, L2-specific information on syntax and semantics replaces the L1 information and the L2 form has strong and direct links with the concept. This third stage, however, may not occur or may be incomplete for many words: that is, there may be fossilization even with plenty of rich input. This is because, when a word is at stage two, input often simply reinforces the L1 information since much L1 transfer is appropriate and successful. Restructuring only takes place in special cases: when the learner notices a mismatch and when there is information available about the nature of the difference between the L1 and L2.

Jiang's model has been discussed with reference to collocations by Yamashita and Jiang (2010), Wolter and Gyllstad (2011) and Wolter and Yamashita (2015, 2017). In the two former papers, the discussion was concerned with collocations as units and the focus was on how a collocational unit may pass through the three stages. However, while learners may acquire some collocations as units, learners tend to focus on individual words rather than larger units, a tendency which may be reinforced by common teaching practices that encourage small-item learning and analytical approaches (Wray, 2008). Thus, learners may more often acquire words individually and then subsequently learn the links a word has with others.

In the two latter papers (Wolter & Yamashita, 2015, 2017), then, an alternative way in which Jiang's model may apply to collocations is considered. Taking the view that learning a collocation in many cases means learning how one word links with another, Jiang's model may operate largely as originally described, but with information on collocational links copied across from the L1 along with information on syntax and semantics at stage two. That is, stage 2 of learning an L2 word means linking an L2 form with the L1 equivalent's syntactic, semantic and collocational details[1]. Following thinking in usage-based and exemplar-based models of learning (Abbot-Smith & Tomasello, 2006; Bybee, 2006; Ellis, 2002), once such a link is

[1] It should be noted that in their studies on the processing of collocations Wolter and Yamashita (2015, 2017) did not find evidence to support the idea that collocational information is copied across from the L1. This raises the possibility that collocational details are not routinely or simply copied across. However, a more complex interaction between L1 influences and L2 exposure cannot be ruled out. It may be, for example, that (1) copying across results only in the potential for a link between two words, which, if activated by L2 exposure, is then formed more readily/strongly than would otherwise be the case; (2) the copying across of collocational information only takes place when, due to communicative needs, a learner generates an L2 collocation via the L1 (i.e. the learner is exposed to their own collocation); or (3) collocational information is copied across, but is then gradually lost if there is no L2 exposure to equivalent collocations.

made, a memory trace may be formed in the lexicon for that collocation, which is entrenched with further exposure/use. This collocational representation can be envisioned as branching off from the process of learning the component word itself. The collocational representation can then fossilize at stage two, or with exposure to appropriate input and the noticing of a mismatch, can progress to stage three with full L2-specific knowledge of the collocation's meaning and usage.

Applied in the above way, Jiang's model provides a useful explanation of the challenges faced by learners. A considerable number of collocations are congruent across the L1 and L2, either because the languages reflect reality or because the languages happen to make use of equivalent words (this is particularly the case for languages with a common root). Consequently, much L1 transfer of collocation knowledge is successful. In this way, L1 transfer of collocation knowledge is an enabler for learners throughout much of their learning, quickly giving learners a base of collocation knowledge. However, transfer is not always appropriate, since there are collocational links between words that are specific to the L2 and there may be subtle differences in usage even for collocations that are basically congruent. In these cases, the collocational link must be formed independently of L1 information and possibly in contrast to L1 information. Thus, L1 transfer may become problematic as learners reach higher levels of proficiency and face more challenging and exacting demands on their language. Might this be why learners of moderate proficiency appear to have knowledge of quite a number of collocations while, equally, problems with collocation are often noted among high proficiency learners?

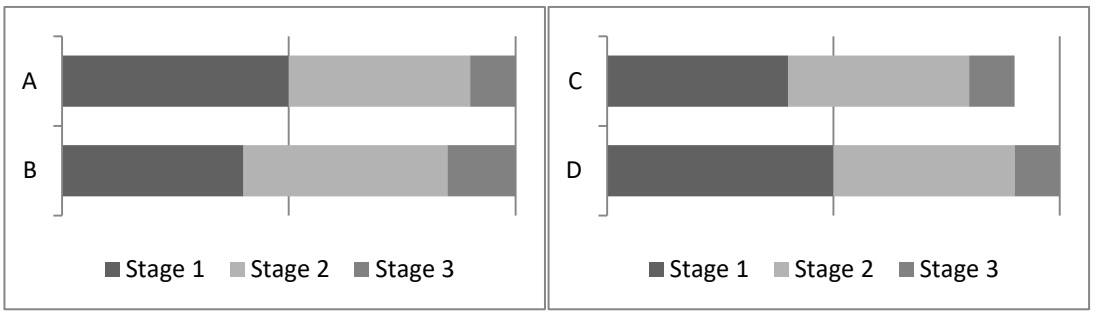
Jiang's model also offers further insights into the finding of no correlation between *LexCombi 2* scores and general proficiency as judged by Yes/No vocabulary size scores (Section 10.3). The key point is that a word at any stage in the process of acquisition will likely be recognised in the Yes/No test, but learners will only be able to supply a canonical collocate for words that have progressed to stage two (if the collocation is congruent) or stage three. This can account for different scenarios in the data.

The data showed, for example, that some learners with similar estimated vocabulary sizes apparently had very different levels of collocation knowledge, while some learners with similar levels of collocation knowledge had very different estimated vocabulary sizes. In the former scenario, all the learners had a similar

number of words at one of the three stages, but some learners had a greater number of words at stage one, and others more words at stages two or three (as illustrated in Figure 11.1, left side). Whichever stage a word was at, it was recognised as a word (and so these learners achieved similar scores on the Yes/No test), but learners were only able to give a canonical collocate for words at stages two or three. In the latter scenario, learners had a different number of words in total across the three stages, and so recognised a different number of words and achieved different scores on the Yes/No test, but had a similar number of words at stages two or three and so produced a similar number of canonical collocates (Figure 11.1, right side). As was discussed earlier, these different scenarios may come about due to the influence of the Japanese education system, the aptitude of learners and learners' interest in learning and using English.

If it is the case that many words and collocates do not reach the third stage in Jiang's model, how does this situation come to pass? As noted above, Jiang says that this stage is reached only when learners notice a mismatch and information on the nature of the difference is available. This may require massive volumes of L2 input for incidental learning and/or sensitive and carefully targeted instruction. With regard to incidental learning of collocations, one oft-noted feature (e.g. Henriksen, 2013) that is believed to make the task difficult is lack of salience. Wible (2008)

Figure 11.1: Possible scenarios based on Jiang's (2000; 2002; 2004) model of lexical development that may account for the lack of correlation between *LexCombi 2* scores and general proficiency.



Note. On the left, Learners A and B have a similar total vocabulary size, but Learner B has more words at stages two and three. On the right, Learners C and D have different total vocabulary sizes, but the same number of words at stages two and three.

argues that for foreign language learners much of the input received is in the form of written text, but, in many cases, the writing system makes orthographic words, rather than chunks, salient. The challenge for learners is to realise that despite the lack of typographical evidence, some parts are best thought of as chunks. The reality of this challenge was demonstrated by Bishop (2004) who found that, during reading, formulaic sequences were looked up by learners significantly less often than single words when neither were highlighted in the text in any way. While these scholars discuss chunks/sequences, the same challenges may well apply to collocations.

On the other hand, there are findings that suggest lack of salience is not a fundamental limitation. Webb, Newton and Chang (2013) showed that incidental learning of collocations through reading can occur, and that frequency of exposure had a clear effect on learning, just as has been found for the incidental learning of single words. Furthermore, Pellicer-Sánchez (2017) directly compared the incidental learning through reading of collocations and of single words and found that, when the number and type of exposures were equal, learning gains for collocations and for single words were similar. This suggests that lack of salience (or any other feature of collocations) does not make the acquisition of collocations inherently more challenging than the learning of single orthographic words.

Another factor behind learners' often incomplete knowledge of collocations may be that, although learners progress towards L1-user norms to an extent, L1-user norms are not the ultimate destination for learners. In Section 10.6, learners' scores under several different scoring approaches were examined, and while relatively little variation was seen (i.e. participants who achieved relatively high or low overall *LexCombi 2* scores did likewise under each scoring approach), some subtle differences between learners were found.

First, it was seen that the effect size for the differences between the collocation knowledge groups was greatest under L2 norms scoring and, while large, not as strong under L1 norms scoring. It was suggested that this is evidence of the direction in which these learners are progressing as their knowledge of collocations develops.

Second, the effect sizes for the two component parts of the corpus-based lists suggested that there were differences in sensitivity to collocation across the three groups of learners. For the representation of English obtained via frequency-based searches for collocations, there was a consistent, strong effect across the three

groups, while for the representation obtained via MI-based searches, the effect was only present as learners moved towards greater levels of collocation knowledge. This finding is of interest since, as mentioned in Section 3.2.3, previous work has found that, in comparison with L1 users, learners/users of English tend to favour frequency-based collocations more and MI-based collocations less (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis, et al., 2008). The findings in Section 10.6 indicate that there may nonetheless be development in learners' sensitivity to MI-based collocations. In this respect, they mirror the findings of Granger and Bestgen (2014) who compared writing by intermediate and advanced learners and found that the advanced essays contained a larger proportion of high MI collocations and a smaller proportion of high *t*-score collocations than the intermediate essays. It may be that learners do become more sensitive to MI-based collocations, but that they do not reach the level of sensitivity shown by L1-user informants.

