Individual profiling of second language learners through word association.

Higginbotham, George Michael

Individual profiling of second language learners through word association

George Michael Higginbotham

A thesis submitted to Swansea University
in fulfilment of the requirements for the degree of
Philosophiae Doctor

2014

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

Abstract

This thesis explores the organisation of second language learners' mental lexicons through the use of word association tests; a reliable measure of which would complement established measures of lexicon size. Following studies with native speakers (Russell & Jenkins, 1954; Ervin, 1961), research with second language learners began in the late 1950's (Lambert, 1956) although much of the methodology and theory had been developed decades before by clinical psychologists (Jung, 1918). Unlike the L1 studies, the L2 studies have been plagued by inconsistent findings, leading some to conclude that the use of word associations to assess L2 learners is unfeasible (Kruse et al., 1987). In an attempt to realise the potential that word association tests have as a method of measuring the organisation of learner lexicons, a series of experiments was conducted. The initial experiment was a replication of Wolter (2001) using a traditional classification system. This was followed by five more experiments that centred around a quite different methodology and approach to data analysis put forward by Fitzpatrick (2006; 2007). The reliability of Fitzpatrick's *individual profiling* approach was tested using various kinds of stimulus words. The results indicate that the word class and frequency of the stimuli have little effect on the reliability of the response profiles generated. Improvements to the methodology and issues that arose during the experiments are discussed. The experiments were all conducted in Japan, with college aged learners between early 2007 and mid-2012. In that six year period, over 20,000 responses were elicited from 213 learners involved in the pilot tests and main experiments.

DECLARATION

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

STATEMENT 1

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

STATEMENT 2

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

SIGNED ....

DATE .....16th January 2014.....

Acknowledgements

While working on this thesis I sometimes felt like a child walking through a field of tall grass, although I usually had a sense that I was moving forward I was not always sure where I was headed or even how far I had come. Clambering up the findings of other researchers allowed me to see a little further and get my bearings, although it was mainly the advice, assistance and occasional push from my seniors that kept me moving in the right direction. Fortunately I had experienced guides to help me out when I needed it. My thanks therefore first of all go to my supervisor Professor Paul Meara, for his patience and perspicacity throughout the preparation of this work. While allowing me the freedom to make mistakes, his depth of knowledge and enthusiasm for research into vocabulary acquisition meant that I didn't stray too far, and in the process learned a lot. Conversations and emails with members of the Swansea University VARG network, many of whom took considerable time to discuss and refine my arguments, were also valuable. In particular, Professor Tess Fitzpatrick and Dr. Ian Munby deserve a special mention for reading and advising on early drafts of chapters in this thesis. For access to the learners that were used in the pilot tests and main experiments I needed to rely on colleagues in neighbouring universities. I therefore owe a debt of gratitude to Dr. Jim Ronald, Dr. Monika Szirmai, Katherine Song, Muneo Hotta, and of course their students, all of whom gave up their time in helping me collect the data. Last but not least, I am deeply appreciative of my family, especially my wife, for all the encouragement over the years. Without a supportive family I would have achieved little, and it is therefore to my family that I dedicate this thesis.

Table of Contents

Chapter Five: Revisiting the effect of word frequency

Chapter Six: The effect of verb stimuli

List of Tables

List of Figures

Chapter One: Introduction, Background, and Overview

1.1 Introduction

As Aitchison's (1987) aptly titled and entertaining introduction to the field states, psycholinguistics is concerned with "words in the mind". Its theoretical and methodological roots stem from two long established areas of research: psychology and linguistics. In this chapter, I will begin with an explanation of two constructs, the *mental lexicon* and *word association* that are fundamental to this thesis. This will be followed by a condensed history of the uses that word association tests have been put to over the last 130 years. The main purpose of this chapter is to explain the potential of word association tests in exploring the mental lexicons of second language learners. As well as introducing key concepts and discussing the background to the methodology, I will also give an overview of the structure of the thesis and the main themes that will be pursued.

1.2 What is the mental lexicon?

The mental lexicon can be thought of as the mental space that stores the words and phrases used in understanding and/or producing language. The most complex of our cognitive functions, language (and its constituent parts), is difficult to conceptualize other than through metaphor. One commonly used metaphor for the mental lexicon is that it resembles a dictionary or thesaurus (Meara, 1978), although as Pavičić Takač (2008:11) points out, such a comparison is unsatisfactory as printed dictionaries are "static, limited and prone to become outdated" with items not necessarily being stored alphabetically. On the contrary, it is widely agreed (Aitchison, 1987; Pavičić Takač, 2008) that the mental lexicon is "connected into semantic networks" and characterized by "fluidity and flexibility". When new words are learned or extra information on existing words is added, there is a need for reorganization. Even for adult (L1) speakers the system must be in a constant state of flux, although we can probably assume the greatest instability would exist in those still at the early stages of competency such as native children and second language (L2) learners. There is a temptation to view the mental lexicon of L2 learners (the focus of this thesis) as being less complex than the lexicon of native speakers, as learners have smaller L2 vocabularies. This is a misconception as the first language is already in place before the learning of any subsequent language; the learner's mental lexicon is a more complex system as it comprises both L1 and L2 items.

Another widely accepted metaphor (Aitchison, 1987:72) is that the lexicon is "a

gigantic multi-dimensional cobweb in which every item is attached to a score of others". In this view the lexicon can be seen as a complex structure with words having multiple links and differing levels of connectivity to other words in the network. In some respects this works quite well, the notion that different parts of the 'web' are more connected than others agrees with the idea of *high* and *low* frequency lexical items. Words that occur rarely in the language, such as *perspicacious,* can be thought of as existing in a more tenuous state at the edge of the web. Common words, such as *clever*, can be thought of as occupying a position closer to the center of the web with a greater number of connections. Despite the attractions the 'web' metaphor holds it does have limits. A recent study by Meara (2011) suggests that the lexicon is far too dense to be thought of as a light diaphanous substance like a cobweb. Meara calculates that for a network of 900 very common words there are over 20,000 connections, with such a density he argues that it might be more correct to think of the lexicon as "bindweed" rather than "gossamer". Whether we choose to view the lexicon as an internal dictionary or an organic network, we ought not to lose sight of the fact that metaphors are simply figurative representations of reality, and when pushed eventually breakdown.

While metaphors are useful in giving a general idea of something as intangible as human thought, as a base on which to build research there is greater merit in working from carefully phrased definitions. For such a widely written about construct though, there have been surprisingly few attempts to precisely pin it down. Of the handful of definitions for the mental lexicon that have been proposed, three will be discussed. These not only help to explain the mental lexicon but also illustrate the different kinds of thinking that the two main research traditions bring to psycholinguistics. The first two definitions are by linguists who view the lexicon from the perspective of the foreign language learner. The third definition is provided by two researchers more concerned with the psychological and neuropsychological side of psycholinguistics.

Definition 1: (Hulstijn, 2000:210)

> *the mental lexicon is a memory system in which a vast number of words, accumulated over time, has been stored.*

The first point about this definition is that it recognizes the sheer size of the mental lexicon. Schmitt (2010:6) estimates an educated native adult speaker's vocabulary will "range between 16,000 and 20,000 word families" and as Bauer & Nation (1993:253) state: "a word family consists of a base word and all its derived and inflected forms". Each of these

forms is connected in various ways (meaning, phonology, orthography, syntactic characteristics) to other words in the network. When we also consider that for many there is also an L2 (and perhaps part of an L3) system being layered on top of this, calling the system 'vast' is no exaggeration. The second point that this definition highlights is that words are not completely acquired in the first-encounter; bits and pieces of word knowledge are gradually 'accumulated' through repeatedly meeting the item in various contexts.

Definition 2: (Richards & Schmidt, 2002:327)

> *the mental lexicon is a person's mental store of words, their meanings and associations*

This definition is useful for two reasons, firstly by specifying 'person' the uniqueness of every individual's lexicon is recognized, a major theme running through this thesis. Secondly, by specifying 'meaning and associations', the multiple aspects of word knowledge (Nation 2001:27) are hinted at. As with the initial definition, being so short it fails to fully account for some important components of the mental lexicon. For example, of the aspects of word knowledge, only 'meaning and associations' are made explicit. Other aspects, such as: what the word sounds like, how it is spelled, its syntactic limitations or frequency of use, are not mentioned. These are surprising omissions given that Richards wrote a seminal paper on the aspects of word knowledge (Richards, 1976). A further problem with this definition is that it does not clearly state that the words are connected in a network. Although by mentioning 'associations' this is perhaps implied, it is easy to interpret this as merely referring to an internal component of the words themselves.

Definition 3: (Jarema & Libben, 2007:2)

> *The mental lexicon is the cognitive system that constitutes the capacity for conscious and unconscious lexical activity.*

Unlike the previous definitions, the issues of size, gradual accumulation, uniqueness of the individual and the aspects of word knowledge are not made explicit. On the positive side, the third definition does however ask us to consider a 'cognitive system', the implication being that words are part of a network – a crucial point. The third definition also specifies 'conscious and unconscious' activity. While it is perhaps redundant to state both, it underlines the fact that in many cases we are able to consciously consider the connections

made between words, but we often do not do this. By doing so we would not be able to express our ideas fluently. The constraints of real-time communication require that a lot of the connections be made so quickly as to be almost automatic (Hulstijn, 2007). I also think the idea of 'capacity' is useful; while there are many connections between words available for us to use it does not necessarily follow that they will all be active and ready for use when demanded. The tip of the tongue phenomenon, a common word that we 'know' but under pressure just don't seem to be able to produce, is a frustrating experience familiar to most.

As noted in the preceding paragraphs, while the three definitions contain valid points they are all, disappointingly, incomplete. However, as these points seem complementary, a logical step would be to combine them into a more inclusive definition. It would also make sense to incorporate the idea that came out of the analyses of the metaphors, that the mental lexicon is a dynamic system. Thus, a more comprehensive (though unfortunately less succinct) definition would be:

> *The mental lexicon is a vast cognitive system that constitutes a mental store of the multiple aspects of word knowledge that an individual gradually accumulates for each word or phrase. There are numerous links between words in this dynamic network, which has the capacity for conscious and unconscious lexical activity.*

Within this definition it ought to be noted that 'word' is used in the broader sense. This goes beyond the everyday meaning of 'word' as 'a string of letters with a space either side', such as *dog*, *watch* or *ear*. As Pawley & Syder (1983) argue, due to the speed with which people communicate it is unlikely that words are stored in the lexicon as single units alone. It is more likely that in addition they will also be stored in larger units such as *dog eared*, *dog watch* or *gone to the dogs*. The term 'word' also covers these multi word units whose meaning cannot be worked out through adding together the individual parts.

Given the current lack of a widely accepted model explaining how the mental lexicon is organized and how new items become integrated, the above definition will need to be refined as our understanding improves. The definition will however serve as a base on which to build this thesis as it includes the components I think are key to its construct.

1.3 What is a word association test?

Now that we have a definition of the mental lexicon to work with we can move on to the next point, how to measure its characteristics. Much of human behaviour can be measured directly by observation, what happens within our brains however is not so easy to discern.

The measurement of thought, and more specifically, a construct such as the mental lexicon requires an indirect approach. Due to the vastness of the lexicon, Meara (1996:50) argues it is impractical to measure every aspect of word knowledge for a representative sample of words; a better approach is therefore to develop tests that cover key "global characteristics". The two dimensions he puts forward are "size and organization". Tests of vocabulary size (The Vocabulary Levels Test, Nation 1983; The Eurocentre's Size Test, Meara & Jones 1990) are widely used and have been checked for reliability and validity (Schmitt, 1996; Schmitt et al., 2001), more importantly they have survived the test of time. With an enhanced version of the VLT - The Vocabulary Size Test (Nation & Beglar, 2007) recently introduced, it seems that measuring the size dimension is well in hand. A widely accepted test of the 'organizational' characteristic has however yet to emerge, although as Meara's (1996) paper suggests, one way to achieve this is through word association tests. It is the analysis and development of reliable word association tests, as a way to understanding the organization of a person's mental lexicon that is the focus of this thesis.

In its simplest form a word association test (WAT) is a task where the person being tested is given a stimulus word and is asked to respond to it with the first word that comes to mind. It is, as Entwistle (1966:1) notes, "a method for gathering data relevant to verbal habits and linguistic development". The main advantages that are cited with WATs are that they are relatively quick to administer, give rich information about a user's knowledge of a word and are uncluttered by context. Deese (1965:39) argues, "the free-association test has survived as a technique of psychological investigation because it is an instrument for detecting the sequences of thought as these seem to exist in their most unconstrained form." More recently, Milton (2009:141) notes that a benefit WATs have over other kinds of language test, is that "people carrying out such tasks are not hindered by the requirement to produce grammatical or well structured grammar". Such tests are therefore well suited for use with language learners, who by definition have difficulty doing language tasks that require comprehension/manipulation of sentences or longer texts. The lack of context is however a double-edged sword; while it only puts a minimal burden on the testee in terms of output, it does not measure language in a particularly authentic or 'communicative' way. Another potential flaw that Milton (2009:141) points out is that "it only works when learners willingly engage with the purpose of the exercise and do not try and maximize their scores".

There are many variations of the WAT, and these will be briefly explained before we explore the history of how the different kinds of WAT have been employed over the

years. A basic distinction we can make is between *free* and *controlled* WATs. In the *free* version, a number of stimulus words (also referred to in the literature as *cue* or *prompt* words) are given to a testee and then the testee is asked to respond to them. This can be done orally, using a paper test or via a computer. Usually the testee is asked to make only one response per stimulus, although a common variation is to ask the testee to give multiple responses. Of the multiple response formats two further variations exist, *continuous* and *continued.* In the continuous test the stimulus is only given once whereas in the continued test the stimulus is repeated many times between responses. This is to ensure that each response is with the main stimulus word rather than responses to the first or second responses – a way of responding known as chaining. The *controlled* word association tests work slightly differently. In this kind of WAT the testees are given stimulus words although their response is limited in some way. A testee could for example be given a stimulus and asked to connect the word with one or more words in a pre-determined list. The "word associates test", developed by Read (1993) for his language learners, follows this format. Other constraints might restrict the response by requiring the testee to respond with a specific word class or perhaps a word in the same lexical group. These variations have been developed over the years and enabled researchers to explore different aspects of the mental lexicon for different purposes. These can be roughly divided into two main research strands, the movement towards native-like *proficiency* and response *type*. As these strands will be dealt with in considerable detail in Chapter 2, in the following paragraphs only a brief outline is provided.

The first of the two main research strands, that has received considerable attention, is how stereotypical learner responses are in contrast to native responses. Many have argued that analysis of responses in this way can be used as a measure of proficiency in the language (Lambert, 1956; Randall, 1980; Piper & Leicester, 1980; Kruse et al., 1987; Read, 1993; Nishiyama, 1996; Schmitt & Meara, 1997; Schmitt, 1998a, 1998b; Greidanus & Nienhuis 2001; Wolter, 2002; Henriksen, 2008; Zareva, 2005, 2007, 2011). Researchers working in this *proficiency* strand measure factors such as the speed, number and also the quality of learner responses and through comparison with norms lists make inferences about the learner's language ability. These studies are based on the assumption that as learners become more proficient, their responses become more native-like. A norms list that has often been used as a benchmark against which to measure native-like ability is the Postman & Keppel list (1970). Recently (Meara & Schmitt, 1997; Schmitt, 1998a; Zareva, 2011) researchers have been using other lists, such as The Edinburgh Associative

Thesaurus (Kiss et al., 1973), or developing their own norms lists.

The second of the two main research strands, is concerned with measuring how the words in the network relate to each other in terms of their semantic or lexical meaning. In this tradition researchers have made inferences about the organization of the mental lexicon for certain groups of people or certain word classes based on the type of responses generated. As this thesis fits into the research strand concerned with analyzing word associations responses based on their lexico-semantic type, it is relevant to consider at this point how these *types* of response might be classified. Although various competing classification systems have been suggested, the traditional method in L1 studies (and also many L2 studies) has been to divide responses into three main groups:

> **Paradigmatic**: meaningful responses which are in the same word class, these *vertically* related responses can often replace each other in a sentence and remain grammatically correct. Synonyms, antonyms, meronyms and hyponyms would fall into this category.
>
> For example: *key→lock* or *dog→puppy*
>
> **Syntagmatic**: meaningful responses that are not (usually) in the same word class but are linked *horizontally* in the sentence. Collocations would fall into this category.
>
> For example: *key→low* or *dog→tired*
>
> **Clang/Phonological**: responses that don't have a meaningful connection with the stimulus but have a similar sound and/or spelling to the stimulus.
>
> For example: *key→keen* or *dog→fog*

Many studies of *type* have focused on the development of the lexico-semantic relationships in native speaker lexicons (Ervin, 1961; Entwistle, 1966; Deese, 1965; Stolz & Tiffany, 1972; Emmerson & Gekoski, 1976; Fitzpatrick 2007). There have also been numerous studies of second language learners (Riegel & Zivian, 1972; Meara, 1978; Politzer, 1978; Söderman, 1993; Wolter, 2001; Orita, 2002; Namei, 2004; Zareva, 2005; 2011; Bagger-Nissen & Henriksen, 2006; Fitzpatrick, 2006; 2009).

1.4 The history of word association testing

Deese (1965) and Entwistle (1966) point out that scholars have been thinking about associations since the time of Aristotle and that the laws of association can be traced back to the 17th century philosopher John Locke. We will however begin our story in the 19th century, with Sir Francis Galton. The definitive 'gentleman scholar', Galton is an unlikely

character to find cited in a modern thesis. Infamous as the founding father of eugenics, his innovative contributions to other fields are consequently often overlooked. Driven by an obsession to measure humans in every aspect he could think of, in pursuit of his belief that some types of humans are superior to others, he made numerous discoveries and inventions: such as the uniqueness of fingerprints. Despite the distasteful direction of his primary field of research, Galton (1883) warrants mentioning here for two reasons. Firstly he can be credited with documenting the first free productive word association test, the form of WAT that is used within this thesis. Galton tested himself on four occasions with 75 stimulus words with about a month between each sitting. He wrote down the first two or three words that he associated with each word within a four second timeframe, using what was then cutting-edge technology (a stopwatch) to record the time. Contrary to his expectation, the associations were marked by a lot of repetitions of the same word. For example, 23% of his responses to stimulus words were the same on all four occasions with a further 21% occurring three out of the four times. As well as recording the quantity of the responses he also devised four qualitative categories for the associations. The second reason for mentioning Galton is his pioneering work in the field of statistics, particularly correlation coefficients: now a standard tool in behavioural science that not only features in many of the papers in the following chapter but also my own experimental work later in this thesis.

Following Galton there were many studies by clinical psychologists such as Jung (1902) who used the free word association test and developed norm lists to diagnose patients. In their 1918 paper, Jung and colleagues explain a method of diagnosing psychopathological conditions through WATs. They created an extensive prompt list of over 200 words from mixed word classes and a detailed classification system. They then tested a lot of people without psychological disorders, to create a response norms list, and then compared these to the responses of their patients in order to identify psychological problems. A stimulus list developed at that time, the Kent & Rosanoff list (1910), became a standard reference. Using this list and following Jung's methodology, Bleuer (1924) for example claimed, "a primary symptom of schizophrenia is a loosening of association". Basing a clinical diagnosis on vaguely defined concepts such as 'a loose association' did not really catch on. In any case, by the late 1920's the work of Freud (1900) had become more accepted, clinical psychologists began to move away from word association tests in favour of dream analysis and testing other kinds of cognitive behaviour.

At about the same time, linguists and psychologists (Woodworth, 1938) began to apply word associations to the study of language development. For a detailed explanation

of the kind of work that was done within an L1 context Cramer's 1968 review is recommended; in this volume she describes over 300 word association studies conducted between 1950 and 1965. Particularly influential projects were initiated by researchers such as Russell & Jenkins (1954), Ervin (1961) and Entwisle (1966). These large-scale L1 studies (typically around 1000 subjects) used word association tests to establish general associative patterns such as the *syntagmatic – paradigmatic shift*. Simply stated this theory holds that children initially make a lot of syntagmatic associations (*pot* in response to *flower*) and then as their lexicon matures they make more paradigmatic associations (*daisy* in response to *flower*). The early L1 studies also laid much of the theoretical groundwork by applying the laws of association to linguistics. Deese (1965:2) informs us that the primary law of association - *the law of contiguity*, can be expressed as "one thought leads to another because it causes another". Of the secondary 'modifying' laws of association: duration, vividness, intensity and frequency, it was frequency (the most accessible of these influences) that formed a central role in many of the studies of this period. The law of frequency tells us that the response to any idea or word is not random but determined by how frequently the stimulus is heard in general and how frequently two words have been heard together in the past. As well as a discussion of the theory of association Deese (1965) details two experiments undertaken at John Hopkins University on 100 native English speakers' associative responses. In the first study he specifically looked at the responses to adjective stimuli. He expected to find (and did) that the most common adjectives are composed of a limited set of polar opposites. Of the 278 adjectives, 29% were polar opposites, *alive* for example generated *dead* 44% of the time and *bad* generated *good* 43% of the time. Frequency was found to be an important variable, the adjectives generating polar opposites were usually high frequency adjectives (e.g. *black→white*), the less frequent adjectives however often generated nouns (e.g. *grand→canyon*). In the second experiment he used noun stimuli, again with university students. The most basic finding was that nouns generated other nouns. An interesting point about Deese's 1965 study is that he rejected the traditional classification used by his peers (if a noun generates a noun it is *paradigmatic* if it generates another form class it is *syntagmatic*) as being too simple to be of any real use. He elected instead to classify the responses into categories aimed at grouping responses with *similar* or *different* nominal characteristics.

While such L1 studies were being conducted it also began to occur to some researchers (Lambert, 1956; Riegel & Zivian, 1972) that the methodology could be applied to bilingual students, the expectation being that adult L2 learners would behave in a similar

way to L1 children. A particularly insightful paper (Meara, 1983) is viewed by many as a good place to begin a discussion of the problems and assumptions involved with using word associations with learners. Primarily, this paper is a review of a series of word association studies done with L2 learners at Birberk College, London in the late 1970's and early 1980's (Meara, 1978; Beck, 1981; Hughes, 1981; Morrison, 1981). The main findings were that learner responses:

differed fairly systematically from native speakers;

were far more unstable than native speakers;

were inhomogeneous;

were often based on phonological connections rather than on meaning.

As this was a "preliminary skirmish" into L2 word associations much of this paper deals with problems in collecting and analyzing the data. These problems, comments and suggestions for future research are summarized in Table 1.1.

Table 1.1 Problems in using word association tests with L2 learners (Meara, 1983)

Problem	Comment/potential solution
It is difficult to classify responses using the syntagmatic/paradigmatic system.	
L1 norms lists (particularly Kent-Rosanoff, 1910) are inadequate for use in L2 studies. This is mainly due to a lot of the words in the list being high frequency words.	Abandon standard lists in favour of carefully considered word lists that suit the research questions
Learner responses are not as stable as native speaker responses.	Need to identify what conditions lead to stable patterns, what the causes of the instability are and how long they last.
Learners make a lot of errors, both in identifying and responding to words.	In analysing responses one needs to consider a learner's language (L1+L2) as an integrated whole rather than a set of discreet components.
A learner's L1 interferes with his/her responses.	
L2 users do not necessarily approach the word association task in the same way as L1 users.	
Learning words is a gradual process, not an all-or-nothing activity.	One-off studies are ineffective measures. Studies ought to include a test-retest strategy over a period of time.
Learner responses are less homogenous than native speakers	
There is a lack of general theoretical models to account for word association responses	

In Table1.1 there are suggestions on how to resolve some of these problems. Three crucial areas are however not addressed: the problem of reliably classifying the responses, the variability within non-native speaker groups and the lack of a broad theoretical model with which to explain the response data. Following Meara's1983 paper, there were however a series of conflicting findings, some studies supported the idea of using WATs as a measure of proficiency (Söderman, 1993) and others rejecting it (Kruse et al., 1987). Some studies found that L2 lexicons shifted from being syntagmatically dominated to paradigmatically dominated with increased proficiency (Piper & Leicester, 1980; Söderman, 1993; Orita, 2002; Namei, 2004). Others had findings that were incompatible with such a shift (Wolter, 2001; Bagger-Nissen & Henriksen, 2006). The optimism of the early 1980's had become bogged down in contradictory research findings, a fresh approach was needed. One researcher who explicitly set out to re-evaluate basic assumptions and resolve Meara's problems (Table 1.1) was Fitzpatrick (2006), who proposed a more precise classification system. In a subsequent paper Fitzpatrick (2007:328) also addressed the problem that learners are not homogenous, proposing an "individual profile" style of analysis.

In Fitzpatrick's work (2006, 2007, 2009), and also studies by Schmitt (1998a) and Wolter (2001), there is serious consideration given to the selection of prompt words and how to analyze the responses generated. Following Meara's arguments it is no longer considered good practice to unquestioningly use the Kent & Rosanoff list (1910) as a source for prompt words. These days, prompt words are usually selected in a principled way in order to provide data that will help to answer specific research questions. In his 1983 paper Meara was one of the first to argue that word association tests could be used as a tool to measure the structure and organization of a learner's mental lexicon and answer questions such as: "what does a learner's mental lexicon look like?" and "how is it different from the mental lexicon of a monolingual native speaker?" At the time such claims might have appeared fanciful due to the numerous problems identified. A quarter of a century (and a few large empirical studies) later Fitzpatrick (2009:52) states: far from being fanciful, they were "prescient". There is still a fair amount of disagreement over how best to collect word association data, interpret the results and apply the findings; but there are currently few who would disagree with the essence of Meara's claims in the early 1980's. It is now widely accepted (Henriksen, 2008; Fitzpatrick, 2009; Zareva & Wolter 2012) that word associations can tell us something useful about a learner's mental lexicon.

1.5 Overview of the thesis

This thesis is primarily an examination into current methods of collecting and analysing word associations for the purpose of describing how L2 mental lexicons are organised. In conjunction with the measures of *size* that are currently available it is argued that a reliable measure of *organisation* would be useful in understanding the global characteristics of a learner's lexicon. In Chapter 2 there will be a detailed review of eleven studies that illustrate how word association research has proceeded over the last 30 years, including an extended discussion of the main achievements and problems in this field. Chapter 3 reports a close replication of an influential study (Wolter, 2001) that holds potential in terms of the theoretical model it proposes and careful attention to methodology. Following this, an alternative approach first suggested by Fitzpatrick (2006), is explored from a number of angles. Fitzpatrick's *individual profiling* idea is taken up in Chapters 4 – 8 with experiments to test the reliability of the construct and evaluate the methodological framework. The main question being addressed is whether Fitzpatrick's approach can generate stable profiles with stimuli from different frequency ranges and different word classes. In Chapter 9 there is a general discussion of the main findings of this series of experiments with suggestions on how research in this area might proceed. The final part of the thesis (Chapter 10) draws together the main points from the general discussion.

Chapter Two: Literature Review

2.1 Introduction

As the main focus of this thesis is with language learners (L2), the large body of research using word associations to understand first language (L1) development, outlined in Chapter 1, will be put to one side unless directly relevant to the discussion. Early L1 word association studies by Galton (1883) and Deese (1965) and also the studies by clinical psychologists (Jung, 1918) are interesting in fleshing out the historical development and uses that word association have been put to over the last century. It is however necessary to now narrow our perspective to studies more pertinent to the experiments in Chapters 3 - 8. This chapter therefore consists of a detailed review of eleven studies that have made contributions to our understanding of the network of words that form a learner's mental lexicon. There are two basic strands of research into the L2 lexicon, studies that argue WA responses can be used as a measure of *proficiency* and studies that use the *type* of response to explore the semantic organisation of the lexicon. The primary focus of this thesis is with *type*, although studies that are mainly concerned with *proficiency* cannot be ignored. Firstly, many of these proficiency studies also examine the type of response and secondly there is a considerable overlap in the methodology and assumptions that underpin them. This overlap can be seen in studies such as Politzer (1978), Söderman (1993b) and Orita (2002) which argue that a shift in response type can discriminate between language users of differing proficiency. Furthermore, studies that are concerned with type of response generally account for learner proficiency as one of the many variables that will affect responses. A discussion of both strands can therefore help fit the studies of response type into the broader framework of research into learner lexicons. Within each of these two main strands the papers are explored in chronological order. By ordering them in such a way there is no intention to imply that the methodologies and findings of these studies follow each other in a well-ordered progression. The main reason for this is that it seems the least complicated way of grouping the studies and allows the reader to follow the development of ideas and methodologies over the last 30 years. The selection of papers reviewed attempts to give a broad view of how the research in this field has progressed (and regressed) with inclusion of papers that reflect both negatively and positively on the use of word associations to answer questions about the mental lexicon of learners.

Following an objective summary of each study within the two main strands and comments on the key points of interest that each paper raises, there will be a general

discussion of some of the more persistent issues. The issues addressed will give greater context to the series of experiments in this thesis and support the approach that was eventually adopted.