This brings us to another question raised in Section 10.7: Why do learners seem to make more progress in acquiring some types of collocation than others? More specifically, given that both the above findings suggest that learners do not necessarily progress towards L1-user norms, why might this be?

First, there are arguments questioning the very notion of "L1-user norms". Larsen-Freeman (2006) argues that one assumption in the field of second language acquisition is that acquisition means increasing conformity to a fixed target. The emergentist perspective challenges this assumption by arguing that a learner's interlanguage cannot converge with the target language, because language is dynamic and constantly evolving. There is, then, not a fixed target, but a moving one. This means that the task for learners is to create a linguistic world, rather than conform to one.

Hulstijn's (2015) account of language proficiency is also informative regarding these issues. This describes two types of "language cognition". "Basic language cognition" (BLC) consists of the linguistic knowledge and skills shared by all the L1 users of a particular language, regardless of age, literacy or education, and thus is restricted to spoken language and the everyday. "Higher language cognition" (HLC) includes less frequent items of vocabulary and grammatical structures, written as well as spoken language and a wide range of topics. There is no limit to HLC, and individuals vary greatly in the extent of their HLC according to intelligence,

education and professional and personal interests. Hulstijn argues that BLC cannot be fully acquired by an L2 learner, particularly in the domains of pronunciation and certain grammatical features, but a learner can gain proficiency in HLC to the same extent as an L1 user with the same intellectual, professional and cultural profile. Typically, when researchers tap into "L1-user norms" against which L2 learners/users will be judged, they actually identify a mix of BLC and HLC features. These "norms" can be questioned in two ways: (1) the BLC features can never be fully acquired by L2 learners and so are unreasonable as a yardstick for measurement; and (2) the HLC features are not in fact "norms" since they are not shared by all L1 users of the language.

L1-user norms may, then, be impossible to truly identify, but the same must also be true of L2-user norms. Indeed, it is likely that L2-user language is more dynamic, while attempting to identify linguistic features shared by all L2 users (setting aside the difficulties in identifying this group) would likely result in a rather narrow view of the language. Ultimately, identifying norms may be an impossible task regardless of whose norms are being identified, but it may be possible, and useful, to produce a representation of the language of a community. While complete conformity with this representation is not to be expected of any individual or group, it may still be useful in providing some sense of the ability of individuals/groups.

A second possibility as to why L1-user norms may not be the ultimate destination for learners is that, consciously or otherwise, they choose not to use certain collocations. There are several discussions of such ideas in work on formulaic language. Wray (2008) says language learners may choose not to use formulaic sequences in order to maintain their identity, and that adult learners may perceive transparent, combinatory language as less risky than formulaic expressions. Kecskes (2007) suggests that formulaic sequences are less used in English as a Lingua Franca (ELF) interactions not because speakers lack knowledge of sequences but because they fear that others may lack it. Accordingly, there is a preference for literal language because this provides common ground between speakers (see also Seidlhofer (2009) for a similar argument and Vetchinnikova (2015) who counters that much of ELF users' language is formulaic even while differing somewhat from L1-user conventions). Finally, Wray (2008) discusses the degree of autonomy in the language as another factor that may explain learners' use of formulaic language.

Wray argues that esoteric contexts/societies, which are tightly bonded, require less explicit exchanges and feature a high level of formulaicity. Exoteric contexts/societies, which are large, open and outward looking, require much more explicit, autonomous language exchanges which can be understood by a range of people out of context, and so feature less formulaicity. This suggests that in ELF contexts, perhaps as large, open and outward looking as possible, language will be more autonomous and feature less formulaicity. The language of successful L2 users may, therefore, feature less formulaicity than English in the L1-user world, as several studies have found (e.g. Fan, 2009; Granger, 1998; Howarth, 1998b; Laufer & Waldman, 2011). However, rather than being a sign of deficiency, as it is often interpreted to be, this may in fact be a sign of learners accommodating to the situational realities they operate in.

Learners responding to *LexCombi 2* are not of course engaged in ELF interaction, and so it may be thought that the factors promoting compositional, transparent language explained above are not relevant. However, their language itself may be shaped by these factors, causing their *LexCombi 2* responses to reflect these factors. Similarly, the L2 norms, compiled from the responses of Japanese users of English, may also be reflective of these factors. Thus it may be that these factors, encouraging more compositional, transparent language, acting both on the L2 norms informants and the learners, may explain why the learners in this study appeared to be progressing more towards the L2 norms than the L1 norms.

11.5 The representation of collocations in the lexicon

In Section 4.5.1, it was noted that participants in my trial of *LexCombi 1* occasionally produced responses to *LexCombi 1* cues that were not independent words but word parts. This was also observed in each subsequent set of data from learners. Mostly, these are responses that form compounds in combination with the cue (e.g. *gold* for the cue FISH, *fall* for WATER). However, there are also derivational affixes (e.g. *non* for the cue SENSE, *neuro* for SCIENCE)[2]. The above

[2] It may further be noted that in the think-aloud data (Sections 5.5-5.7) there were two examples (*researcher* for the cue RESEARCH, *supporter* for SUPPORT) of derivations of the cues occurring to participants which were not ultimately given as responses, whether as the derived form or as the derivational affix itself. It may then be that derivations occur as potential responses to participants more regularly than is apparent from the responses themselves. It is not clear at this stage what factors lead participants to reject such responses or to respond with the derivational affix alone.

example word pairs, and those given below, occur in both the COCA and the BNC in the form of two separate words with very low frequency and are far more frequent joined together as single words.

The initial assumption, when responses that form compounds with the cue were encountered, was that the responses reflected limited knowledge on the part of learners regarding whether these compounds are conventionally written as single orthographic words or not. However, as Section 7.6.3 reported, such responses were also frequently seen in the norms data gathered for the purpose of scoring *LexCombi*. Both the British L1 users and the Japanese users of English who acted as informants for the two sets of norms lists (see Section 7.2) gave such responses, and indeed gave many of the same responses as each other. As with the learners, the majority of these responses form compounds in combination with the cue (e.g. *basket* for the cue BALL, *case* for BOOK, *sight* for EYE), but there were also a number of derivational affixes (e.g. *less* for the cue AGE, *ship* for LEADER, *ful* for POWER)[3].

It is perhaps notable that neither the learners nor the norming informants provided inflectional affixes as responses. This may reflect the fundamental division between inflection on the one hand and derivation and compounding on the other; the former creating grammatical variants of a word, the latter creating new words (Marslen-Wilson, 2007). The one apparent exception to this is that in the data collected with Barfield's original set of cues, the cue INTEREST did on occasion illicit the response *interested* from learners. It would appear in this case that an inflectional affix was added to the cue to create a response. However, it might be argued that though this response involves an inflectional affix, it functions somewhat like a derivational affix in that it creates a different part of speech, the form *interested* typically functioning as an adjective rather than a verb.

On occasion, then, Japanese learners of English, advanced L2 users of English and L1 users all give responses that form multi-morphemic words with the cues[4].

[3] There was, in addition, one case of a response which combines with the cue to form a word unrelated to the cue: the response *re* for the cue MEMBER. While bearing in mind that items on the norms lists were given by at least two informants, the fact that there is only this single instance of such a response means it can perhaps be considered idiosyncratic.

[4] The tendency for such responses to be given does not seem to have been noted in previous studies on collocations. This is probably because many studies have not collected data in a form that would bring such a tendency to light. Most corpus-based studies of collocation (see Section 3.2.1) and elicitation studies (see Section 3.2.2) begin with a notion of collocation that involves orthographic words rather than morphemes. Consequently, they do not conduct searches or select items that have

What is the significance of these responses?

Section 7.6.3 suggested that they might in part reflect the *LexCombi* task. It may be thought, for example, that the adapted *LexCombi* format, in which the cue word appears with gaps to the left and right in which a response should be placed, leads to such responses (even though the gaps for responses are separated from the cue by a space). Yet, in the case of learners at least, this type of response was also noted in data collected with the original *LexCombi* format (Section 4.5.1). Another possibility is that such responses occurred because respondents were struggling to give the full quota of three responses for each cue: that is, these responses were given when respondents could produce no further "conventional" responses. If this were the case, these responses might be expected to more often occur as the final response given by a participant. However, examining the original test booklets of learners and of the norming informants, there was no such tendency. Indeed, responses that form multi-morphemic words with the cues were given alongside and intermingled with "conventional" responses. The impression is that the various respondents who completed the *LexCombi* task made no distinction between the two types of responses. It might be asked, therefore, whether, in terms of their representation in the lexicon, there are parallels between collocations and multi-morphemic words.