2.2 The *Proficiency* response strand

The initial strand, papers that are more concerned with using word association tests as a measure of proficiency, begins with Politzer (1978). This is by no means the first study to view word association behaviour as a potential measure of learner proficiency, that distinction can be attributed to Lambert and colleagues (Lambert, 1956; Lambert & Moore, 1966). The decision to start with Politzer was made due its comparative recency and the large (and lingering) impact on the field. Also, this paper is particularly interesting for two further reasons. It was the first to suggest that the kind of word associations made by L2 learners are similar to those made by L1 children (in that they make a lot of syntagmatic links) and there is an attempt to link associative patterns with pedagogy. The problems highlighted in this study reveal that studying the L2 context is not merely a case of applying methodologies and assumptions that seem to work with native speakers. An equally influential paper by Kruse, Pankhurst and Sharwood-Smith (1987) is next on the list. This requires inclusion due to the severe criticism of the use of word association response data in general, which led to the basic methodology falling out of favour for about a decade. Some research did however persevere with this methodology, one study from this period (Söderman, 1993b) is included. This study attempted to work on the ideas initially put forward by Politzer. Although still problematic this study certainly represents a step forward in terms of experimental design. The work by Söderman is important as there is an attempt to look at lower frequency stimulus words and compare the findings with more frequent stimulus items. Along with word class, the frequency of the stimulus has long been assumed (Deese, 1965; Cramer, 1968) to be a key variable in determining response. Until this point however studies had generally limited themselves to high frequency stimulus items taken from the Kent-Rosanoff list (1910). As it is quite likely that low frequency stimuli will behave differently to high frequency stimuli the generalisations that can be made from such early studies on response behaviour are limited. The next study, Schmitt (1998a) is included due to it being one of the few studies to have attempted to track word association responses over time. As is argued in Chapter 8, it does not do justice to the incremental nature of vocabulary acquisition to only view snapshots of a learner's lexicon. A lexicon is dynamic by nature with a learner's lexicon being especially

so; words are integrated, reinforced and forgotten all the time as the learner's lexicon continually reorganises itself. If we are to truly understand the complexity of the processes at work, studies such as Schmitt (1998a) which track the development of words over time (and in detail) are likely to lead to enlightenment. The next paper, Orita (2002) is relevant to the experiments presented in this thesis as it is in a Japanese context using participants of a similar background and ability. Despite being exemplary in some respects of how word association studies ought to be done it also highlights some issues that had yet to be adequately addressed: how stimulus words are selected and how responses are classified. Finally in this section we look at a recent, large scale Danish study that includes two measures of word association (Henriksen, 2008). This is particularly interesting as it builds on the work of Schmitt (1998a) in developing a categorisation method that 'scores' word association. To a certain extent it also takes on board the idea that when viewed from an individual perspective, word association data can give important insights into how a learner's lexicon is structured. This is a key theme that we will come back to repeatedly throughout this thesis.

2.3 Politzer 1978

2.3.1 Summary

This study builds on L1 research findings (Brown & Berko, 1960; Entwistle, 1966) that young children give a high proportion of syntagmatic responses and when their lexicon matures (between five and eight) they begin to make more paradigmatic responses. Working on the assumption that the proportion of paradigmatic and syntagmatic responses is indicative of language proficiency Politzer gave word association and language proficiency tests to his French learners. He posed three questions.

> *-How does the ratio of paradigmatic and syntagmatic responses of French students in French compare with the same ratio in their native language (English)?*
>
> *-What is the relation of paradigmatic and syntagmatic responses and achievement in French?*
>
> *-Are there any teaching behaviours which favour the establishment of either paradigmatic or syntagmatic responses by the student?* (Politzer, 1978:204)

The 203 first year French students were given two 20-item word association tests; the items were from various word classes. The first test used French items and the second test used the same items but in English. The tests were two days apart and the order of the

items was randomised for the second test. The responses to these 40 stimulus words were categorized, the important categories being paradigmatic and syntagmatic. As previous L1 studies led the author to expect, the responses to the L2 stimuli showed a syntagmatic dominance (on average 9.7 of the 20 responses were syntagmatic and 5.3 were paradigmatic). The remaining responses (25%) were not classified; Politzer notes that they were a mixture of "clang - purely acoustic" or "non-responses". The responses to the L1 stimuli showed a paradigmatic dominance (on average 14.3 of the 20 responses were paradigmatic and 4.9 syntagmatic). To test the second question the L2 word association responses were correlated with scores from an L2 language test. The results (Table 2.1) show significant correlations between the number of syntagmatic responses and the scores on the Listening, Reading, and Speaking sections of the test. All sections of the test had significant correlations with the number of paradigmatic responses.

Table 2.1 Correlations between L2 responses and language tests (Politzer, 1978)

	Number of syntagmatic responses	Number of paradigmatic responses
Listening	0.26**	0.20**
Reading	0.14*	0.28**
Grammar	0.10	0.34**
Writing	0.10	0.29**
Free writing	0.04	0.26**
Speaking	0.27**	0.21**

**p< .01, *p< .05

To test the third question the L2 responses were correlated with data on teaching behaviour taken from a different study done by the author on the same classes. It was found that some teaching methods, such as dialogue drills generate syntagmatic responses (a correlation of 0.56) and that some methods, such as substitution drills, generate paradigmatic responses (0.55). The translation method was found to be "counterproductive", he argues that the native language "inhibits thinking in the foreign language" and consequently hinders the development of L2 associations. Politzer concludes that L2 beginners' dominant responses are syntagmatic with any paradigmatic responses being associated with the learning of writing and grammar skills. He argues that as learners improve there is a shift in association type and that this can be used as an indicator of L2 proficiency.

2.3.2 Critique of Politzer 1978

In this section I will discuss some serious problems with the evidence that this study presents, focusing particularly on the low correlation values and also the stimulus items used in the word association test. As well as these negative points I will also comment on an interesting (though poorly supported) idea that came out of this study.

A major problem with this widely cited and influential study is that the evidence presented is very weak and offers scant support for the conclusions. In this study Politzer calculates the correlation coefficients between various sets of data. The significant correlations cited as evidence that a particular set of word association responses are linked in various ways to language ability range from 0.14 to 0.34 (Table 2.1). It is natural to question such values, which even within the behavioural sciences are very low. As Cohen et al. (2006:202) explain, the reason that such low correlation values are deemed "statistically significant" is that significance is determined by the number of subjects. In this case the number of subjects (203) is high, meaning that correlations as low as 0.14 have statistical significance. Even though they are statistically significant it would have made more sense to interpret all these values as showing that there was very little relationship between any of the language measures and the responses. Cohen et al. (2006:202) argue that "when correlations are around 0.40, crude group predictions may be possible" but that between the 0.20 and 0.35 range they are "of no value". With the correlations between the word association responses and the teaching behaviours we do see some higher values (the highest being 0.58 between paradigmatic responses and the number of drills per minute) although even these are not particularly convincing.

As well as the low correlation values there are problems with the proficiency measure, one of these is that some parts of the language test are more (or less) syntagmatically biased than others. The listening section for example requires students to produce learned sentence patterns, a syntagmatic task, whereas the reading section requires substitution of items, a paradigmatic task. It therefore seems quite likely that the kind of task influenced the type of response given. A further problem is that each section of the main test only comprises a few questions. The proficiency sub skills (listening, reading, grammar, writing, speaking) were being judged on only a small demonstration of ability in each area. As the reliability of these test data is questionable, we must also view the correlations that they are based on with caution.

The problems with the measures of language proficiency, one half of the correlations, are compounded by problems with the word association measure, the other

half of the correlations. The problem here is with the stimulus items. Firstly, they are not (as stated) a translation of each other, the English list is not an English version of the initial French list. There are some items in the second list that are not in the first, and the item *we* is included twice in the English version! A more serious problem is that the number of items is small (20 per test), although this might not have mattered so much if the selection of items had been a bit better. Unfortunately the selection of items (it is not reported how this was done) is poor and it is this which undermines the study the most. Many of the items included have a very strong primary link to just one other word. As Meara (1983) notes the word *white* for example will generally give *black* in English and *good* will probably give *bad*. Similarly, in this test *blanc* will give *noir*, not only because *blanc* and *noir* are strongly linked, but because both these words are used in the same stimulus list leading to a priming effect. Consequently, a lot of the responses generated from these stimulus words simply show the strength of the link that particular pairs of words have with each other rather than telling us something useful about the response characteristics of the learners. The use of personal pronouns (*we* in the English list and *nous* in the French) are also problematic as French pronouns will give a verb response *nous→sommes* (syntagmatic) but in English it is likely for pronouns to generate responses from the same class, *we* might give *you* (paradigmatic). As there were only 20 items in each list, these problematic stimuli would have seriously skewed the data. A more useful list of stimulus words would have contained not only many more items but items that had been carefully selected so as to give a variety of potential responses (both syntagmatic and paradigmatic). These problems with the stimuli invalidate the measure of learners' response preferences.

Putting aside the serious weaknesses with the study, one idea that does seem to make sense (and had the strongest correlation values in Politzer's data) is that the way we are taught an L2 dictates to a certain extent the kind of characteristic responses that we give. According to Politzer, those who have for example been taught using a method that promotes a lot of memorisation of dialogues can be expected to give a lot of syntagmatic responses to a word association test. Unfortunately the data does not really support this idea, for the reasons given above and also because Politzer did not isolate the language learned through this classroom practice from general language learning. It would have been more reasonable to conclude from his data that spoken language in general encouraged the syntagmatic responses rather than just the dialogues used in these classes. Still, the idea that *how language is taught is reflected in word association responses* does warrant more serious consideration. If we turn the idea around we might even entertain the

notion that a student, who has a tendency to give an unusually high number of syntagmatic responses, might acquire L2 vocabulary more efficiently from doing activities (like memorising dialogues) that reinforce this predisposition.

Despite its many weaknesses this paper was extremely influential, and not in a positive way. Although not the first study of L2 word associations, it was the first to suggest that L2 learners are similar to L1 children in that they make a lot of syntagmatic responses. Consequently it led research down what "turned out to be a blind alley" (Meara 2009:98). As Meara explains, "Politzer was a Big Name in Applied Linguistics at the time, and his results were taken very much at face value by L2 researchers, and not subjected to much critical analysis". However, when these results are subjected to critical analysis they are found to be seriously flawed, lacking both validity and reliability.

2.4 Kruse, Pankhurst & Sharwood-Smith 1987

2.4.1 Summary

Based on a methodology used in Randall (1980) this experiment was conducted in order to establish whether word association generated reliable data and also to assess the viability of using word association tests as a measure of second language proficiency. In the experiment 15 Dutch students of English were asked to make multiple (up to 12) responses to ten stimulus words selected from the Postman and Keppel (1970) norms list. As a control, seven native speakers of English were also asked to do the word association test. In an attempt to limit the effect of word stimulus type, one word was selected from each of ten word categories. The words in the norms list were initially divided into ten categories based on the stereotypy of the native speaker responses. The first category consisted of words with extremely high primary responses, such as *man* which associates with the word *woman* to the exclusion of most other words. The last category (10) consisted of words that are not stereotypical, words that have a very low primary response. The following words were selected: *man, high, sickness, short, fruit, mutton, priest, eating, comfort,* and *anger.* To determine their English proficiency level, each student was given a cloze style English test (students had to fill in 50 words deleted from a text) as well as a grammar test. To measure the reliability of the word association test, the students were asked to repeat the test two weeks later. A high correlation coefficient ($r = >0.8$) between the two data sets was expected in order to confirm its reliability. The computer-administered tests (participants saw the word on the screen for 30 seconds and typed in their responses) were analysed in two ways. The first was to count the number of responses each student made to the

stimulus words. It was expected that as proficiency increased the number of responses would increase. The second was to measure how stereotypical the responses were when compared to the norms list. It was expected that learners would give more varied (less stereotypical) responses than natives. Stereotypy was calculated in two ways, an unweighted measure and a weighted measure. The weighted measure gave more weight to stereotypical responses that were in the same order as the native responses. In the weighted measure each response was given a score (from 12 to 1) depending on how stereotypical it was in the norms list (the most typical response got a score of 12, the second 11, etc.), this was then multiplied by the order in which it appeared in the participant's response (if it was the first response it was multiplied by 12). Therefore if the primary response to a stimulus was also the primary response in the norms list the score for that response would be 144. The scores for each of the responses to the nine words were summed and the averages calculated.

There were three main sets of findings. Firstly (Table 2.2), the mean scores did not show any significant difference between the learners and the native speakers for any of the measures.

Table 2.2 Mean scores, SDs, and theoretical maximum scoring (Kruse et al., 1987)

Test 1			Test 2		Native speaker		
	Mean	SD	Mean	SD	Mean	SD	Theoretical Maximum
A:	76.8	17.9	82.8	19.1	79.9	14.2	108
B:	23.4	7.3	22.9	5.7	25.7	7.2	108
C:	1457	377	1542	337	1509	414	15,552

Test A = Response score Test B = Non-weighted stereotypy Test C= Weighted stereotypy

The second main finding was that when the tests in the first word association test were compared with the second test the correlations (Table 2.3) were below expectations. The authors argue that this demonstrates that the word association measure is unreliable.

Table 2.3 Reliability coefficients on two test sessions (Kruse et al.,1987)

Test A: Response score	r = 0.759	p<0.005
Test B: Non-weighted stereotypy	r = 0.658	p<0.005
Test C: Weighted stereotypy	r = 0.554	p<0.025

As shown in Table 2.4, the third main finding was that there were "disappointing" correlations between the word association scores and the measures of proficiency.

Table 2.4 Correlations between association and proficiency (Kruse et al., 1987)

	Cloze	Grammar monitoring
Test A: Response score	r = 0.441*	r = 0.576**
Test B: Non-weighted stereotypy	r = 0.547**	r = 0.296 ns
Test C: Weighted stereotypy	r = 0.535**	r = 0.147 ns

* p<0.005 ** p<0.025

The authors conclude from these results that word association is not a reliable measure of proficiency and dismiss word association tests in general as they "do not show much promise for the specific role created for them in L2 research".

2.4.2 Critique of Kruse et al. 1987

The first point to make about this paper is that it has had an enormous impact on studies of word association. The impact was a negative one, which is evident from the lack of studies attempting to link word association responses with proficiency for many years after. It was not until much later that researchers (Schmitt & Meara, 1997; Schmitt, 1998a; Wolter, 2002) began to pick up this thread again. Research along the proficiency strand of word association all but dried up for about ten years as a consequence of these findings. A closer inspection of Kruse et al. (1987) however reveals that their findings are problematic from a number of perspectives, and didn't warrant such a dramatic rejection of the word association methodology. Many of these problems stem from uncritically borrowing a methodology from Randall (1980).