In studies of collocation, compounds conventionally written as two words may or may not be considered collocations. Granger and Paquot (2008) state that in the phraseological tradition such items are not usually considered collocations, whereas in frequency-based work those that meet the frequency criteria are routinely included. Compounds conventionally written as a single orthographic word, however, as well as derivations, are rarely considered collocations. One exception is Van der Wouden (1997) who talks of "collocations below the word level" (p. 20). These are words that include elements (affixes or other morphemes) that occur in a limited number of words: for example, the *cran* in *cranberry* or the plural affix *-en* in *oxen*. In line with the phraseological tradition, Van der Wouden considers these elements collocations because their distribution is restricted and idiosyncratic.

---

the potential to reveal this phenomenon. Only with more open tasks, such as that used by Nordquist (2009) described below or the classic word association task, would this phenomenon be potentially observable, and, while Nordquist does not seem to have observed it, in word association data, such as the *Edinburgh Associative Thesaurus* (Kiss, et al., 1973), some instances from among the above examples can be found, such as the response *basket* for the cue BALL, *case* for BOOK and *sight* for EYE.

Outside of work on collocation, derivation and compounding are both major areas of study. On derivations, there has been much debate about their processing and representation. Summarising the research, Marslen-Wilson (2007) states that there is now strong evidence that "all potentially morphologically complex words undergo an initial obligatory process of segmentation into their morphemic components, irrespective of whether the words actually are morphologically complex" (p. 184). However, despite this segmentation process, it appears that derivations are stored holistically in the lexicon, though there are differences in how they are stored depending on their semantic transparency. Derivations which are transparent and feature productive affixes are stored in a form which makes their structure clear, while derivations which are opaque are stored without a decompositional morphemic structure.

Similar debates about computation and storage have taken place with respect to compounds. Compounds are considered to be "at the crossroads of storage and computation" (Libben, 2005, p. 269): computation, it is thought, must occur because compounding is a productive word-formation process, while storage must occur because most compounds cannot be understood solely through their constituent parts. Several models of compound processing have been developed and while these models disagree about whether constituent parts are first accessed before the whole is processed or whether the whole is processed before constituent parts are activated, there does appear to be agreement that constituent parts are activated at some point (Li, Jiang, & Gor, 2017).

If compounds are accepted as one form of formulaic language, the idea that their component parts are activated at some point seems to challenge Wray's claims regarding formulaic language. Wray (2002) argued that formulaic sequences are "stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p. 9) and later developed the idea of the morpheme-equivalent unit: "a word or word string . . . that is processed like a morpheme, that is, without recourse to any form-meaning matching of any sub-parts it may have" (2008, p. 12).

Several studies have suggested that evidence of faster processing of formulaic sequences supports the idea of holistic storage (Jiang & Nekrasova, 2007; Tremblay, et al., 2011). Stronger support for Wray's position can be seen in Sosa and

Macfarlane's (2002) study, which investigated the response times of participants to the word *of* as it occurred in combinations of varying frequency levels. Participants were found to have slower reaction times when *of* occurred within high-frequency collocations than when it occurred in less frequent phrases, which was seen as indicating holistic storage of the collocations such that recognition of *of* was somewhat delayed. Kapatsinski and Radicke (2009) pointed out that Sosa and Macfarlane's results could have been explained by articulatory reduction of *of* in highly frequent combinations. They therefore conducted a similar study, but focused on the particle *up* since it does not undergo much articulatory reduction, and similarly found slower reaction times to *up* occurring within high-frequency combinations.

Wray (2008) has also addressed previous suggestions, with respect to idioms, that formulaic sequences are processed compositionally. She argues that the research provides evidence "only that idioms *can* be compositional" (p. 30), and describes experiments as encouraging "artificial engagements with the form" in that "if you ask people to look at or listen to idioms out of a communicative context and they know they are going to be tested on what they see or hear, it is hardly surprising that they attend to them in a different way from their approach in, say, a conversation or reading a book" (p. 32).

Siyanova-Chanturia (2015), however, sees evidence of the constituents of compounds being activated in processing as one line of evidence against Wray's (2002) conception of holistically stored, unanalysed formulaic sequences. She points to a range of evidence showing that the component parts of various types of formulaic sequence are activated in processing. In addition to findings regarding compounds, she notes: (1) Arnon and Cohen Priva's (2013) investigation of the duration of words in speech, which found that word duration within three-word phrases was affected by word frequency as well as by phrasal frequency; (2) Konopka and Bock's (2009) and Snider and Arnon's (2012) studies of idiom priming, which show that idioms prime their syntactic structures; (3) Sprenger, Levelt and Kempen's (2006) study, which shows that idioms are primed by their constituent words; and (4) Molinaro, Canal, Vespignani, Pesciarelli and Cacciari's (2013) study of complex prepositions in Italian (e.g. *nelle mani dei* (= *in the hands of*)), which argues that the individual words in such prepositions retain their semantic

and syntactic properties.

Siyanova-Chanturia notes the findings of Sosa and Macfarlane (2002) and Kapatsinski and Radicke (2009), but points to methodological problems with both studies and the fact they are limited to considering a single form. Siyanova-Chanturia also notes Wray's (2008) criticisms of experimental research, but counters that there is evidence supporting her argument from well-designed studies in which participants have no knowledge of the study's purpose and in which plenty of distractors are used so that the target sequences are not noticeable, as well as from studies that use sensitive methodologies such as eye-tracking and event-related brain potentials (ERPs). Siyanova-Chanturia concludes that while there is clear evidence for fast processing of formulaic sequences, this does not necessarily mean that sequences are holistically stored, and that in fact the evidence overall does not suggest they are stored as unanalysed wholes.

While mentioned several times by Siyanova-Chanturia as one type of formulaic sequence, evidence from studies of collocations is not considered in her paper. Yet this evidence is clearly relevant. First, though not studies of online processing, two elicitation studies provide interesting findings. McGee (2009) asked participants to give what they considered to be the most frequent noun collocate for 20 common adjectives, and compared these responses with BNC data. McGee found that frequent collocations which typically occur within a larger unit were often not produced by participants. For example, many of the most frequent noun collocates in the BNC for the adjective *important* which were not produced by his participants were those that often occur in the frame *DET ADJ NOUN of NP*, such as the noun *aspect*, which often occurs with *important* in the frame *an important aspect of NP*. McGee suggests collocates occurring within such longer frames are to a degree hidden from memory searches.

In Nordquist's (2009) elicitation study participants were asked to produce one utterance each in response to 12 cue words. She found that while the utterances produced reflected participants' knowledge of collocations and the language was produced mostly in accordance with the idiom principle, the most frequent collocations for the 12 cues did not generally occur. Nordquist suggests, following Bybee's (2002) linear fusion hypothesis, that this is because these most frequent collocations are autonomous from their component parts.

Evidence from research on collocation processing also points to collocations being largely autonomous from their constituent words. Wolter and Gyllstad (2013), investigating reaction times to adjective + noun collocations, found effects for collocation frequency and length, but not for the frequency of the component adjectives or nouns. Sonbul (2015), in an eye-tracking study with adjective + noun collocations, found effects for collocation frequency and for noun frequency on first pass reading times, but not for adjective frequency. Gyllstad and Wolter (2016), looking at reaction times to verb + noun collocations, found effects for collocation frequency but not for the frequency of the verbs or of the nouns. Finally, Wolter and Yamashita (2017), in a reaction time study featuring adjective + noun collocations, found effects for adjective frequency, noun frequency and collocation frequency, albeit that the last of these was the strongest. In two studies, therefore, an effect was found for the frequency of component words, but all four studies found effects for collocation frequency. The signs of component words having an effect on processing are then rather varied in the case of collocations.

The signals, then, seem to be mixed regarding holistic storage and the processing of formulaic sequences, with evidence indicating that the component parts of sequences are activated in processing and evidence that sequences are to some extent autonomous from their internal components. What could account for this?

One possibility is that the relationship between sequences and their component parts is similar to that between words and the letters that form them. We appear to have implicit knowledge of the frequencies of English letters and of letter sequences (Miller, Bruner, & Postman, 1954); yet letters in themselves are not meaningful. Presumably, then, as we process a word (i.e. a sequence of letters) as a meaningful unit, there is at the same time some activation of the component parts (i.e. the individual letters and letter combinations) even though these elements do not contribute units of meaning that are combined to produce the meaning of the whole. Perhaps, in processing formulaic sequences, the component parts are similarly activated while not becoming contributors to meaning, meaning being processed for the unit as a whole. This account of the process would seem to explain the evidence for activation of internal components, while also according with Wray's (2008) views in that in defining the morpheme-equivalent unit she talks of there being no "form-meaning matching of any sub-parts" (p. 12).