The first problem is with the participants. Randall's study (26 learners and 16 natives) was not particularly large but as Kruse et al. only used about half the number of

learners and natives it is difficult to have much confidence in their findings. Notwithstanding the small sample size, the group of learners (third year Dutch university English majors) are not, I would argue, particularly representative of English learners. The general high level of English proficiency in Holland, even with those who don't need it in their profession, is a striking feature of this country. English competence has long been viewed as a required skill by Dutch people, it is a small country surrounded by more dominant languages and cannot survive (economically) with its native language alone. Consequently strong efforts are made by the education system to formally study English. When coupled with an input rich environment (many un-dubbed TV programs and films are aired in English), close proximity to England (geographically and culturally) and many similarities between the two languages, the result is a general high level of English competence. As the learners within this study would be viewed as advanced, even within such a high ability environment, it is difficult to see how these findings apply to more average learners studying in less ideal contexts. When we also consider that the stimulus words used are all fairly frequent (and so would have been very familiar to these advanced learners), it is little wonder that this study failed to find any difference between 'learner' and 'native' responses.

The next problem is with the stimulus items, as has already been noted these are all likely to have been well integrated items within even the learner's lexicons. Again, quantity is an issue, only nine out of the ten original items made it to the analysis stage (as opposed to 50 in Randall, 1980). As well as the small number of items the selection procedure is questionable. The decision to include items that usually elicited responses on a cline from very stereotypical to non-stereotypical is difficult to understand. This supposed control for stereotypy meant that other (more obvious) response variables (frequency, word class, emotionality, concreteness, L1 cognates) were not explicitly accounted for. Consequently stimulus words such as *high* and *short* were included; for reasons explained in Meara (1983) these will probably give their opposites (*low* and *tall*) irrespective of proficiency, and so be of little value. A further problem with these stimulus words is that there is a great deal of similarity for five of them in the L1 and L2, not only in terms of meaning, but also phonology and orthography (Table 2.5). Were there around 50 - 100 items, we might overlook this handful of questionable stimulus items, but as there are only nine this is a serious problem.

Table 2.5 An L1/L2 comparison of five stimulus words used in Kruse et al. (1987)

English (L2)	Dutch (L1)	comment
sickness	ziekte	similar initial sound
short	kort	similar stem sound
fruit	fruit	same spelling with different pronunciation
comfort	comfort	same spelling with different pronunciation
priest	priester	similar spelling

Even if we buy into the idea of selecting stimulus words based on the stereotypy of native responses, choosing from a wide range of stereotypy (ten bands) makes interpreting the results unnecessarily complex. It would have made more sense to select stimulus words that all generated native responses within a similar range of stereotypy. They could have for example chosen only stimulus words that previous studies or norms lists had noted as being very stereotypical. Alternatively they could have gone the other way and only selected words that were not very stereotypical. By having a mix it is difficult to understand the results. We might want to ask whether the responses given are due to the students' proficiency levels or the stereotypical nature of the particular stimulus words used.

This leads us into a related problem, how stereotypy is viewed. The authors categorise a very stereotypical answer (one that matches the primary association in the norms lists) as being the answer which is most native-like and therefore highest scoring. This is also the position taken by Schmitt (1998a). However, it could equally be argued that a non-stereotypical answer represents the highest level of response, as a rare response may be showing a deeper level of understanding than normal. A more recent study (Henriksen, 2008) does indeed take this view. In Henriksen's associative hierarchy the highest score is given to "low frequency, non-canonical but semantically related" responses. For the stimulus *bread* for example Henriksen views responses such as *grainy* or *flour* as demonstrating higher knowledge than a stereotypical (canonical in her terms) response such as *food* or *water*. It might also be noted that Lambert (1956) also assumed lower proficiency learners would exhibit stereotypical responses due to the smaller size of their vocabularies. The view that 'native-like' stereotypical responses represent high proficiency is therefore debatable.

Given these fundamental problems it is actually surprising that the retest data gave

correlations as high as they did, while the authors dismiss them as "disappointing" this interpretation seems rather harsh. If we put aside the flaws highlighted so far and imagine that the data came from a much larger sample of learners responding to a larger and more suitable stimulus list then the correlations of 0.554, 0.658 and 0.759 do, in my view, show moderate relationships. It is unclear where the 0.8 threshold comes from, although it seems particularly strict when considered alongside the low correlation values that Politzer (1978) viewed as "significant". Another good reason for not expecting the tests to correlate so well is that students were instructed to "treat the second session as a separate test, as a fresh start...they were instructed not to try specifically to give the same or different responses." (Kruse et al, 1987:153). By telling students to think back to the responses they gave the first time round the test instructions probably affected the responses somewhat. Although it is unlikely that the students would have remembered many of their initial responses (two weeks before), if they had then this in itself would make any comparison questionable.

Due to the many problems with the way the research was done in this study, it cannot be taken as evidence that word associations are an unreliable measurement tool or that word associations and proficiency are unrelated. Were it not for its continuing influence, the conclusions drawn by Kruse and colleagues might well be ignored. Unfortunately as this paper is still cited as fuel for the argument that word association data is inconsistent (Zareva, 2007; Zareva & Wolter, 2012) it is necessary to explain the main problems in detail and refute the negative image it created.

2.5 Söderman 1993b

2.5.1 Summary

This study aims to see if evidence can be found for the existence of a *syntagmatic - paradigmatic shift* in foreign language learners as their L2 proficiency increases. Söderman's initial assumption, based on L1 word association studies, is that native speakers shift from word association responses that are syntagmatically dominated when they are children to a more mature pattern of paradigmatically dominated responses. As language learners and native children both have immature lexicons the expectation is that less proficient language learners will make more "child-like" syntagmatic responses. The results of two experiments are discussed.

In the first experiment four groups of students studying English in Finland were given 100 stimulus words in a free word association test. The stimulus words, from the

Kent-Rosanoff list (1910), were mostly nouns. There were 28 students in each group of; 7th Graders (3rd year of English study), 2nd Form Gymnasium (7th year of English), 1st year University and Advanced University English Students. The students were asked to write down one response to each of the 100 items, their responses were classified as either:

Paradigmatic	*in the same word form class as the stimulus*
Syntagmatic	*in a different form class to the stimulus*
Clang	*phonetically related but with no semantic relation*
Other	*responses in the L1, responses influenced by the L1, anomalous responses and repetitions*

(Söderman 1993b:157)

It was expected that as proficiency increased there would be fewer *syntagmatic* responses. The results confirmed this, as the age/proficiency of the students increased so did the proportion of paradigmatic responses.

The author was aware that the results, while possibly showing that there was a shift in response type as a function of proficiency/age, could also be a function of the particular words tested. In the second experiment this idea, that it is the actual words themselves that dictate the response type, was tested. There were two kinds of stimuli, high frequency words and low frequency words. In this follow up experiment 28 advanced learners and 28 native speakers were given 64 stimulus words to respond to. Using Hofland and Johansson's 1982 frequency list, 32 frequent words and 32 infrequent words were selected as stimulus items (all but 4 were adjectives). It was hypothesised that:

- *The frequent words would result in a larger proportion of paradigmatic responses than syntagmatic responses and that clang/other responses would be few.*
- *The infrequent words would result in a larger proportion of syntagmatic responses and that the number of clang/other responses would increase.*

The author argues that the results confirmed these two hypotheses, with both native and advanced learners making more paradigmatic responses (62.7% and 52.6% respectively) to the frequent words. With the infrequent words the proportion of paradigmatic responses dropped to 44.3 % for the natives and 30.3% for the advanced learners. It was expected that both groups would respond similarly to the high frequency words but that there would be a difference in response with the less frequent words, again this was confirmed. The small number of paradigmatic responses to the infrequent stimulus words was however unexpected, this conflicted with previous studies. An analysis of variance showed that although there was no significant group effect (F=3.48), the effect of frequency was highly

significant (F=86.54).

Söderman (1993: 168) concludes that although there does seem to be a shift in response type as language proficiency increases she challenges the idea of a syntagmatic-paradigmatic shift "...the more general concept of a shift in response type has been preferred here to the widely accepted term syntagmatic-paradigmatic shift." She also questions the idea that a *shift* occurs at a specific age or phase of learning across the whole mental lexicon but that individual words are likely to shift at different times.

2.5.2 Critique of Söderman 1993b

Considering the two unsatisfactory studies reviewed so far (Politzer, 1978; Kruse et al., 1987) this study marks a move to more rigorously apply and test the findings of L1 studies to the L2 context. An important point to come out of this study is the support it gives to the idea that language proficiency is linked in some way to responses on a word association test; specifically, that more proficient users give a higher proportion of paradigmatic responses. This study adds weight to Politzer's suggestion that L2 responses are similar to the responses given by L1 children. It also stands out as being one of the few studies that specifically examine responses to low frequency stimuli. Another point of interest is that the stimulus items used in the second experiment were mainly adjectives; a word class that has only received limited attention (Piper & Leicester, 1980) in word association studies. In this section I intend to look at three problematic areas, the lack of a test to define the proficiency of the groups in the 1st experiment, the implications of using stimulus lists of mixed word class, and how the stimulus words were selected.

The first problem is the claim that more proficient students give more paradigmatic responses. It seems premature of Söderman to jump to the conclusion that proficiency is linked to the type of response as there was no consistent measure of proficiency. In the second experiment the advanced learner's proficiency was directly measured but in the first experiment it was merely assumed that students who had studied longer were of a higher proficiency. There wasn't a test of English proficiency to confirm that the 7th Graders were indeed at a lower level than the other groups. As the effect of proficiency (F=6) in the 1st experiment is not particularly dramatic, one wonders precisely how much each group differed in terms of their English proficiency. The Gymnasium group and the 1st year University group for example gave virtually the same amount of paradigmatic responses (58% and 58.4% respectively). This may be due to their proficiency being relatively close, but without a measure of proficiency this cannot be confirmed.

The second problem is the decision to use a stimulus set comprising of a variety of word classes in the first experiment and then use a stimulus set comprising of mostly adjectives in the second experiment. Early work in this field by Deese (1965) suggests that word class has an effect on responses, the use of two stimulus sets with widely differing proportions of word class therefore means that the results from the 1st experiment cannot really be viewed alongside the results from the 2nd experiment. There are various reasons to expect different response types from different word classes. Let us consider adjectives versus nouns. An important point is that adjectives are a much smaller group than nouns; therefore the number of potential paradigmatic responses is lower. Due to the large size of the noun word class, when a person is asked to make a response to a noun such as *photograph* there are a number of common synonyms or near synonyms available (*picture, image, snap, portrait*), these would be classified as paradigmatic as they are nouns. There are also a large number of potential syntagmatic associations for *photograph*; such as, *family, group, colour* or *passport*. On the other hand when one is asked to make a response to an adjective such as *sick,* although there are synonyms available (*ailing, ill, feeble, queasy*), they are not nearly so frequent and therefore less likely to be generated. Even though a contrasting response such as *well* (paradigmatic) might be given it should be noted that as the primary function of adjectives is to modify nouns a more likely response to an adjective stimulus would be a noun that collocates strongly with that adjective. In the case of *sick* we might expect a response of *joke* (as in *a sick joke*) or *dog* (as in the idiom *as sick as a dog*), such responses would be classified as syntagmatic. Other likely responses to *sick* would be *doctor, nurse* or *hospital* as they all belong in the same lexical group. These are also nouns so such responses would also be classed as syntagmatic. Given that the stimulus items were mainly adjectives in Söderman's second experiment (94%) it does not seem so surprising that there were more syntagmatic responses when compared to the 1st experiment or previous L1 experiments. Comparing the results of a word association experiment that use a mixed word class stimulus set with the results of an experiment that use a single word class stimulus set is problematic when the general effect of each word class on responses is unknown.

The third problem is how the stimulus words were selected. In the first experiment the Kent & Rosanoff (1910) word list was used in its entirety. In the second experiment a different source had to be used in order to identify low frequency words. It is reported that words occurring more than 50 times per million in Hofland and Johansson's (1982) corpus were considered frequent and words occurring under 10 times per million were considered

low frequency. It ought to be noted that this corpus is now obsolete due to the availability of much larger corpora. The 100 million word British National Corpus (BNC) which was completed a year after Söderman's paper or the massive 425 million word Corpus Of Contemporary American English (COCA), are now generally used to identify the frequencies of words within the English language. Using a large corpus to select words from is logical if one wants to use words of a particular frequency and Söderman deserves credit for making the effort to do this rather than limiting the study to words in existing stimulus lists. The selection of words within the 50 occurrences per million seems quite a broad band though and makes me think that some of them might not be so frequent. Are these words within the top 2000 most frequent words, which researchers such as Nation (2001) would rate as high frequency? Unfortunately Söderman does not report which words were used so we cannot compare them with modern lists and verify that they correspond to what we now generally consider as low and high frequency words.

The two experiments reported in this paper are not without problems but are interesting from a number of angles. There is an attempt to explore responses to low frequency items and also adjectives, previously these groups of stimuli had been neglected in favour of high frequency noun stimuli. A surprising point about this paper is that it only has a small number of citations (21 Google Scholar citations as of May 2013); this is partly due to the fact that there are two very similar papers in circulation. The other paper, published in the same year (Söderman 1993a) is also based on data collected for Söderman's 1992 PhD thesis. As a consequence, the apparent paucity of citations is misleading as they are split between the two collections in which the research is published. The other reason for this fairly large research project only receiving limited attention is that both papers were published in obscure, difficult to obtain collections.

2.6 Schmitt 1998a

2.6.1 Summary

This longitudinal study attempts to identify the stages learners go through when they learn individual words and whether it is possible to make links between certain types of word knowledge. One of the main goals was to identify a developmental hierarchy for word knowledge types. To do this Schmitt tested the depth of knowledge that three international students at a British University had for 11 target words over one year. Polysemous words covering a range of frequency levels and word classes were included in the study. The cohort was tested three times during the year, every six months (T1, T2 and T3), to see

how their knowledge of these words improved. The testing session involved a two-hour interview with each student, consisting of tasks that would elicit demonstrations of four types of word knowledge. The demonstrations of word knowledge for these 11 words by the three students were compared against data collected from native speakers. The four types of word knowledge were:

Spelling A four-point scale was used: 0 indicating no knowledge and 3 indicating perfect spelling.

Associations Three responses were asked for each word. A four-point scale was used to score each set of responses:
0 - none of the responses were on the norms lists
1 - some responses were on the norms list, although infrequent
2 - responses were similar to responses on the norms list
3 - all responses were in the top half of the norms list.

Grammatical information Students were asked to indicate the word class of each item and then conjugate them into the other word classes. A four point scale was used: 0 indicating no understanding of the word class and 3 indicating students understood its word class and could make all the other potential word classes.

Meaning A meaning proportion was calculated for each word which took into account the various meanings each word had and whether the student could generate a meaning *productively* or merely to a *receptive* level (needing a hint).
0 indicated no meanings were known despite hints.
0.5 indicated that there was some combination of partial/half/full known meanings.
0.8 indicated that students could explain most meanings without prompting.
1.0 indicated all word meanings were fully explained without the need for hints.

In order to avoid the problematic high frequency words, the target words were selected from the University Word List. The words chosen ranged from relatively unknown to well known:

brood, spur, (relatively unknown)
circulate, convert, launch, plot, trace
abandon, dedicate, illuminate, suspend (well known)

These words were piloted with similar students to see if in fact they were good examples of 'well known' or 'unknown' words and also to determine how long it would take to test knowledge of these words.

The main results were that even for advanced students studying at a British university their knowledge of the different meaning senses for these words was

surprisingly limited; the meaning proportion for each word was usually below 0.5 and rarely exceeded the 0.8 threshold. With 74 cases of an improvement in the meaning sense as opposed to 29 cases of the meaning sense regressing, there was some overall progression. It ought to be noted though that the vast majority of cases (263) remained stable, it was quite common for student's knowledge of the meanings of these words to remain unchanged. The measure of spelling did not reveal so much, these students were of a high enough level that they could spell most words based on phonology, even when they were unsure of the meaning. For those words that were initially misspelt, the data does however show an improvement over time. With the word association measure the responses generally became more native-like over time. Of the 33 cases of association (3 learners x 11 words) 23 cases showed "stability or progress" with only 10 cases becoming less native-like, "backsliding". The general progression in association scores was reflected by similar increases in their meaning scores and grammar scores.

Schmitt (1998a: 309) concluded that although "no evidence of a developmental hierarchy for word knowledge types" could be seen from this study, some of the knowledge types seemed to be inter-related. Another point to come out of this study was that over a six-month period students didn't forget words that had been learned to a productive level. However words that had been learned to a receptive level were sometimes forgotten. He also found that the verb and noun form of words seemed to be better retained than adjective and adverb forms and therefore suggested that adjectives and adverbs require more explicit study.

2.6.2 Critique of Schmitt 1998a

The design of this study stands out from other studies in vocabulary acquisition for a number of reasons. Firstly, rather than trying to make generalisations about the global knowledge of a large group of learners' lexicons, it tracks the development of just a few words with a few students and measure their development intensively from a variety of angles. As with recent studies that attempted a similar approach (Churchill, 2007; Crossley et al, 2009), detailed and rich information is obtained for each word. Of course, this depth of information is at the expense of breadth. Such studies do not really allow us to extrapolate the data and make broad generalisations about particular kinds of learners or particular kinds of words, but then that is not their intention. By closely examining the progress a learner makes with a word we can compare numerous aspects of word knowledge development and build up a detailed picture of that learner's understanding for

that item. Each additional aspect that is measured helps us to improve the resolution of this picture and allows us to more precisely identify how far the individual has progressed along the word knowledge continuum (from first encounter with a new word until full acquisition) at each testing point. The second striking feature of this study is that it is longitudinal, given that vocabulary is learned in an incremental fashion, this is in some ways more appropriate than the snapshot style experiments that have so far been exemplified. The third remarkable point is the development of a native norms list against which to compare the learner word association responses, as opposed to the use of pre-existing norms lists such as the Minnesota Word Association Norms (Postman & Keppel, 1970). Using lists such as this, developed from samples with little relation to language learners, is therefore questionable. Schmitt's norms list, developed from responses by British university students, increases the validity of his study as it represents the native community within which these particular students were studying. As well as these positive points, two further areas will be addressed, one of them relating to the study as a whole and the other to the word association component.

The first point is with the length of the study. While I agree that words are learned incrementally and therefore a longitudinal study is appropriate, it is unfortunate that this study stops after one year. In such a short period I would not expect (even for students studying at a university within the target environment) to meet these fairly infrequent words enough times; I would therefore not expect a noticeable improvement in knowledge for these words. Unsurprisingly, with a gap of just six months between testing there was not a lot of change detected in students' word knowledge. As Schmitt himself notes (1998:300) there is considerable "inertia" to overcome when learning words, acquisition is a slow process. Had T1, T2 and T3 been at intervals of a year rather than six months we might have expected more substantial changes. Future studies taking up this longitudinal approach to word development would be well advised to allow greater gaps between testing sessions. As well as giving students more opportunities for word growth between test times, an experiment conducted over a longer period may also be more reliable. A concern I have with Schmitt's experiment is that the close attention paid to these 11 words in the initial testing sessions is likely to have been responsible for some of the development that is reported in later testing sessions. The influence of prior test sessions on the learner's word knowledge would be less of an issue with a larger gap between test points and would allow greater confidence to be put in the claims made about word development.

With the word association component there are two areas of interest. The first of

these is the scoring system, explained in greater detail in Schmitt (1998b). This system assumes that a response that is more frequent in the norms list demonstrates a higher proficiency than a response that is less frequent in the norms list. A basic problem with this assumption is that the ultimate goal for many learners is not to become native-like. As Meara notes in his 1983 paper, learners are actually moving towards becoming bilingual so should probably be compared with highly proficient bilinguals. Even if we accept that comparison with native speakers is a valid measure of associative competence it is debatable whether it is valid to score one response higher than another due to its frequency in a norms list. It seems just as likely to me that a lower frequency response indicates a similar (or higher) level of understanding of the word. Had the learners (and natives in the norms list) been instructed to *give the response that you think most people would give* then perhaps we would have a good reason to grade the more idiosyncratic responses lower than high frequency responses. However, the instructions were "Please give the first three words you think of when you hear the word_____" (Schmitt, 1998: 294). In these instructions there is no indication that the respondents ought to try to give stereotypical responses, I would therefore argue that an idiosyncratic response (as long as it is meaningful) does not necessarily represent a lower level of competence. To exemplify this let's imagine that Person A and Person B give the following responses to the stimulus word *plot*:

Person A: *plot* →*plan, land, play*

Person B: *plot* →*thickens, insurrection, scatter-graph*

For the purposes of this example (and in the absence of the data from Schmitt's norms list) I will use data from The Edinburgh Associative Thesaurus (Kiss et al., 1973) as a guide. According to the EAT norms list *plan, land* and *play* are all in the upper half of the native norms list, *thickens* however is ranked 17th, *insurrection* 35th and *scatter-graph* is off list. Using Schmitt's scheme we would therefore award Person A three points for this word and Person B one point. I would however argue that Person B probably has a greater knowledge of *plot* as he is demonstrating that he can use this word in the fairly abstract collocation *the plot thickens*, knows a rare synonym and is aware of its use in mathematics. Although Person B does not respond with the more common responses, there is no evidence that he doesn't know them. In fact, if he can respond to *plot* in this way I would assume that he does know these more common links. Given the lack of any requirement in the task instructions to respond in a stereotypical way Person B seems to be unfairly penalised for making less frequent associations. This problem with the scoring system is

well demonstrated by the fact that of the three native speaker responses (the control) 6% of their responses were judged to have been *un-native-like* and 12% of responses were judged as having only *a minimal amount of NS-like associations.* It is hard to understand how a native speaker response can ever be categorised as *un-native-like*; all responses by a native speaker are (by definition) native-like.

The second issue that the word association measure raises is the decision to ask for three responses to each stimulus word. Eliciting multiple word association responses does have advantages; it gives participants a better chance to demonstrate their knowledge of an item. For learners, many of these items are only partially known and so by only asking for one response we may be denying them a fair chance to display their knowledge. A problem with asking participants to give a single response to a prompt word is that it may fail to pick up on partially integrated words. As Schmitt puts it "multiple responses better capture the richness of a subject's association network". Wolter (2002) puts a similar argument forward in his attempt to develop a word association test as a measure of proficiency. It should be noted though that limiting the choice of responses to three is an arbitrary one. For the learner who gives untypical responses to a stimulus word, such as Person B in the example above, we cannot discount the possibility that this learner would have given highly stereotypical responses on the 4^{th} and 5^{th} attempt. Another disadvantage to this method is that by focusing on an item for an extended period of time we have the possibility of the respondent's initial response influencing subsequent responses: chaining.

Even though the scoring system in the word association measure is questionable, this study is noteworthy for a number of positive reasons. Not only did Schmitt develop an innovative methodology, but his approach to data collection was careful and well planned. Each measurement tool was trialled prior to the main experiment and the target words (though limited in number) were selected on a principled basis. Through attacking research questions that large group studies have failed to solve with experimental designs that dig deeper into the vocabulary knowledge of a few individuals, new insights can be gained. Another important contribution that this paper makes is that it demonstrates the possibility of measuring multiple levels of word knowledge at the same time, thus giving a deep understanding of how well someone knows a word. As prior to this paper measurements of word knowledge had merely skimmed the surface, measuring one or perhaps two word knowledge types, this is I believe a significant development.

2.7 Orita 2002

2.7.1 Summary

In this study Orita is interested in the changes in response type as proficiency increases. Five groups were given 60 high frequency stimulus words, the stimuli were an equal number of nouns, adjectives and verbs. There were: 74 novice low students (ages 13-14), 79 novice high students (ages 16-17), 71 non English major university - intermediate students (ages 19-21), 73 English major university - advanced students and 53 native English adults. The stimuli were read aloud and the students had to write down the first word they thought of. The responses were classified as syntagmatic, paradigmatic, phonological, other or no-response. The study attempted to answer the following questions:

> *Are syntagmatic - paradigmatic shifts found in the whole results as English proficiency advances?*
> *Do particular words undergo idiosyncratic development unrelated to English proficiency?*
> *For the native speakers, do any words produce an exceptionally large number of syntagmatic or phonological responses?*
> *For the least proficient, do any words produce an exceptionally large number of paradigmatic responses?* (Orita, 2002: 113)

The results showed that, as expected, the number of no-responses was high with the least proficient group (18%), lower with the two higher ability groups and decreased to a negligible 0.2% with the native speakers. Some evidence was found for a syntagmatic – paradigmatic shift across the proficiency groups. The responses did not distinguish between the three lowest ability groups although they did show a shift in response patterns from the lower proficiency to the advanced group and also from the advanced to the native group. The lowest group gave similar proportions of responses (around 66% syntagmatic and 29% paradigmatic) which then changed to 60.3% syntagmatic and 36.9% paradigmatic with the advanced group and 50.2% syntagmatic and 47% paradigmatic with the natives.

To answer the second question, each stimulus word was re-categorised depending on the proportion of responses that it received (Table 2.6). There were four categories:

Standard	a similar ratio of syntagmatic and paradigmatic responses to the overall results
Divergent	different to the response patterns of natives
Other	not standard or divergent
No differ	the proportion of responses in each category were similar for each group.

Table 2.6 The number of each stimulus type (Orita, 2002)

	Number of words
Standard	20
Divergent	9
Other	15
No differ	16
Total	60

Of the 60 words, around half of them (Standard and Divergent) showed a shift in line with expectations whereas the other half did not (Other and No-diff). The author concludes that "not every word follows the same path or undergoes a shift; indeed, individual words seem to develop in their own way".

In answer to question three, although phonological responses were rare for the natives in this study, four words were identified which produced an exceptionally high number of syntagmatic responses; *jump, ball, window,* and *sky.* Another three stimulus words *mother, dog* and *Sunday* were also identified as eliciting very high numbers of paradigmatic responses (usually one strong primary response) regardless of proficiency level.

2.7.2 Critique of Orita 2002

The key point to come out of this study is the finding that some stimulus words generate responses that are good indicators of proficiency whereas some words (such as *father, dog, Sunday, jump, ball, window, sky* in this study) do not. Stimulus selection is therefore an important step in the design of any word association study. Due to the frequency of the stimulus used, from this study alone we cannot extend this statement to words beyond the 1000 most frequent word range. It is possible that stimulus word selection will be less of an issue with lower frequency items, the need for careful selection of very high frequency words seems clear though. Despite the positive points to come out of this study, such as the large number of subjects (276) and stimulus items (60) and controlling for the frequency and word class of the stimuli, there are two areas of concern. The first is with the classification of the responses and the second is with how the proficiency of the students was assessed.

The first problem is with the initial classification of responses into two broad groups (syntagmatic and paradigmatic); it is unclear how Orita knew precisely what respondents were thinking when they made their responses. *Sky* for example is reported to have elicited a lot of stereotypical responses: *blue, cloud, high, fly, limit, bird.* Orita

classifies all these as syntagmatic responses, and this may well be the case for many of them as they all collocate very strongly with the stimulus word. However, it seems to me that some of these could equally have been classified as phonological, *high* and *fly* for example both rhyme with *sky*. It is also conceivable that both phonological and collocational factors could have been equally responsible for some of these responses. In another example Orita reports that, regardless of the group, most people (63%) gave the stereotypical response of *father* to the stimulus *mother*. These responses were all interpreted as paradigmatic, presumably because *mother* is a co-ordinate of *father*. As with the *sky* example, this classification is again quite likely to be correct, although we cannot discount the possibility that some of these responses could have been *syntagmatic*. When we search for these two items in a concordancing program based on a large corpus, such as COCA (Davies, 2008), we find that other than the grammatical items *My* and *Her* these two words have the strongest collocation. Sentences such as *There is no easy way to tell a mother and father that their child is dead.* are not uncommon and can be found in a wide variety of genres. Correct classification is of crucial importance, as the initial stage of the analysis also impacts on the second re-classification into "standard" or "divergent" responses. Although Orita used two judges, it is not clear how this would have helped in ambiguous cases, such as *mother→father*. It is likely that (as they were both working to the same classification procedures) both the judges would have misinterpreted the responses in the same way. Without some kind of introspective check (such as an interview or questionnaire) it is hard to say for certain what these learners were thinking when they made their responses.

The second point of concern is that no common test of language ability was given to the participants of this study. Given that the main purpose of this study is to see how responses to words change as student proficiency increases, the lack of a direct measure of ability is difficult to understand. The proficiency of the participants within each group was assumed to be roughly the same based on how long they had studied English and a questionnaire asking about how they had fared on other tests of general English ability such as TOEFL or TOEIC. While this may have been a reasonable way to roughly divide the groups it would have been preferable to have had a specific measure of vocabulary that could have been applied to all the groups. Nation's Vocabulary Levels Test (1983) or perhaps his more recent Vocabulary Size Test (Nation & Beglar, 2007) would seem fit for the purpose as they specifically focus on measuring vocabulary knowledge and would be capable of testing the full range of students in the study (beginner to advanced). Use of a

standardized test such as the VLT would also allow easier comparison with other studies and facilitate replication.