A second possibility is that the processing of formulaic sequences has parallels with the processing of derivations. As outlined above, it is thought that any word which is potentially morphologically complex initially undergoes an obligatory process of segmentation which identifies its morphemic structure. This structure is not, however, subsequently analysed to produce meaning, for derived words are stored in the lexicon and meaning is listed in the entry, this being the case whether the derivation is semantically transparent or opaque (Marslen-Wilson, 2007). In the processing of derivations, there is, then, a separation between the activation of components and access to meaning. This would appear to match the evidence regarding formulaic sequences. As the findings reviewed by Siyanova-Chanturia showed, there seems to be activation of the components of formulaic sequences, but, paralleling the representation of derivations, there may nonetheless be holistic storage of the formulaic sequences which comprises a form-meaning link.

Both the above explanations suggest that while there is activation of the component parts of sequences with respect to their recognition, the processing of meaning is not componential but holistic. Interestingly, among the evidence cited by Siyanova-Chanturia against the idea of holistic storage, semantic access is hardly mentioned. Only the study by Molinaro et al. (2013) makes claims involving semantics. This study, conducted in Italian with Italian L1 participants, compared the processing of complex prepositions such as *nelle mani dei* (*in the hands of*) with modified forms such as *nelle affidabili mani dei* (*in the capable hands of*). It was found that the modified forms did not lose their functional role, and on this basis it was suggested that the constituent words in these forms "still maintain their semantic and syntactic properties" (p. 783). However, in English modifications of the type included in the study are far from unusual (e.g. the phrase *in contrast to* appears in the COCA 4,706 times, while *in stark contrast to* has 395 occurrences and *in sharp contrast to* 340), and the same may be true of Italian given that prior to the experiment all the sentences were rated for naturalness and those with modifications were rated just as highly as those without. Wray (2008) notes, with respect to idioms, that the creation of alternatives does involve contemplation of an idiom's parts, but argues that once alternatives are established, a frame may exist which allows variation in a particular slot. Given the naturalness ratings for the sentences including the modified phrases in Molinaro et al.'s study, it seems likely that this

process has occurred. It is not necessarily the case therefore that these sequences are processed componentially for meaning.

Also interesting with regard to the idea of a division between recognition and access to meaning are the collocation processing studies mentioned above. It will be recalled that in Sonbul's (2015) eye-tracking study and Wolter and Yamashita's (2017) study of acceptability judgements, there were indications of an effect for the frequency of the component words in collocations, while in the Wolter and Gyllstad (2013) and Gyllstad and Wolter (2016) lexical decision studies no effects for the frequency of component words were seen. Interestingly, Sonbul's finding was with regard to first pass reading times, which are believed to be a measure of initial processing. Her finding may therefore be about recognition, rather than access to meaning. With regard to lexical decision, as used by Wolter and Gyllstad and Gyllstad and Wolter, there is some debate about the issue, but it is argued that the task involves semantic access (Dilkina, McClelland, & Plaut, 2010). The pattern of findings in these three studies also then fit with the above suggestions.

Wolter and Yamashita's (2017) study, however, does not fit into this picture. The acceptability judgement task their study featured might be expected to involve access to meaning, and yet effects were found for the frequency of the component words as well as for the frequency of the collocations. One explanation for this might be that judging acceptability requires participants both to recognise the forms and access meaning, and so it is possible that the frequency of the words had a role in speeding up recognition of the form of the collocation and that subsequently the frequency of the collocation affected access to meaning. If this were the case, however, similar findings of both word frequency and collocation frequency having effects should also have been observed in Wolter and Gyllstad (2013) and Gyllstad and Wolter (2016). It is not clear at this stage how the contrasting findings of these studies can be reconciled.

This section began with the observation that L2 learners, L2 users and L1 users all provided some responses that form multi-morphemic words in combination with *LexCombi* cues. The impression gained from the data was that these informants treated collocations, compounds and derivations as equivalent. The above account has suggested that there may also be equivalences in their processing. In particular, it suggests that the recognition of collocations (and formulaic sequences) may involve

the activation of component words, but that there is holistic storage of collocations in the lexicon and access to meaning is not componential but holistic.

Chapter 12 Conclusions

At the start of this thesis, it was observed that while collocation is often characterised as an area of considerable difficulty for L2 learners, learners also make use of collocations in their speech and writing from quite an early stage. A great deal of the existing research on L2 learners' knowledge of collocations is, however, focused on learners of advanced proficiency, and previous efforts to investigate learners' knowledge of collocations have suffered from a number of other limitations. The primary concern of this thesis has, therefore, been the development of a tool for eliciting L2 learners' productive knowledge of collocations.

Beginning with an instrument designed by Barfield (2009) named *LexCombi*, which was considered to have potential, a series of studies were conducted to trial and develop the instrument further. This involved adapting the instrument's format and undertaking a rigorous process of cue selection to ensure the elicitation of more valid data, along with considering how the concept of collocation may best be operationalised to allow the scoring of learners' responses. This sustained, iterative development of the instrument is one important contribution of this thesis.

The result is a version, *LexCombi 2*, which is suitable for use with a wide range of learners, and which has the potential for use in both longitudinal and cross-sectional studies. Chapter 10 reported an initial investigation of data collected with *LexCombi 2*, the instrument in its final form. From this investigation and the discussion of findings reported across this thesis, a number of conclusions may be drawn, as outlined below. As with any empirical work, there are limitations to these findings, and these are also discussed.

The first conclusion is that the relationship between productive knowledge of collocations and general L2 proficiency is unclear. This thesis found no relationship between these two variables, and the reliability of this finding is supported by the fact that it was seen in two separate data sets (once a cue set was used which was not selected in a biased manner; see Section 10.3). However, this finding may reflect limitations in the range of learners it was possible to collect data from or in the proficiency measure which was used.

With regard to the former, while data was collected from groups of learners displaying quite a wide range in proficiency, it would be interesting to see if the

above finding is maintained were data collected from both learners of less and of more advanced proficiency. Learners of lower proficiency are of interest since the learners included in this study were seen to have productive knowledge of a reasonable number of collocations and this knowledge must have developed in some way. Data from less proficient learners would presumably show the development of this knowledge. In like manner, data from learners of more advanced proficiency might reveal whether or to what extent the lack of correlation with productive knowledge of colocations continues. Also interesting would be a longitudinal study to see if productive knowledge of collocations progresses as learners develop their general proficiency, and at what point progress in their knowledge of collocations takes place. Section 10.3.2 meanwhile suggested that the Japanese education system, with its emphasis on high-stakes examinations focused on receptive skills, may also in part explain the lack of correlation found. Exploring the relationship among learners from a different L1 background with an education system that has a more productive orientation to English might then have a different outcome.

The final point above is also related to the second issue: the proficiency measure which was used. A Yes/No test, while highly practical and a good proxy for general proficiency, is nonetheless a measure of passive recognition of vocabulary, yet, as Section 11.3 pointed out, there are good reasons to think that productive knowledge of collocations should, in particular, correlate with productive language skills. Investigating data obtained with a measure of productive proficiency or indeed a general measure which provides a breakdown by different skills, such as the IELTS test, would therefore be of interest. There are then a number of potentially fruitful avenues for further investigation in this area.

Second, it is concluded that learners seem to develop principally towards the type of collocational competence displayed by advanced L2 users of English with the same L1 background rather than towards L1-user competence. This conclusion was reached since, although there were significant differences between groups of learners in terms of their *LexCombi 2* responses scored both against the L1-user norms and against the L2-user norms, the effect size was larger in the latter case (Section 10.6). A caveat to this finding, however, is that the L1-user norms, for practical reasons, were based on the responses of British informants, whereas English education in Japan is primarily oriented towards American English. It may be that the learners in

this study are progressing more towards a US L1-user norm than a British L1-user norm. However, while there certainly are differences in collocation between British and American English, these differences are relatively slight and it seems unlikely that they could have had a major impact on this finding.

Assuming, then, that learners do develop most strongly towards the collocational competence of advanced L2 users, this may be due to the nature of L1-user competence, to choices made by learners and/or a consequence of the realities of English as a Lingua Franca communication, as discussed in Section 11.4. As with the first conclusion, it would be interesting to explore whether this conclusion holds with learners from other L1 backgrounds. This might enable greater insight into the relative importance of the three factors above. That is, were a number of investigations carried out with English learners of different L1 backgrounds and the same finding made, it would suggest that the more general factors – the nature of L1-user competence and the realities of English as a Lingua Franca communication – are perhaps of more consequence than the more specific factor of learners' choices. Whichever factor or combination of factors is responsible, future studies of L2 collocation should carefully consider how best to determine the standard against which learners are measured and recognise that L1-user knowledge may not be the desired or actual end-point of learning.

A third conclusion is that learners and users of English seem to treat collocations as akin to multi-morphemic words. On the one hand, this conclusion is limited in that the evidence for it was indirect and rather dependent on inference. On the other hand, similar behaviour, in terms of the types of responses which were given and how these were interspersed, was observed in the data from learners, from L2 users of English and from L1 users of English, meaning it is more likely to be a general phenomenon.