An additional test might also have tried to measure students' ability to recognise and use the specific stimulus words. It is not obvious that all students within a particular age range (length of study) would have had a similar level of knowledge for each of the words; so this ought to have been explicitly tested. Even if we accept Orita's assumption that students within for example the intermediate group (around ten years of formal English study) were similar in their general English abilities they probably all had different learning backgrounds. There would have been various paths towards such an "intermediate" level; some would have studied abroad, some not, some would have watched a lot of English movies or read a lot and some not. Each student would have experienced the stimulus words in differing contexts and differing amounts. For some words a student would have had the full range of receptive and productive knowledge that Nation details (2001:27) but for other words many of these word aspects might not yet have been fully acquired. A measure such as the Vocabulary Knowledge Test (Wesche & Paribakht, 1996) could have been used to test the depth of knowledge that each student had for each word. Although time consuming (especially for a group as large as Orita's) this would have allowed for more precise comparisons between word association responses and proficiency.

Orita's study has many good points, particularly in terms of the number of students, the number of stimulus items used and the consideration given to the frequency and word class of the stimulus items. This study is not however entirely without problems, there is a query over whether all the responses were correctly classified and also with how the proficiency of the participants was determined. Such problems are not serious enough to dismiss the study outright; they do however cast a shadow over the conclusions that come out of it.

2.8 Henriksen 2008

2.8.1 Summary

Henriksen's study of declarative lexical knowledge was part of a larger project (Albrechtsen et al., 2008) conducted in Denmark that examined how language is processed in both the L1 and L2. In the main project three researchers focused on different areas of language learning: declarative lexical knowledge, lexical inferencing and writing. For each of these areas the researchers looked at how the same Danish students processed their L2,

English. There were three age levels, grade 7 (three years English study), grade 10 (six years study) and grade 13 (nine years study). There were 30 students at each level and over the period of two weeks they were given a series of language tests. One distinctive aspect of the project as a whole was the use of think-aloud methodology combined with retrospective analysis by the participants in an attempt to "get as close as is presently possible to what goes on in our informants minds" (Henriksen, 2008: 10). Another important point about this project was that each of the language tests had a parallel activity in the learners L1, allowing direct comparison between L1 and L2 performance.

In the section on declarative lexical knowledge there were two measures, a productive measure (a free word association test - WAT) and a receptive measure (word correction test – WCT). In the WAT, 48 items (24 nouns and 24 adjectives) were read out to the students at 15 second intervals and they were instructed to write down the first two words that came to mind. The WAT used a novel categorisation system based on three fundamental points identified in prior word association studies. As language proficiency increases:

there is a shift from form to meaning based responses
there is an increase in canonical responses (*bread→butter*)
there is an increase in low frequency responses

The reason for abandoning the traditional Paradigmatic / Syntagmatic classification system was due to concerns with coding. Henriksen (2008: 46) notes it is "extremely difficult to find clear and objective criteria for categorising a specific response". As well as using a new classification method, the data responses were scored, when added together these scores gave an "overall word association score" that showed how native-like each individual's responses were (see Table 2.7). This was a development of the scoring system initially used in Schmitt (1998a). In Henriksen's scoring system it ought to be noted that the term *canonical* is preferred to the term *stereotypical* that is usually used in the literature.

Table 2.7 The categorisation and scoring system (Henriksen, 2008)

Main Category	Sub Category	Example for stimulus bread	score
Lack of a form or semantic link (unqualified)	Empty		0
	Repetition	bread	0
	Translation	brød	0
	Ragbag	paper	0
Form	Formal	red	1
	Chain	table	1
Semantic link	High frequency non-canonical	white, birds	2
	High frequency canonical	food, water	3
	Low frequency canonical	toast, loaf	4
	Low frequency non-canonical	grainy, flour	5

The criteria for deciding whether a response was canonical or not was based on data collected from native speakers. 127 UK university students and 108 Danish university students were recruited to make the norm lists, each group could be viewed as consisting of proficient users as all students were studying their respective native language as their major subject. For an English response to be classified as *low frequency* it needed to be a word which was beyond the 5000 most frequent words in the British National Corpus, the same threshold was used with a Danish corpus to classify the Danish responses as either high or low frequency.

The final point concerning the methodology is that the WAT was followed by a retrospective task (20 minutes) where students were asked to expand/qualify their choice of responses in order to aid classification. As can be seen in Table 2.8 the WAT results confirmed the expectation that students would get higher scores in their L1 than their L2, the youngest group for example averaged 217.89 (from a possible 240) in their first language and 151.83 in English. There was also a clear progression across the ages tested, as students got older their scores increased in both their L1 and their L2.

Table 2.8 Results of two measures of declarative lexical knowledge. (Henriksen, 2008)

			G7	G10	G13
Overall Word Association Score (Max: 240)	L2	Mean	151.83	208.65	221.77
		SD	49.76	25.26	28.95
	L1	Mean	217.89	238.65	239.17
		SD	24.23	19.82	31.85
Word Connection Score	L2	Mean	64.89%	71.29%	70.60%
		SD	8.15	4.6	5.36
	L1	Mean	71.09%	74.83%	74.46%
		SD	6.89	3.44	5.04

Looking at the details, the results show the youngest students gave a large number of unqualified responses in both their L1 (5.54%) and L2 (21.42%). With the older students there was a decrease in this type of response (the 10th graders gave 1.76% unqualified responses in their L1 and 6.23% in their L2) and a corresponding increase in the number of semantically related responses. It was also found that the older students gave more low frequency responses to the stimuli than the younger students. These results were predicted from previous free word association studies, which led Henriksen to conclude that her word association score effectively differentiated between the L1 and L2 and was also sensitive enough to detect changes to the learners' lexical networks as they become more proficient in both languages.

In the receptive measure – the word connection task (WCT), 24 stimulus words were given and for each of these words participants were instructed to indicate the five strongest links between the stimulus and ten potential associates. The stimulus words were the first half of the word list used in the WAT, 12 nouns and 12 adjectives. Comparing the L1 and L2 data (Table 2.8) the results were as predicted, the learners made more 'correct' links in their native language than in English; grade 7 for example made 64.89% of the possible links in English and 71.09% in Danish. Unlike the WAT the WCT measure did not show a clear progression with age. It was able to distinguish between the most proficient and least proficient groups, however it could not distinguish between the grade 10 and grade 13 groups. Although it was expected that the 10th graders would not be able to make as many canonical links as the 13th graders the data did not show any significant difference between the two groups.

The results of the lexical knowledge study were compared to some of the other measures in the project and also general measures of language proficiency, these findings were mixed. Strong correlations were found between the L2 WAT scores and reading

ability for the grade 7 (r=0.749) and grade 10 students (r=0.722) with moderate correlations (r=0.492) being found for the grade 13 students. The English WAT data also correlated positively with the Vocabulary Level's Test (Nation 2001) for all the age groups. Contrary to expectations though, there were no significant correlations between the groups' scores in the L1 writing tasks and the L1 WAT. In the L2 WAT only the 7th grade scores showed any significant correlations with the scores derived from the L2 writing task. Another surprising finding was that the WAT data did not correlate with the WCT data. As both tests used the same stimuli it was expected that students who got high word association scores would also make a lot of connections in the WCT test.

2.8.2 Critique of Henriksen 2008

In this section I initially intend to focus on some of the positive aspects of Henriksen's research into lexical network development, look at some problems with the stimulus words used and explore issues surrounding the word association score she proposes.

Firstly, this project as a whole has many strong points, the attempt to collect and compare data from both the student's L1 and L2 at three different ages and also the broad scope of the project which used tests designed to measure different knowledge aspects of the same words being particularly striking. There is a wide acceptance in the literature (Meara, 1983; Wesche & Paribakht, 1996; Schmitt, 1998a; Nation 2001) that in order to be confident that someone knows a word it is not enough to merely ask for a translation into their L1 or ask the person to give a synonym for that word. For us to confidently say a word is *known* there needs to be a demonstration of various kinds of knowledge for that word, such as: what it means, how it is pronounced, how it is spelt, what words it usually collocates with and what restrictions there are on its use. The aspects of knowledge framework put forward by Richards (1976) and then developed further by Nation (2001) suggests that to be fully competent with a word, not only do we need a receptive understanding but we also need to be able to use it productively. Any attempt at measuring all the 18 aspects that these studies identify for a representative sample of words is, as Meara & Wolter (2004: 88) note, "fundamentally doomed" due to logistics. They calculate that measuring 50 words in such a comprehensive way might require 600 test items. It ought to be noted that, as some of these aspects are inter-related it is probably not necessary to measure them all to get a good understanding of how well someone knows a word. We might speculate that measuring a few of these aspects that are fundamentally different might be enough. Other than Schmitt (1998a) who measured four aspects, there

are unfortunately few studies which take on the challenge of developing a way of measuring multiple aspects of word knowledge. Consequently projects involving a considerable number of participants (140) and a decent sample of words (48) such as this which go beyond measuring only one aspect of word knowledge are to be applauded.

This project is also notable for another important reason. With the attempt to develop lexical profiles of individuals' language ability from the battery of tests that each student was given we can see a move away from analysing language development in terms of group norms. Other than Fitzpatrick (2007, 2009) there are few studies that attempt to look at network development from this individual perspective. There are however exceptions, such as Churchill (2007) and Meara (2011). Meara for example calculated the number of connections within his own mental lexicon over a six month period at various frequency levels. In that study he used a computer program to generate 5000 random word pairs from the JACET (2003) word list, he then judged whether there was a link or not between the pairs. As Fitzpatrick (2007) notes, the analysis of word association data from a group perspective has produced a lot of conflicting findings, which have not led to a clear picture of how the mental lexicon is structured. Case-study style investigations such as this, which attempt to analyse vocabulary development and lexical growth from an individual perspective, are more likely to further our understanding of how lexicons are structured and how they develop. It should be noted though that as the design of Henriksen's study is essentially cross-sectional, with the main WAT and WCT findings *grouped* by age, it cannot be argued that Henriksen and colleagues are analysing their data from a truly *individual* perspective.

An important part of Henriksen's contribution to the project is the "overall word association score", a measure of how developed a person's lexical network has become. The fact that the overall word association score correlated well with both Nation's Levels Test and the reading measure is a positive sign. If her productive word association score does prove to be reliable it is likely to be a considerable breakthrough, it should be noted though that it is not entirely unproblematic. The fact that it did not correlate well with the writing test is however a cause for concern. As she is arguing that her WAT is in itself a measure of competence, it would seem logical to expect that it would correlate fairly well with more established measures of productive ability.

The most likely source of the inconsistent findings in this paper are the stimulus words. One problem is that only high frequency words were used. Many of these have very strong primary links, such as *woman,* so it is not so surprising that the WAT was found to

be less effective at predicting language ability than had been hoped. Even though highly proficient respondents have the ability to respond to the word *woman* with lower frequency words (higher scoring) such as *feminine* or *dowager* they do not usually do this. According to The Edinburgh Associative Thesaurus norms list (Kiss et al., 1973) *man* is given 59% of the time to the cue word *woman* by native speakers. There is little reason to expect the higher ability students to respond to an item such as *woman* with lower frequency words, as this is something that even native speakers rarely do. As has already been argued in the discussion of Schmitt (1998a), the practice of assigning a score for a word association response based on how stereotypical it is on a norms list is problematic. Interestingly, in Henriksen's system it is the unusual responses that get the highest score, as opposed to Schmitt's system that awards the highest score to the most stereotypical responses.

Another problem with the stimulus words is their high frequency. If the stimulus words had been lower frequency or the list had been cross checked against a norms list such as EAT (and words with extremely strong primary connections replaced with words that had the potential to elicit a range of responses) then the WAT would have been better at distinguishing between respondents of differing proficiency. As it is, of the 48 stimulus words over half of them (according to the EAT data norms list) have >25% of their associations to just one other word with 6 of the 48 having >50% of their associations to just one word. In Meara & Fitzpatrick (2000) a threshold of 15% was applied to filter out stimulus words that are strongly linked to just one word. Were we to apply this stricter limit to Henriksen's list then we would only be left with nine words to analyse! While we cannot assume that L1 norms lists will correctly identify all of the words which will be unhelpful in L2 research, it seems prudent to replace those words that the native norms lists identify as having extremely strong ties to just one word. The stimulus words Henriksen used were taken from the Kent-Rosanoff list developed over a hundred years ago (1910), to diagnose psychological problems. One benefit of using this list is that studies (Postman & Keppel, 1970) have already compiled native speaker norms from these words – a convenient benchmark against with which to compare responses and measure native-likeness. As Henriksen did not make use of such a norms list, but developed her own for each language, there seems little value in selecting from this source. As discussed in Meara (1983) this list has a number of drawbacks, it would have therefore been more logical to have selected from a word list devised for linguistic purposes (e.g. BNC or COCA; Davies, 2008). These sources provide a wider range of potential stimulus words to choose from and can be searched for particular frequency ranges or word classes. Had she selected from a

wider range of frequencies then I think this would also have helped the WAT measure to distinguish between the different proficiencies examined. One might also speculate that another reason the WAT measure did not correlate well with the measure of writing was because the WAT stimuli were very familiar to all proficiency levels, these were words that were probably all well integrated parts of the students' lexicons. Even at the lower levels of proficiency there would have been few peripheral (newly integrated) words, consequently responses to this easy task were often of a similar type. The writing task on the other hand was a more open activity where students had the opportunity to work at a level closer to limit of their ability: it was harder and therefore more discriminating.

A final point about the stimulus words is that they were uneven in terms of word class, there were 24 nouns and 24 adjectives but no other word classes. Considering a paper by the same author (Bagger-Nissen & Henriksen, 2006), which specifically examined the role of word class (nouns, adjectives and verbs) in word association tests, it seems odd that the stimuli were unbalanced in this respect. The main findings of the 2006 paper were that in both the L1 and L2 word class had a significant effect, "nouns trigger more paradigmatic responses than adjectives". As there is also other evidence to suggest word class has an effect (Deese 1965 and Entwistle 1966 mention this in their L1 studies) we might have expected Henriksen to have included a stimulus list more representative of the proportion of word classes in the languages involved (e.g. 20 nouns, 15 verbs, 15 adjectives, 5 adverbs and 3 prepositions). To have achieved such a mix it would have been necessary to select from a broader source than the Kent-Rosanoff list, as this list is mostly (60%) nouns. Alternatively the study could have been limited to just one word class (e.g. 48 nouns) where the bias could be accounted for.

An issue discussed in the review of Schmitt (1998a), that strongly influenced the approach taken in these experiments, was how to create a score to represent associative development. A problem not touched upon in that section, is that when allocating a single score to a series of responses the details become obscured. If we imagine that two students both score 180 points in the WAT it is probable that they would have achieved this score in different ways. One student could have made 36 *low frequency non-canonical* responses (5 points each) and then translated the rest (0 points). The second student could have made a variety of responses, 6 *high frequent canonical,* 11 *low frequent canonical,* 21 *high frequent non-canonical* and 10 *low frequent non-canonical.* The score in itself does not give us a very detailed picture of the individual's response characteristics and it seems misleading to judge both students as being similar in terms of the development of their

mental lexicons. This criticism can also be applied to others who have also attempted to use composite word association scores (Kruse et al.,1987; Wolter, 2002). However, as Henriksen uses these scores alongside other lexical measures within a larger vocabulary proficiency framework this is less of an issue. The different perspectives on the depth and breadth of an individual's vocabulary knowledge gained from the other measures give context to the WAT score. When viewed as part of a battery of tests that combine to give a broad view of an individual's grasp of vocabulary Henriksen's overall word association score is I think meaningful.

With this study, Henriksen makes a useful contribution to a project impressive in terms of size, breadth, its theoretical underpinnings and the innovative measures it employs. While Henriksen's classification system, WAT score and method of analysis have great potential they will need to be trialled using a better selection of stimulus words before any confident claims can be made about them. The main problem with the word association measure, which seriously undermines the findings, is the stimulus items used.

2.9 *Type* response strand

In the *type* strand we begin with Sökmen (1993), which along with Söderman (1993a) came out when there was little research being done using word associations: following the negative conclusions of Kruse et al. (1987). Sökmen (1993) is interesting due to an innovative categorisation system and the direct application of word association research findings to pedagogy. While Politzer (1978) also attempted to apply word association findings to the teaching context as has already been argued, his study is far from convincing. Next we examine a paper by Wolter (2001) that addresses some of the common problems that researchers face with word associations. Following Wolter (2001), Bagger-Nissen & Henriksen (2006) is considered, in this study it is argued that the word class of the stimulus is an important variable in determining responses. The effect of word class is taken up in Chapters 4 – 7, this paper therefore forms an important backdrop to the experimental work.

An alternative approach is provided by Fitzpatrick 2007, in this paper she explores word association data from an altogether different perspective, focusing on individuals rather than groups. The "individual profiling" approach proposed offers a way forward in a field that in recent years had virtually ground to a halt due to a series of conflicting findings. While the experimental work within this particular paper is concerned with native speakers, it is part of a series (Fitzpatrick 2006, 2007, 2009) that examines the responses of

L2 learners and native speakers. An exploration of the methodology and alternative approach to data analysis from this series of papers forms a large part of this thesis. Following a review of Fitzpatrick's work we look at one more recent contribution to the discussion on word association (Zareva, 2011) that identifies some key gaps in the research. This paper is unusual in that it does not fall so neatly into the type or proficiency strands that have so far been identified but attempts to give attention to both. Zareva (2011) explores variables that are assumed to affect the type of response (word class and word frequency) alongside proficiency.

2.10 Sökmen 1993

2.10.1 Summary

Using a 50 item free word association test, Sökmen elicited the responses of 198 ESL learners studying in America. These learners were of mixed ability and originated from various countries; 108 Japanese, 18 Korean, 16 Chinese, 13 Arabic, and 43 others. The stimulus words were selected from the Kent-Rosanoff (1910) list and consisted of 30 nouns, 19 adjectives and 1 verb. The purpose of this study was to demonstrate how patterns in associations might be used by teachers to help students acquire words more effectively. The main question she asked was "Which associations are useful to teach?" The responses were analysed on various levels. Initially they were divided into eight "word class" categories, see Table 2.9. As shown in the 'Examples' column, *beautiful* in response to *woman* would not be considered as a collocation (as in, a beautiful woman); collocations being defined as those constructions that are made from left to right, such as *street→car*.

The main finding from this analysis was that *affective* responses dominated - for these learners they accounted for 47% of the 9049 responses. Collocations (17%) and contrasts (12%) were also significant categories.

The second analysis was by parts of speech. The findings were that nouns generally elicit nouns (68.36%) confirming previous studies such as Deese (1965), and also that adjectives and verbs are more likely to stimulate responses which form syntactic units, such as *deep→kiss* or *eating→rice*. The responses were also analysed by how often nouns, adjectives and verbs elicited the responses from the word class categories in Table 2.9. This revealed that noun stimuli usually generate *affective* responses (56.35%), adjectives usually generate either *collocations* or *affective* responses (35% each) and that verbs also generate *collocations* and *affective* responses (47% and 40% respectively).

Table 2.9 Classification by word class categories (Sökmen, 1993)

Word class categories	Explanation	Examples	No of responses
Affective	a visual image, opinion, emotional response or personal past experience	*dark→scared, sickness→hospital*	4,284
Collocations	words which commonly go together from left to right (not from right to left)	*street→car* ~~*woman→beautiful*~~	1,540
Contrasts	Opposites	*quick→slow, doctor→patient*	1,157
Coordinates	words in equal rank and importance	*bath→shower salt→sugar*	839
Supra/subordinate classifications	words that show category relationships up or down	*fruit→apple, bread→food*	652
Synonyms	words with similar meanings	*boy→guy*	474
Nonsense	the coder could not determine the relationship	*scissors→honesty*	76
Word forms		*sickness→sick, deep→depth*	27

The third level of analysis was to compare the first three responses to the Minnesota word norms (Postman & Keppel, 1970). The responses matched well, 90% of the responses were in the top three of the norms list and 48% of these had exactly the same primary response. For example, both the learners and the norms list gave *hot* as the top response to *cold*. She concludes from this that native norms "could be useful for planning vocabulary teaching for ESL students".

Finally the responses were analysed in terms of the learners' backgrounds. Age was found to have no real effect, although gender, ESL level, education and language background did. When analysed by a t-test, men were found to be more likely to give *verb* responses and women more likely to give a native *primary* response or an *adjective* response. When comparing the three ability levels, beginners were seen to give more *contrasts* than the intermediate groups who in turn gave more *contrasts* than the advanced students. There was also an increase in *verb* and also *affective* responses as proficiency increased. With regards to their background; Chinese students gave relatively high numbers of *verb* responses but fewer *collocation* or *noun* responses, whereas Arab speakers gave a lot of *classifications*. Both Japanese and Korean learners gave few *verb* responses. It was also shown that the most educated students were more likely to give *word form* responses.

Due to the overall dominance of *affective* responses (which increased with proficiency), Sökmen argues that language teachers ought to develop activities that promote emotional or personal associations to facilitate the acquisition of words. She also notes that, in general, giving synonyms or learning word forms is less useful than other tasks such as collocational or contrasting activities. She concludes the paper by speculating that words might be more effectively taught if the teacher capitalizes on the results that came out of the analysis of learners' gender and background. For example, men could initially be made to focus on verb associations and women on adjective associations. Similarly, Chinese speakers might benefit from being initially taught which verbs associate with a word whereas Arab speakers might benefit from being taught how the word fits into the supra/subordinate hierarchy.

2.10.2 Critique of Sökmen 1993

Some of the early L2 studies (Politzer 1978; Kruse et al., 1987) were unsatisfactory due to a small number of test items and/or a small number of participants. As with another study from this era (Söderman, 1993a) a more suitable number of students (198) was sampled and a decent number of stimulus items (50) used. Another positive feature is that Sökmen developed her own classification system; she did not simply assume that the system developed for L1 studies (the syntagmatic - paradigmatic division) would be applicable to the L2 context. By using a classification system that has been designed to fit the purpose of the study (measuring L2 word associations) an effort has been made to increase validity. Another interesting point is the attempt to apply the findings directly to pedagogy. Other than Politzer (1978) most word association studies do not usually explain how language teachers might use their findings. That said however there are still some concerns with the classification system, the choice of stimulus words and how the proficiency of the learners was measured, these three points will be discussed below.

Within her categorisation system there are a couple of problematic categories. The first of these is the *affective* response category, responses which were made due to "a visual image, opinion or personal past experience". The fact that such a high percentage of responses (47%) were put in this category immediately suggests that something is not quite right. A simple explanation for this category being so dominant might be that it was less restrictive than others and so perhaps functioned as a waste bin for the responses that didn't fit nicely into the more clearly defined categories (such as *collocations* or *contrasts*). I imagine that in many cases it would have been difficult for the coder to be sure that

learners were making "personal" responses, based only on the word association data. As the students were not asked to elaborate on their responses the coder would have had to rely on intuition, an educated guess based on knowledge of the students involved and other local information. The thinking behind some of these responses might have been obvious; people's names, local places or perhaps responses relating to a popular TV personality. However, it is not unlikely that some associations would have been categorised as *affective* due to an over imaginative interpretation on the part of the coder. For example with the stimulus/response of *table→study* (Sökmen classified this as *affective)* she was assuming that the respondent was thinking along the lines of *Last night I studied at the table in my bedroom.* Clearly there are other potential interpretations, *table* and *study* could be seen as part of a lexical set of words (study, book, pencil, table, chair) that many people would cluster together as *coordinates*. Alternatively they could be seen as belonging to a hierarchy; *education → school → study → table → book.* The student could also have been thinking that these words collocate (study-table) as they might collocate in this way in their L1. Given that both words are homonyms it is in fact very difficult to be sure of what the student was thinking and (unlikely as it is) we cannot rule out the possibility that the student didn't know the word *table* and made a lucky guess. Without some kind of introspective measure (asking the respondent to reconstruct their thoughts, either verbally or by questionnaire) we are in very subjective territory with this particular category. A teacher who knows her students well, may be able to easily see the link and correctly classify it, but (as in this case) when students are from a variety of backgrounds the likelihood of misclassification is high.

The other worrying category is the *collocation* category, which only allows left to right collocations. Although such a tight restriction makes responses easier to assign, it leaves out a lot of other potential collocations such as *moon→blue* (that collocate right to left) or more distant collocations that are parts of larger formulaic units (*quiet→mouse*, from the idiom *as quiet as a mouse*). Following work by Pawley & Syder (1983) and Wray (2002) collocations and formulaic language are now thought of "as being as important as individual words" (Schmitt, 2010). A categorisation system that cannot deal with responses derived from multi-word units or right to left collocations is therefore questionable. If such responses were not categorised as *collocations* though we have to wonder where they were put, it seems likely that many ended up in the *affective* category. Given the importance of categorising items correctly it is rather surprising that there was no trial of this new system prior to the main experiment. A pilot study which included an introspective measure (or

used a group whose response behaviour was known) would have helped identify poorly defined categories, such as the affective category. It would have also provided a more principled method of selecting stimulus items than the random selection approach that was adopted. This brings us to the next point of concern, the choice of stimulus items.

As many of the problems with the stimulus items stem from using the Kent Rosanoff list as a source, commented on in Chapter 1 and the critique of Henriksen (2008), many of the points brought out in those discussion also apply here. Suffice to say, Sökmen's finding that the norms list matched the learner responses is unsurprising. Had lower frequency items been used, more consideration given to potentially problematic words (L2 cognates, homonyms and words which generate very stereotypical responses) then the findings might have been quite different. Consequently we cannot read much into Sökmen's assertion that native norms list can be used to help plan L2 studies, unless of course the students are of such a low level that the teacher is actually teaching the words on the Kent-Rosanoff list.

Another problem with choosing stimulus words from this list, not covered in other discussions but particularly relevant here, is that it has an unequal number of items per word class. Consisting of 60% nouns and 25% adjectives there are only six verbs from which to choose. Of these, only one verb (eating) made it into Sökmen's stimulus list, which puts into question the reliability of her analysis by parts of speech. As there were only 172 responses to this verb, as opposed to the 3,408 responses to adjectives and 4792 responses to nouns, her conclusions (especially those concerning responses to verbs) are questionable. A further confounding factor is that of the verbs available *eating* was a particularly unfortunate choice as this can also be used as an adjective: an *eating apple* as opposed to a *cooking apple*. One wonders why this particular analysis was even attempted given the obvious inequality of the lists and the fact that the study was already quite complex. The problem of using unequal group sizes can also be seen within the analysis of learner backgrounds. Most students were Japanese (55%) with the other nationalities all being under 10%, this means statements such as "Arabic speakers may respond better to vocabulary taught with verb associations" are not well supported. Again, as this level of analysis (although interesting) is not essential to the main question it might have been better to have left the effects of learner background to a follow up study. If for example Sökmen had simply concentrated on the Japanese learners (108 is still a good size) the stimulus words could have been fairly easily screened for possible L1 cognates and (as there is only one culture to consider) the coder would stand a better chance of correctly categorizing the more ambiguous responses.

The final point of concern is with how proficiency was determined. Within the study there were no tests of general language proficiency or vocabulary to confirm the three ability levels. We can assume that none of the students were complete beginners (the University of Washington web page stipulates students must have an IELTS score of 5.5 to enter the pre-course English program) although beyond that it is difficult to know what their levels were. We might speculate that some kind of placement test was given to divide students into these groups but this is not reported and so it seems proficiency was determined by length of study. As many of these learners would have begun their studies at different levels (and common sense tells us that some people learn faster than others) this does not seem satisfactory and precludes an exact replication or comparison with similar studies.

Despite reservations on how some responses were classified, the selection of stimulus words and how learner ability was measured this study does give some impetus to the idea that word association tests can be used as a pedagogical tool. Given that this study is quite good in terms of numbers, the fact that the stimulus words were poorly chosen (and that some responses may have been miscategorised) does not necessarily lead us to reject the findings. The concerns raised in this discussion do however mean that Sökmen's conclusions need to be viewed with caution.

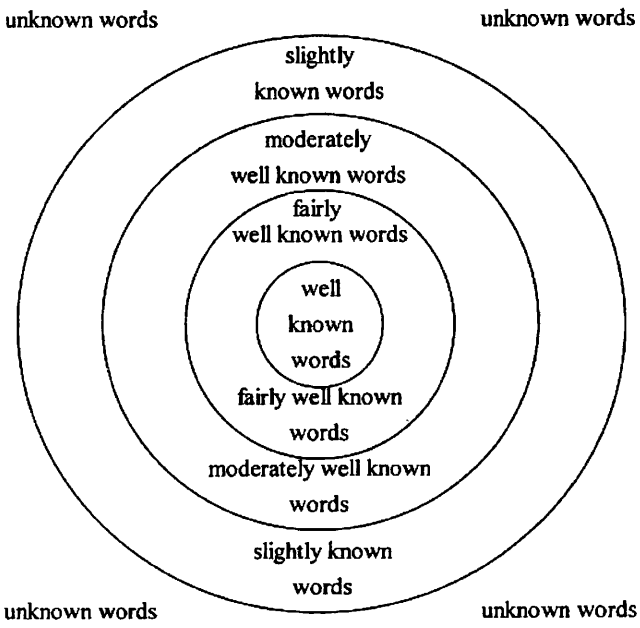
2.11 Wolter 2001

2.11.1 Summary

This study introduces and tests a model of the mental lexicon - the Depth of Individual Word Knowledge model (DIWK) - that claims to be able to accommodate both L1 and L2 lexicons. DIWK is based on a number of similarities that previous studies have consistently shown these two types of lexicon to have. It also builds on the assumption that a key factor in how well words are integrated into the lexicon is how well they are known. If words are relatively unknown it is assumed they only have a limited number of links to other words in the network and if they are well known they will have a greater number and more complex set of connections to other words. The model was tested by examining two things. Firstly, how a person associates a set of stimulus words (whether responses are paradigmatic, syntagmatic or phonological), and secondly by measuring the depth of knowledge that individual has for each of these words. The DIWK model (Fig 2.1) predicts that words at the *core* of an individual's lexicon will have strong paradigmatic links, those at the *periphery* will have strong syntagmatic links and those *slightly known* words in the

outer rings will have strong phonological links. The reason for assuming that the well known 'core' words will have strong paradigmatic associations comes from L1 studies (Ervin, 1961; Entwisle 1966) and also L2 studies (Piper & Leicester, 1980; Söderman, 1993).