If there are parallels between multi-morphemic words and collocations, there may be much to learn from explorations and models of the acquisition and processing of such words, about which there is a rich literature. After first seeking confirmation of the parallels between multi-morphemic words and collocations suggested by this thesis, and assuming that such confirmation is found, it may be possible to adopt or adapt research methods used in L1 research on multi-morphemic words and take forward the investigation of the acquisition and processing of

collocations by L2 learners. Similarly, it may be possible to use models developed in L1 research in this area to guide and direct investigations of L2 learners' acquisition and processing of collocations.

There remains much to be learnt about the development of L2 learners' productive knowledge of collocations. The work reported in this thesis provides an instrument that may facilitate future research and some insights that it is hoped can stimulate further enquiry.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23*(3), 275-290.

Adolphs, S., & Durrow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 107-126). Amsterdam: John Benjamins.

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*(3), 349-371. doi: 10.1177/0023830913484891

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67-82. doi: 10.1016/j.jml.2009.09.005

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal, 47*(1), 56-63. doi: 10.1093/elt/47.1.56

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101-114. doi: 10.1016/0346-251x(93)90010-e

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19*(3), 241-248. doi: 10.1111/j.1467-9280.2008.02075.x

Barfield, A. (2009a). Exploring productive L2 collocation knowledge. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 95-110). Bristol: Multilingual Matters.

Barfield, A. (2009b). Following individuals' L2 collocation development over time. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 208-223). Basingstoke: Palgrave Macmillan.

Barfield, A., & Gyllstad, H. (2009). Introduction: Researching L2 collocation knowledge and development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 1-18). Basingstoke: Palgrave Macmillan.

Barnbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). *Collocation: Applications and implications*. Basingstoke: Palgrave Macmillan.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing, 18*(3), 235-274. doi: 10.1177/026553220101800301

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118. doi: 10.1177/0265532209340194

Bell, H. (2009). The messy little details: A longitudinal case study of the emerging lexicon. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 111-127). Bristol: Multilingual Matters.

Benson, M., Benson, E., & Ilson, R. (1986). *Lexicographic description of English*. Amsterdam: John Benjamins.

Benson, M., Benson, E., & Ilson, R. (Eds.). (1997). *The BBI dictionary of English word combinations* (Revised ed.). Amsterdam: John Benjamins.

Benson, M., Benson, E., & Ilson, R. (Eds.). (2009). *The BBI combinatory dictionary of English* (Third ed.). Amsterdam: John Benjamins Publishing Company.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System, 69*, 65-78. doi: 10.1016/j.system.2017.08.004

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26*, 28-41. doi: 10.1016/j.jslw.2014.09.004

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson*

(pp. 181-190). Amsterdam: Rodopi.
Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. doi: 10.1093/applin/25.3.371
Bishop, H. (2004). The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 227-248). Amsterdam: John Benjamins Publishing Company.
Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research, 10*(3), 245-261. doi: 10.1191/1362168806lr195oa
Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Basingstoke: Palgrave Macmillan.
Boers, F., & Webb, S. (2015). Gauging the semantic transparency of idioms: Do natives and learners see eye to eye? In R. Heredia & A. Cieślicka (Eds.), *Bilingual figurative language processing* (pp. 368-392). Cambridge: Cambridge University Press.
Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (Second ed.). New York: Routledge.
Bonk, W. J. (2000). Testing ESL learners' knowledge of collocations. Educational Resources Information Center.
Booth, P. (2013). Vocabulary knowledge in relation to memory and analysis: An approximate replication of Milton's (2007) study on lexical profiles and learning style. *Language Teaching, 46*(3), 335-354. doi: 10.1017/S0261444813000049
Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 109-132). Amsterdam: John Benjamins.
Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711-733.
Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The Companion to language assessment* (pp. 1079-1097). Hoboken, NJ: John Wiley & Sons, Inc.
Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics, 30*(2), 236-252. doi: 10.1093/applin/amn039
Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science, 9*(3), 542-551. doi: 10.1111/tops.12274
Churchill, E. (2008). A dynamic systems account of learning a word: From ecology to form relations. *Applied Linguistics, 29*(3), 339-358. doi: 10.1093/applin/amm019
Clear, J. (1993). From Firth principles: Computational tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 271-292). Amsterdam: John Benjamins.
Cogo, A., & Dewey, M. (2012). *Analysing English as a Lingua Franca: A corpus-driven investigation*. New York: Continuum.
Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61. doi: 10.1017/S0267190512000074
Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal, 52*(1), 57-63. doi: 10.1093/elt/52.1.57
Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185-209. doi: 10.2307/3587717
Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 126-139). Harlow: Longman.
Cowie, A. P. (1998a). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 1-20). Oxford: Oxford University Press.
Cowie, A. P. (1998b). A.S. Hornby, 1898–1998: A centenary tribute. *International Journal*

of Lexicography, 11(4), 251-263. doi: 10.1093/ijl/11.4.251
Crowther, J., Dignen, S., & Lea, D. (Eds.). (2002). *Oxford Collocations Dictionary*. Oxford: Oxford University Press.
Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.
Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. http://www.americancorpus.org
Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. London: Routledge.
Davis, H. G. (2001). *Words: An integrational approach*. Richmond: Curzon.
Dewaele, J-M. (2018). Why the dichotomy 'L1 versus LX user' is better than 'native versus non-native speaker'. *Applied Linguistics*, *39*(2), 236–240. doi: 10.1093/applin/amw055
Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research, 1365*, 66-81. doi: 10.1016/j.brainres.2010.09.057
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. doi: 10.1016/j.esp.2009.02.002
Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics, 19*(4), 443-477. doi: 10.1075/ijcl.19.4.01dur
Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, 6*(2), 125-155. doi: 10.1515/cllt.2010.006
Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*(2), 157-177. doi: 10.1515/iral.2009.007
Educational Testing Service. (n.d.). TOEIC Listening and Reading Test Scores and the CEFR levels Retrieved May 31, 2017, from https://www.etsglobal.org/Global/Eng/Research/CEFR
Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143-188. doi: 10.1017/s0272263102002024
Ellis, N. C., Frey, E., & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In U. Römer & R. Schulze (Eds.), *Exploring the lexis-grammar interface* (pp. 89-114). Amsterdam: John Benjamins.
Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61-78. doi: 10.1515/cllt.2009.003
Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375-396. doi: 10.1002/j.1545-7249.2008.tb00137.x
Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, Massachusetts: The MIT Press.
Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Færch & G. Kasper (Eds.), *Introspection in second language research* (pp. 24-53). Clevedon: Multilingual Matters.
Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1212-1248). Berlin: Mouton de Gruyter.
Eyckmans, J. (2009). Toward an assessment of learners' receptive and productive syntagmatic knowledge. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language* (pp. 139-152). Basingstoke: Palgrave Macmillan.
Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in Yes/No vocabulary tests. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 59-76). Cambridge:

Cambridge University Press.
Fan, M. (2009). An exploratory study of collocational use by ESL students – A task based approach. *System, 37*(1), 110-123. doi: 10.1016/j.system.2008.06.004
Field, A. (2009). *Discovering statistics using SPSS* (Third ed.). London: Sage Publications.
Firth, J. R. (1951). Modes of meaning. *Essays and studies*, 118-149. (Reprinted in J.R. Firth (1957). *Papers in linguistics: 1934-1951* (pp. 1190-1215). London: Oxford University Press.).
Firth, J. R. (1952/3). Linguistic analysis as a study of meaning. (Printed in F.R. Palmer (Ed.) (1968). *Selected papers of J.R. Firth, 1952-1959* (pp. 1912-1926). Harlow: Longman.).
Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 168-205. (Reprinted in F.R. Palmer (Ed.) (1968). *Selected papers of J.R. Firth, 1952-1959* (pp. 1168-1205). Harlow: Longman.).
Fitzpatrick, T. (2007). Word Association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319-331. doi: 10.1111/j.1473-4192.2007.00172.x
Fitzpatrick, T. (2009). Word association profiles in a first and second language: Puzzles and problems. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 38-52). Bristol: Multilingual Matters.
Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. J. (2015). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics, 36*(1), 23-50. doi: 10.1093/applin/amt020
Forsberg, F., & Fant, L. (2010). Idiomatically speaking: Effects of task variation on formulaic language in highly proficient users of L2 French and Spanish. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 47-70). London: Continuum.
Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*, 155-179. doi: 10.1111/lang.12225
Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *International Journal of Applied Linguistics, 166*(1), 94-126. doi: 10.1075/itl.166.1.03fer
Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 145-159). Oxford: Oxford University Press.
Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3), 229-252.
Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27-49). Amsterdam: John Benjamins Publishing Company.
Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 21-33). Basingstoke: Palgrave Macmillan.
Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language, 36*(1), 3-15. doi: 10.1016/0093-934X(89)90048-5
Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. (Doctoral dissertation), Lund University.
Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning, 66*(2), 296-323. doi: 10.1111/lang.12143
Handl, S., & Graf, E. (2010). Collocation, anchoring, and the mental lexicon – an ontogenetic perspective. In H.-J. Schmid & S. Handl (Eds.), *Cognitive foundations of linguistic usage patterns* (pp. 119-147). Berlin: De Gruyter Mouton.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, Massachusetts: The MIT Press.
HarperCollins. (2004). Collins Wordbanks Online, from http://www.collins.co.uk/corpus/corpussearch.aspx
Hausmann, F. J. (1991). Collocations in monolingual and bilingual English dictionaries. In V. Ivir & D. Kalogjera (Eds.), *Languages in contact and contrast: Essays in contact linguistics* (pp. 225-236). Berlin: De Gruyter.
Hausmann, F. J. (1999). Semiotaxis and learners' dictionaries. In T. Herbst & K. Popp (Eds.) *The perfect learners' dictionary(?)* (pp. 205-211). Tübingen: Max Niemeyer Verlag.
Henriksen, B. (2013). Research on L2 learners' collocational competence and development – a progress report. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*. EUROSLA.
Hernández, M., Costa, A., & Arnon, I. (2016). More than words: Multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience, 31*(6), 785-800. doi: 10.1080/23273798.2016.1152389
Hill, J., & Morgan, L. (Eds.). (1997). *LTP Dictionary of Selected Collocations*. Hove: LTP.
Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
Howarth, P. (1998a). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. doi: 10.1093/applin/19.1.24
Howarth, P. (1998b). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 161-186). Oxford: Oxford University Press.
Hsu, J.-y., & Chiu, C. Y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL Journal, 10*(1), 181-204.
Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing, 19*(3), 227-245. doi: 10.1191/0265532202lt229oa
Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers*. Amsterdam: John Benjamins.
Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
Hunston, S., & Francis, G. (2000). *Pattern grammar*. Amsterdam: John Benjamins.
Imao, Y., & Brown, D. (2014). CollCheck. Osaka: Osaka University.
Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., & Tono, Y. (2003). *JACET8000: JACET list of 8000 basic words* Tokyo: JACET.
Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*(1), 47-77. doi: 10.1093/applin/21.1.47
Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 24*(4), 617-637. doi: 10.1017/S0272263102004047
Jiang, N. (2004). Semantic transfer and development in adult L2 vocabulary acquisition. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 101-126). Amsterdam: John Benjamins.
Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*(3), 433-445. doi: 10.1111/j.1540-4781.2007.00589.x
Jones, S., & Sinclair, J. (1974). English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie, XXXIV*(I), 15-61. (Reprinted in W. Teubert & R. Krishnamurthy (Eds.) (2007). *Corpus linguistics* (pp. 2223-2269). London: Routledge.).
Kapatsinski, V., & Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (Eds.),

Formulaic language (pp. 499-522). Amsterdam: John Benjamins.
Kecskes, I. (2007). Formulaic language in English Lingua Franca. In I. Kecskes & L. R. Horn (Eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (pp. 191-218). Berlin: Mouton de Gruyter.
Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly, 37*(3), 467-487. doi: 10.2307/3588400
Kennedy, G. (2005). *Collocational patterning with high frequency verbs in the British National Corpus.* Paper presented at the American Association of Applied Corpus Linguistics Conference, University of Michigan, Ann Arbor.
Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263-276. doi: 10.1515/cllt.2005.1.2.263
Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153-165). Edinburgh: University Press.
Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Clarendon Press.
Koizumi, R. (2015). Second language vocabulary assessment studies: Validity evidence and future directions. *Vocabulary Learning and Instruction, 4*(1), 36-46. doi: 10.7820/vli.v04.1.koizumi
Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive psychology, 58*(1), 68-101. doi: 10.1016/j.cogpsych.2008.05.002
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786. doi: 10.1002/tesq.194
Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-619. doi: 10.1093/applin/aml029
Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647-672. doi: 10.1111/j.1467-9922.2010.00621.x
Lenko-Szymanska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*(2), 225-251. doi: 10.1075/ijcl.19.2.04len
Lewis, M. (1993). *The lexical approach*. Hove: Language Teaching Publications.
Lewis, M. (1997). *Implementing the lexical approach*. Hove: Language Teaching Publications.
Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.
Li, M. A. N., Jiang, N., & Gor, K. (2017). L1 and L2 processing of compound words: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition, 20*(2), 384-402. doi: 10.1017/s1366728915000681
Libben, G. (2005). Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *The Canadian Journal of Linguistics/La revue canadienne de linguistique, 50*, 267-283. doi: 10.1017/s000841310000373x
Libben, G. (2014). The nature of compounds: A psychocentric perspective. *Cognitive Neuropsychology, 31*(1-2), 8-25. doi: 10.1080/02643294.2013.874994
Linacre, J. M. (2014). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.
Linacre, J. M. (2015). Winsteps® Rasch measurement computer program (Version 3.9). Beaverton, Oregon: Winsteps.com.
Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Basingstoke: Palgrave Macmillan.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT press.
Marslen-Wilson, W. (2007). Morphological processes in language comprehension. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 175-193). Oxford: Oxford University Press.
Mason, O. (1997). *The weight of words: An investigation of lexical gravity.* Paper presented at the Proceedings of PALC '97.
Mauranen, A. (2011). Learners and users – Who do we want corpus data from? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 155-171). Amsterdam: John Benjamins.
McCarthy, M., & O'Dell, F. (2005). *English collocations in use*. Cambridge: Cambridge University Press.
McEnery, T., & Hardie, A. (2012). *Corpus linguistics*. Cambridge: Cambridge University Press.
McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory, 5*(1), 79-103. doi: 10.1515/cllt.2009.004
McIntosh, C. (Ed.) (2009). *Oxford Collocations Dictionary* (Second ed.). Oxford: Oxford University Press.
McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
Meara, P. (1983). Word associations in a foreign language: A report on the Birkbeck Vocabulary Project. *Nottingham Linguistics Circular, 11*, 29-38.
Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge University Press: Cambridge.
Meara, P. (2004). Modelling vocabulary loss. *Applied Linguistics, 25*(2), 137-155. doi: 10.1093/applin/25.2.137
Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics, 27*(4), 620-644. doi: 10.1093/applin/aml030
Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5-19.
Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142-154. doi: 10.1177/026553228700400202
Meara, P., & Fitzpatrick, T. (2000). *Lex30*: An improved method of assessing productive vocabulary in L2. *System, 28*(1), 19-30. doi: 10.1016/s0346-251x(99)00058-5
Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society: Papers from the Annual Meeting of the British Association for Applied Linguistics* (pp. 80-87). London: British Association for Applied Linguistics.
Meara, P., & Milton, J. (2003). X_Lex, The Swansea Levels Test. Newbury: Express.
Meara, P., & Miralpeix, I. (2006). Y_Lex: The Swansea Advanced Vocabulary Test. v2.05. Swansea: Lognostics.
Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. *Angles on the English-speaking world, 4*, 85-96.
Michelbacher, L., Evert, S., & Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory, 7*(2), 245-276. doi: 10.1515/cllt.2011.012
Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics, 32*(2), 129-148. doi: 10.1093/applin/amq035
Miller, G. A., Bruner, J. S., & Postman, L. (1954). Familarity of letter sequences and tachistoscopic identification. *The Journal of General Psychology, 50*, 129-139.
Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 47-58). Cambridge: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse & M. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83-97). Bristol: Multilingual Matters.
Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73-98. doi: 10.1191/0265532206lt321oa
Molinaro, N., Canal, P., Vespignani, F., Pesciarelli, F., & Cacciari, C. (2013). Are complex function words processed as semantically empty strings? A reading time and ERP study of collocational complex prepositions. *Language and Cognitive Processes, 28*(6), 762-788. doi: 10.1080/01690965.2012.665465
Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory, 5*(2), 175-200. doi: 10.1515/cllt.2009.008
Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12-25.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.
Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). Amsterdam: John Benjamins.
Nation, I. S. P. (2006a). BNC-based word lists. Wellington: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation
Nation, I. S. P. (2006b). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La revue canadienne des langues vivantes, 63*(1), 59-81. doi: 10.3138/cmlr.63.1.59
Nation, I. S. P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.
Neff van Aertselaer, J. (2008). Contrasting English-Spanish interpersonal discourse phrases: A corpus study. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 85-99). Amsterdam: John Benjamins.
Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. doi: 10.1093/applin/24.2.223
Nesselhauf, N. (2004). What are collocations? In D. J. Allerton, N. Nesselhauf & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application*. Basel: Schwabe Verlag.
Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Company.
Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research, 21*(3), 298-320. doi: 10.1177/1362168816639619
Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory, 5*(1), 105-130. doi: 10.1515/cllt.2009.005
O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18*(1), 83-108. doi: 10.1075/ijcl.18.1.07odo
Ooi, V. B. Y. (2000). Asian or Western realities? Collocations in Singaporean-Malaysian English. In J. M. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English* (pp. 73-89). Amsterdam: Rodopi.
Oxford Dictionaries. (n.d.). The Oxford English Corpus. Retrieved August 25, 2016, from https://en.oxforddictionaries.com/explore/the-oxford-english-corpus

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.
Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149. doi: 10.1017/S0267190512000098
Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-225). London: Longman.
Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research*, *21*(3), 381–402. doi:10.1177/1362168815618428
Peters, E. (2009). Learning collocations through attention-drawing techniques: A qualitative and quantitative analysis. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 194-207). Basingstoke: Palgrave Macmillan.
Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research, 20*(1), 113-138. doi: 10.1177/1362168814568131
Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning, 65*(S1), 37-75. doi: 10.1111/lang.12112
Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.
Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209-227). Amsterdam: John Benjamins.
Revier, R. L. (2009). Evaluating a new test of *whole* English collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 125-138). Basingstoke: Palgrave Macmillan.
Richards, K. (2003). *Qualitative inquiry in TESOL*. Basingstoke: Palgrave Macmillan.
Rundell, M., & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 257-281). Amsterdam: John Benjamins.
Schmid, H.-J. (2003). Collocation: Hard to pin down, but bloody useful. *Zeitschrift für Anglistik und Amerikanistik, 51*(3), 235-258.
Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281-317. doi: 10.1111/1467-9922.00042
Schmitt, N. (Ed.). (2004). *Formulaic sequences*. Amsterdam: John Benjamins.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.
Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951. doi: 10.1111/lang.12077
Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 1-22). Amsterdam: John Benjamins.
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. doi: 10.1177/026553220101800103
Seidlhofer, B. (2005). English as a lingua franca. *ELT Journal, 59*(4), 339-341. doi: 10.1093/elt/cci064
Seidlhofer, B. (2009). Accommodation and the idiom principle in English as a lingua franca. *Intercultural Pragmatics, 6*(2), 195-215. doi: 10.1515/iprg.2009.011
Shillaw, J. (1999). *The application of the Rasch model to Yes/No vocabulary tests.* (Doctoral dissertation), University of Wales, Swansea.
Shin, D., & Nation, I. S. P. (2008). Beyond single words: The most frequent collocations in

spoken English. *ELT Journal, 62*(4), 339-348. doi: 10.1093/elt/ccm091
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. doi: 10.1093/applin/amp058
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (2008). The phrase, the whole phrase and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407-410). Amsterdam: John Benjamins.
Sinclair, J., Jones, S., & Daley, R. (1970). English lexical studies. In R. Krishnamurthy (Ed.), *English collocation studies: The OSTI report* (pp. 2-204). London: Continuum.
Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 64*(3), 429-458. doi: 10.3138/cmlr.64.3.429
Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory, 11*(2), 285-301. doi: 10.1515/cllt-2014-0016
Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251-272. doi: 10.1177/0267658310382068
Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics, 36*(5), 549-569. doi: 10.1093/applin/amt054
Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, *65*, 533-562. doi: 10.1111/lang.12125
Skandera, P. (Ed.). (2007). *Phraseology and culture in English*. Berlin: Mouton de Gruyter.
Smith, M. C. (2015). Word categories. In J. Taylor (Ed.), *The Oxford handbook of the word* (pp. 175-195). Oxford: Oxford University Press.
Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. In S. T. Gries & D. Divjak (Eds.), *Frequency effects in language* (pp. 127-163). Berlin: Mouton de Gruyter.
Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition, 18*(3), 419-437. doi: 10.1017/s1366728914000674
Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language, 83*(2), 227-236. doi: 10.1016/S0093-934X(02)00032-9
Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54*(2), 161-184. doi: 10.1016/j.jml.2005.11.001
Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal, 36*(2), 139-152. doi: 10.1080/09571730802389975
Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics in Language Teaching, 49*(4), 321-343. doi: 10.1515/iral.2011.017
Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language, 2*(1), 23-55.
Taylor, J. R. (2012). *The mental corpus*. Oxford: Oxford University Press.
Teddiman, L. (2012). Conversion and the lexicon: Comparing evidence from corpora and experimentation. In D. Divjak & S. T. Gries (Eds.), *Frequency effects in language representation* (pp. 235-253). Berlin: De Gruyter Mouton.
Tennant, A. (2004). Disordered thresholds: An example from the Functional Independence Measure. *Rasch Measurement Transactions, 17*(4), 945-948.
Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569-613. doi: 10.1111/j.1467-9922.2010.00622.x

van der Wouden, T. (1997). *Negative contexts: Collocation, polarity and multiple negation*. London: Routledge.
Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234. doi: 10.1017/s0142716401002041
Vetchinnikova, S. (2015). Usage-based recycling or creative exploitation of the shared code? The case of phraseological patterning. *Journal of English as a Lingua Franca, 4*(2), 223-252. doi: 10.1515/jelf-2015-0019
VOICE. (2013). The Vienna-Oxford International Corpus of English (version POS XML 2.0).
Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design.* (Doctoral dissertation), Iowa State University.
Walker, C. P. (2011). A corpus-based study of the linguistic features and processes which influence the way collocations are formed: Some implications for the learning of collocations. *TESOL Quarterly, 45*(2), 291-312. doi: 10.5054/tq.2011.247710
Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME journal, 32*, 201-232.
Webb, S. , Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*, 91-120. doi: 10.1111/j.1467-9922.2012.00729.x
West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.
Wible, D. (2008). Multiword expressions and the digital turn. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 163-181). Amsterdam: John Benjamins Publishing Company.
Widdowson, H. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-25. doi: 10.1093/applin/21.1.3
Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.
Wilks, C., & Meara, P. (2002). Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research, 18*(4), 303-324. doi: 10.1191/0267658302sr203oa
Wilson, M. (2005). *Constructing measures: An Item Response Modeling approach*. Lawrence Erlbaum Associates.
Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. *Studies in Second Language Acquisition, 23*(1), 41-69. doi: 10.1017/s0272263101001024
Wolter, B. (2005). *V_Links: A new approach to assessing depth of word knowledge.* (Doctoral dissertation), University of Wales, Swansea.
Wolter, B. (2009). Meaning-last vocabulary acquisition and collocational productivity. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 128-140). Bristol: Multilingual Matters.
Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics, 32*(4), 430-449. doi: 10.1093/applin/amr011
Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocations processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition, 35*(3), 451-482. doi: 10.1017/S0272263113000107
Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics, 36*(5), 1193-1221. doi: 10.1017/s0142716414000113
Wolter, B., & Yamashita, J. (2017). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, 1-22. doi: 10.1017/s0272263117000237
Wood, D. (2010). *Formulaic language and second language speech fluency*. London:

Continuum.
Wood, D. (2015). *Fundamentals of formulaic language*. London: Bloomsbury.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 249-268). Amsterdam: John Benjamins.
Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.
Wray, A. (2009). Conclusion: Navigating L2 collocation research. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 232-244). Basingstoke: Palgrave Macmillan.
Wray, A. (2015). Why are we so sure we know what a word is? In J. Taylor (Ed.), *The Oxford handbook of the word* (pp. 725-750). Oxford: Oxford University Press.
Wray, A. (2017). Formulaic sequences as a regulatory mechanism for cognitive perturbations during the achievement of social goals. *Topics in Cognitive Science, 9*(3), 569-587. doi: 10.1111/tops.12257
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.
Xiao, R. (2015). Collocation. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 106-124). Cambridge: Cambridge University Press.
Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly, 44*(4), 647-668. doi: 10.5054/tq.2010.235998

Appendix A: Original *LexCombi* instructions.

These instructions were provided to participants in Japanese.

The task is very simple. You will look at 30 words. For each word, all you need to do is write down **3 collocations** – three words that you most expect to see together with the word if you were to read it.

Example 1. For the word BOOK, some typical collocations are: *new*, *interesting*, *title*, *review*, *library*, *phone*, *story*, *read*, *write*, *borrow*, *copy*, *publish*.
Example 2. For the word SCHOOL, some typical collocations are: *language*, *private*, *building*, *library*, *uniform*, *go to*, *attend*, *graduate*, *finish*, *teacher*, *festival*.

NOTE:

- There are no right or wrong answers, so don't worry.
- The collocations can come before or after the word. For example:
  book title – interesting book – borrow a book – book review
  language school – go to school – school teacher – school festival
- Any part of speech is OK (noun, verb, adjective, adverb, etc.).
- You should write down your collocations quickly. You have 30 seconds for each word.
- Please DO NOT write your name or student number anywhere on this paper.
- If you have taken an English test (e.g. Eiken, TOEIC) within the last 12 months, please fill in the following:
  Test name: Date taken (month and year): Score/Result:

Please practice with these 3 words. Write down 3 collocations for each word. You have 30 seconds for each word.

holiday	1	2	3
letter	1	2	3
university	1	2	3

Please write down 3 collocations for each of these words.

house	1	2	3
issue	1	2	3
research	1	2	3
power	1	2	3
question	1	2	3
voice	1	2	3
experience	1	2	3
health	1	2	3
value	1	2	3
reason	1	2	3
example	1	2	3
child	1	2	3
war	1	2	3
paper	1	2	3
police	1	2	3

PLEASE GO TO THE NEXT PAGE.

car	1	2	3
death	1	2	3
government	1	2	3
life	1	2	3
role	1	2	3
problem	1	2	3
law	1	2	3
support	1	2	3
body	1	2	3
interest	1	2	3
decision	1	2	3
work	1	2	3
country	1	2	3
family	1	2	3
friend	1	2	3

END

Thank you for your help.

Appendix B: Modified *LexCombi* instructions.

These instructions were provided to participants in Japanese.

The task is very simple. You will look at 30 words. For each word, all you need to do is write down **3 collocations** – three phrases that you most expect to see in conjunction with the word.

Example 1. For the word BOOK, some typical collocations are: *a new book*, *an interesting book*, *the book's title*, *a book review*, *a book from the library*, *the phone book*, *read a book*, *write a book*, *borrow a book*, *copy some pages from a book*, *publish a book*.
Example 2. For the word SCHOOL, some typical collocations are: *a language school*, *a private school*, *the school buildings*, *the school library*, *my school uniform*, *go to school*, *attend school*, *graduate from school*, *finish school*, *a school teacher*, *the school festival*.

NOTE:

- There are no right or wrong answers, so don't worry.
- Any part of speech is OK (noun, verb, adjective, adverb, etc.).
- You should write down your collocations quickly. You have 30 seconds for each word.
- If you have taken an English test (e.g. Eiken, TOEIC) within the last 12 months, please fill in the following:
  Test name: Date taken (month and year): Score/Result:

- The 30 words are all nouns.
- Enter one phrase in each box (three phrases in total for each word).

Appendix C: Yes/No Vocabulary Test (*X_Lex*)

These instructions were provided to participants in Japanese.

Please look at these words. Some of these words are real English words and some are invented but are made to look like real words. Please tick the words that you know or can use. Here is an example.

dog ✓

cliff	both	relative	with	century
pardoe	upward	arrow	cup	provide
person	feeling	discuss	feel	round
moreover	treadaway	pedestrian	staircase	park
stream	question	bullet	darrock	anxious
path	publish	tower	woman	crisis
mercy	outlet	believe	trunk	sumption
fine	weather	gillen	press	produce
tube	independent	conduct	wheel	drum
normal	arrive	cantileen	rake	horozone
difficult	whole	juice	that	impress
frequid	perform	fishlock	daily	hyslop
fade	dam	pity	waygood	instead
mount	probably	early	violent	brighten
hobrow	insult	signal	jug	headlong
everywhere	permission	gazard	diamond	reasonable
candlin	table	dish	effect	military
contract	refer	chart	kennard	earn
litholect	limp	nod	antique	manomize
stand	sweat	feeble	essential	market
deny	alden	trick	sorrow	oak
gumm	cardboard	group	manage	horobin
humble	before	associate	slip	sandy
lessen	shot	gentle	mud	boil

Appendix D: Adapted *LexCombi* instructions

These instructions were provided to participants in Japanese.

The task is very simple. You will look at 30 words. For each word, all you need to do is write down 3 collocations, something that would be used either before or after the word. Please write each response on the line next to the word in the box.

Example 1. For the word BOOK, some typical collocations are:
new, *interesting*, *title*, *review*, *library*, *phone*, *story*, *read*, *write*, *borrow*, *copy*, *publish*.

Example 2. For the word SCHOOL, some typical collocations are:
language, *private*, *building*, *library*, *uniform*, *go*, *attend*, *graduate*, *finish*, *teacher*, *festival*.

Write each collocation next to the word, either to the left or the right as appropriate. Write only one word in each box.

For example, for the words BOOK and SCHOOL:

new book _____	*read* book _____	_____ book *title*
go school _____	_____ school *uniform*	*private* school _____

NOTE:

1. There are no right or wrong answers, so don't worry.
2. The collocations can come before or after the word. For example:
   book title – interesting book – borrow a book – book review
   language school – go to school – school teacher – school festival
3. Any part of speech is OK (noun, verb, adjective, adverb, etc.).
4. You should write down your collocations quickly. You have 30 seconds for each word.
5. All 30 words are nouns.
6. If you have taken an English test (e.g. Eiken, TOEIC) within the last 12 months, please fill in the following:

Test name:

Date taken (month and year):

Score/Result:

7. Write only one word in each box (three words in total).
8. Write your student number here.

Student number:

Please practice with these 3 words. Write down 3 collocations for each word. Write each collocation either before or after the word. You have 30 seconds for each word.

_____ holiday _____	_____ holiday _____	_____ holiday _____
_____ letter _____	_____ letter _____	_____ letter _____
_____ university _____	_____ university _____	_____ university _____

Please write down 3 collocations for each of these words. Write each collocation either before or after the word.

_____ house _____	_____ house _____	_____ house _____
_____ issue _____	_____ issue _____	_____ issue _____
_____ research _____	_____ research _____	_____ research _____
_____ power _____	_____ power _____	_____ power _____
_____ question _____	_____ question _____	_____ question _____
_____ voice _____	_____ voice _____	_____ voice _____
_____ experience _____	_____ experience _____	_____ experience _____
_____ health _____	_____ health _____	_____ health _____
_____ value _____	_____ value _____	_____ value _____
_____ reason _____	_____ reason _____	_____ reason _____
_____ example _____	_____ example _____	_____ example _____
_____ child _____	_____ child _____	_____ child _____
_____ war _____	_____ war _____	_____ war _____
_____ paper _____	_____ paper _____	_____ paper _____
_____ police _____	_____ police _____	_____ police _____

PLEASE GO TO THE NEXT PAGE.

_____ car _____	_____ car _____	_____ car _____
_____ death _____	_____ death _____	_____ death _____
_____ government _____	_____ government _____	_____ government _____
_____ life _____	_____ life _____	_____ life _____
_____ role _____	_____ role _____	_____ role _____
_____ problem _____	_____ problem _____	_____ problem _____
_____ law _____	_____ law _____	_____ law _____
_____ support _____	_____ support _____	_____ support _____
_____ body _____	_____ body _____	_____ body _____
_____ interest _____	_____ interest _____	_____ interest _____
_____ decision _____	_____ decision _____	_____ decision _____
_____ work _____	_____ work _____	_____ work _____
_____ country _____	_____ country _____	_____ country _____
_____ family _____	_____ family _____	_____ family _____
_____ friend _____	_____ friend _____	_____ friend _____

END

Thank you for your help.

Appendix E: Yes/No Vocabulary Test (*X_Lex*/*Y_Lex*)

These instructions were provided to participants in Japanese.

Please look at these words. Some of these words are real English words and some are invented but are made to look like real words. Please tick the words that you know or can use. Here is an example.

dog ✓

bullet	effect	tactical	jug	ale
park	haze	woman	deny	press
megalodic	accumulation	elector	fictional	mount
violent	recoup	frequid	infer	dish
feeling	difficult	treadaway	chart	sumption
that	gorge	question	indigo	murray
impress	reasonable	juice	crocodile	oak
pinch	requined	ferdy	outlet	relative
abergy	fallow	publish	ordeal	lustrous
ashill	feeble	fender	believe	substitution
gumm	brendation	pout	provide	moreover
nicked	garge	modular	candlin	rugged
jumble	mud	overthrow	scruffy	animation
cantileen	blunder	illuminate	vessy	arrow
refuge	limp	hyslop	evaculate	authorize
litholect	group	battered	market	biblical
chipboard	ledge	spurt	person	pardoe
lethargic	drum	contract	crisis	daily
gentle	unfold	gillen	contrive	permission
refer	trunk	insult	maiden	horozone
upward	with	gaze	arrive	constrain
lessen	excellence	innoculism	cup	dam
normal	perrate	whole	brighten	shuffle
seclusion	manage	table	century	geological
before	horobin	sorrow	anxious	pulity
gale	allay	cripple	both	thoughtful

obsolete	signal	bonfire	upheaval	cigar
sweat	calcite	interviewer	shot	staircase
sandy	hobrow	headlong	stunt	antique
cliff	psychiatrist	kilp	wrinkle	oar
waygood	eternal	idleness	fade	peninsula
trait	rubber	fiver	scornful	slip
crust	nod	perform	mercy	overshadowed
endorsement	harness	fishlock	diamond	privileged
trick	wheel	restless	instead	catastrophe
gazard	cardboard	fine	earn	humble
early	encoding	mauve	summon	bridesmaid
pity	anger	sinner	feel	demonstrator
shipment	produce	manomize	path	pedestrian
unsightly	cupoid	dicey	round	cloakery
stand	disguise	harmonical	warden	politicure
stream	kennard	independent	probably	banquet
dispatch	straddling	tube	darrock	cocky
contributor	military	tower	weather	workman
reap	slap	rake	traitor	twig
draconian	discuss	essential	mandatory	spalding
misfortune	everywhere	treggle	alden	infect
conduct	ramp	unsuccessful	associate	boil