Fig 2.1 The depth of individual word knowledge model (Wolter, 2001:48)



To test the model, word association tests were given orally to 9 native English speakers (NS's) and 12 non-native speakers (NNS's). Their depth of vocabulary knowledge was also assessed by an oral interview, respondents were asked to rate their own level of knowledge for each word and (if they knew them) give examples of how these words might be used. The five point Vocabulary Knowledge Scale (VKS) developed by Wesche & Paribakht (1996) was used to measure word knowledge. The "well known words" in DIWK correspond to a VKS score of V, the word can be used in a sentence. The "fairly well known words" correspond to a VKS score of IV, a synonym or translation for the word can be made. The "moderately well known words" correspond to a VKS score of III, the respondent *thinks* they know a synonym or translation but is not confident. The "slightly known words" are words that have been heard but the meaning has been forgotten (VKS level II). The area outside of the rings, the "unknown words", corresponds to VKS I "I don't remember having heard this word before".

There were two related hypotheses:

The L2 mental lexicon is structurally similar to the L1 mental lexicon of a native speaker.

Depth of word knowledge is a key component for determining the degree of integration for the individual words that make up the structure of both the L1 and the L2 mental lexicon.

The study tested both how NS's and NNS's responded to prompt words selected from the Bank of English Corpus, a wide range of word frequencies were used. The NNS's were given a list that contained 45 words. Three words (a noun, an adjective and a verb) were selected at 500 word intervals from between 1000 - 8600 of the most frequent words. It was assumed that some of these words would be well known, some would be on the fringes of the learner's knowledge and some not known at all. The NS's were also given this same prompt word list (PWL1) to respond to and in addition given another word list (PWL2) which contained 45 words selected at regular intervals from the 9000 – 39,000 frequency range. As with the NNS's the NS's were confronted with a variety of words, these were later classified as: well known, only partially known and completely unknown.

Wolter concluded that there was not enough evidence to support the notion that L1 and L2 mental lexicons are structurally similar for words that are well known. For example, with words classified as VKS V, *well known words*, NSs gave 48.9% paradigmatic responses whereas NNSs gave 35.4% paradigmatic responses. At the other extreme, words that were *not well known* did however show similar patterns, both groups in the VKS I and II categories showed no statistical difference in their tendency to produce a lot of clang responses. With words classified as VKS II, *slightly known words*, the native speakers made 78.1% clang responses and the NNSs made 65.4% clang responses. Therefore it is argued that phonology plays an important role when words are only partially known, a role which seems to decrease as word knowledge improves. Wolter concludes that the first hypothesis did not get a clear answer.

The second hypothesis that depth of word knowledge is an important component for determining the degree of integration for individual words into the mental lexicon is supported by the results. If we consider the clang responses that were so dominant with the unknown and vaguely known words (VKS I & II) they can be seen to gradually disappear at the expense of more meaningful responses (paradigmatic and syntagmatic) as depth of knowledge increases (VKS IV &V).

An interesting finding was that overall syntagmatic responses were more numerous

than previous studies would have predicted, which led the author to challenge the traditional idea of a syntagmatic – paradigmatic shift. He proposes that rather than thinking of a shift from syntagmatic to paradigmatic responses as the lexicon develops we ought to be thinking of a shift from "semantically meaningless responses to semantically meaningful responses".

2.11.2 Critique of Wolter 2001

In this section there are two main areas discussed, some positive aspects of the study, and on a more negative note, problems with the experimental work that support the claims made in favour of the proposed model.

This paper has many strong points, which explains why it has become extremely influential (148 hits in Google scholar as of May 2013). Perhaps the main reason for this is that Wolter gives hope to what Schmitt (2010:36) calls the "Holy Grail of vocabulary studies", the search for a model which can explain the broad structure of the mental lexicon. The lack of such a model has hindered research in psycholinguistics for decades. The development of a model which gives a realistic view of how words are integrated into a lexicon would represent a major breakthrough. An interesting point to note about Wolter's model is that it shares a feature of other long-standing models; it is uncluttered and therefore easy to understand at a glance. One classic model that looks very similar to DIWK is Burgess' concentric ring model (1924), a model of how cities are organised. It might be argued that models of this kind oversimplify complex structures such as mental lexicons or cities. DIWK for example does not show explicitly how words progress through their various stages, from *unknown* to *well known*. That said though, the simplicity of these kinds of models give them an enduring appeal.

As well as the introduction of DIWK this study is marked by its use of both high and low frequency stimulus words. A limitation on how the findings of previous studies can be generalised is that they usually only use high frequency stimuli. As Wolter notes, the assumption that the patterns observed using high frequency stimuli will be similar to low frequency stimuli does not seem sound. Most research using word association tests has been restricted to high frequency stimuli, often selected from standard lists. One widely used list is the Kent-Rosonoff list (1910), initially developed as a diagnostic tool for clinical psychologists. Notwithstanding its age or original purpose, as noted in Chapter 1, Meara (1983) gives three specific reasons for abandoning the use of this list when measuring the response patterns of L2 users. Wolter's decision to avoid lists such as the

Kent-Rosonoff list and make his own based on a principled selection criteria (no L2 cognates, no words that can be used in more than one word class etc.) therefore give his stimulus words greater validity. As noted by Meara (1983:31) "tried and trusted tools which work for L1 situations are rarely wholly appropriate for L2 situations".

Another important point this paper makes is to challenge the idea that there is a shift in how the lexicon is structured when it reaches a certain point of maturity, the so called syntagmatic – paradigmatic shift. The obsession with this idea is evident in much of the L1 and early L2 word association literature and detracts attention from other equally interesting issues. Wolter makes a strong case for putting aside the seemingly fruitless task of proving whether the syntagmatic- paradigmatic shift is a real phenomenon or not and concentrating on other issues, such as how individual words integrate into the lexicon. An integration that seems more likely to occur at different rates for different words, rather than a holistic or across-the-board *shift* in how people think. If the model in this paper is indeed robust enough to sufficiently explain how words in both the L1 and L2 are integrated into the lexicon then Wolter can be credited as having made a significant contribution. Unfortunately it seems a little premature to hail this paper in such terms due to questions concerning the reliability of the methodology used to test the model. There are I believe two main problems, the size of the samples and the method of categorisation.

The biggest problem with this study is with the size of the sample data. A sample of nine seems too small to make generalisations about the response patterns of native speakers. Similarly, 12 is not a representative group of non-native speakers. Considering the growing evidence (Wilks & Meara, 2007; Fitzpatrick, 2007) that there are a wide variety of behaviour types (even within NS groups) I would not expect a sample as small as this to give stable results. Wolter reports that in his pilot study there was an individual who persisted in giving unusual responses, a native speaker who gave predominantly syntagmatic responses. As this individual was not part of the main study his responses were not included in analysis. However, had this individual's responses been included, then given the small sample size, the results would have been quite different. It may in fact be the case that syntagmatically dominated native speakers are more common than we think. From a sample of nine we cannot really estimate how common these syntagmatically dominant respondents are or make a decision to exclude (or not) such individuals as exceptions. If on the other hand, around 100 native speakers had been sampled and still only one showed this syntagmatic response characteristic we would have a more valid reason to ignore this individual. The issue of small sample sizes, Deese (1965)

recommends a minimum of 50 for L1 studies, brings us to a related problem with the methodology, it is very time consuming. It is likely to take up to an hour per individual using Wolter's oral data collection approach, consequently attempting a sample of 50 - 100 would be a major undertaking. An experimental design that allowed more than one participant per hour to be evaluated would enable more participants to be included and thus improve reliability.

The second main area of concern is with the categorisation system. A positive point concerning this is that Wolter makes an attempt to define and set out detailed procedures for classifying responses as either *paradigmatic, syntagmatic, clang* or *no response*. This is an improvement from studies such as Söderman (1993a), which typically define paradigmatic responses in simple terms such as "words that are in the same word class", and then assume that everything left over is syntagmatic. Despite the clear definitions, some responses were still found to be difficult to categorise as it was often possible to make a connection (albeit a distant one) between a stimulus and response where none was intended. Wolter for example experienced problems with how to classify 'tolerate' as a response to 'confine', the decision to classify this response as neither paradigmatic or syntagmatic was ultimately a subjective one. As with Schmitt (1998a) and Orita (2002), he tackled this problem of subjectivity by using two judges; as two heads are thought to be better than one, this seems a sensible solution. That said though, the second rater's opinion is also subjective and while it might agree with the first it is very likely (especially when second guessing learners from a different cultural background) that they are both misinterpretations. Just as two wrongs don't make a right, two subjective opinions don't necessarily equate with objectivity. A further problem with using such broad categories is that many of the details become obscured. Within the paradigmatic category we might for example be curious to know how many responses were synonyms and how many were not synonymous but had some other semantic association. Unfortunately these broad categories do not allow us to sift through the finer grains and pick out such interesting details.

Despite the potentially ground breaking nature of the model proposed in this paper the experimental work to support this model is disappointing. As outlined above the main concerns stem from the small number of participants and the difficulties in accurately classifying responses.

2.12 Bagger-Nissen & Henriksen 2006

2.12.1 Summary

In the conclusion to Miller and Fellbaum's (1991: 227) description of the how the mental lexicon is organised they note "striking differences in the relational structure for words in different semantic categories". Added to this, native speaker word association studies (Deese, 1965; Entwistle, 1966) have long argued that the word class of the stimulus word has an effect on the type of response. Such studies suggest that when nouns are used as stimulus words in a word association test they produce a disproportionately high number of paradigmatic responses. Given then that there are good reasons to suspect word class will have some kind of an effect on word association responses in general, this study looks at whether this is the case for non-native speakers. As well as exploring the effect of word class, this study also re-examined the assumption that the adult L1 mental lexicon is paradigmatically structured whereas the L2 mental lexicon is predominantly syntagmatically structured. Three hypotheses were made:

> *In the L1 word association test, the proportion of paradigmatic responses will be larger than the proportion of syntagmatic responses.*
>
> *In the L2 test, the proportion of syntagmatic responses will be larger than the proportion of paradigmatic responses.*
>
> *Nouns will elicit more paradigmatic responses than verbs and adjectives in both tests.*
> (Bagger-Nissen & Henriksen, 2006: 391)

To test these hypotheses 25 Danish high school students were asked to make two written associative responses to 45 stimulus words within 20 minutes. They did this in both their native language (Danish) and their second language (English). Stimulus words were equally selected from three word classes (noun, verb and adjective), 15 items per word class. The frequency levels were controlled by choosing words from the 2000 and 3000 word levels of Nation's Vocabulary Levels Test (Schmitt, 2000; Nation, 2001). The tests were repeated a month later, students who initially took the test in their L1 took the second test in their L2, and vice-versa. Responses were classified using the traditional broad classification system of; paradigmatic, syntagmatic, phonological or other.

The main findings (summarized in Table 2.10) show that contrary to the expectations of previous L1 studies there was a preference for these Danish students to give mainly syntagmatic responses when responding in their L1, hypothesis 1 was therefore rejected. As high school age students' L1 mental lexicons should be fairly mature, this leads the authors to question the syntagmatic – paradigmatic shift. The second

hypothesis, that students will produce more syntagmatic responses in their L2, was accepted, the results match other studies (Wolter, 2001). The third hypothesis, that nouns would trigger paradigmatic responses was partially supported.

Table 2.10 Responses to two 45 item word association tests in L1 and L2

N=25	Paradigmatic responses			Syntagmatic responses		
	noun	adjective	verb	noun	adjective	verb
L1 (Danish)	43.9%	29.3%	25.1%	43.5%	58.8%	59.7%
L1 combined	**32.8%**			**54%**		
L2 (English)	28.5%	23.5%	15.5%	43%	51.5%	43.6%
L2 combined	**22.5%**			**46%**		

(adapted from Bagger-Nissen & Henriksen, 2006)

In the L1 test, nouns (43.9%) were over represented within the paradigmatic responses category; if they were equal we would have expected them to be nearer to the combined value of 32.8%. With the syntagmatic responses, nouns were under represented (43.5%) which is again far from the combined value of 54%. With the L2 data though the proportions of word classes were more equally represented, although it might be noted that there were less paradigmatic responses to the verbs than expected. This study concludes that the word class of the stimulus has an effect on the type of response that is generated.

2.12.2 Critique of Bagger-Nissen & Henriksen 2006

This experiment claims to support the assertion that the word class of a stimulus word has an effect on the type of response given in a word association test. It also casts doubt over the concept of a syntagmatic-paradigmatic shift. These findings cannot however be accepted due a series of problems in how the tests were implemented, the classification of responses, the number of items used in the test and the selection of stimulus items.

The first issue I wish to look at concerns the task given to the students. To write 90 associations to a list of words in 20 minutes seems a simple enough task in an L1 but in an L2 when many of these words are likely to be partially known, this is quite a challenge. Using the English version of the list I tried this WA test in my own L1 (English) and managed it in a little under 17 minutes, I think I would be hard pressed to achieve this in my L2. I would probably waste time worrying about how to spell words or dither over

some of the words that I didn't know very well. At best I imagine the students in this study struggled to achieve the task in the time available, having to skip a few (the data shows an increase of 10% in the No Response category from their attempt in their L1). At worst the test was completed in such a rush that most of the stimulus words were not properly read before the response was given. If the words were all within the ability range of the students then the high rate of No Response and Other (31.2%) in the L2 responses does indicate a sense of panic. If the responses were ill considered due to time pressures then this puts into question the whole L2 data set. Although it is understandable to attempt to complete this tedious (for the students) task as quickly as possible, the harder task requirement of responding in an L2 should have been recognised and a more reasonable time period given.

The second issue is that as well as the L2 word association task being hard to complete in the time given, there is also the general problem of classification. The problem of correctly classifying responses into these broad categories is mentioned by other authors (Meara, 1983; Wolter, 2001; Fitzpatrick, 2006) and has already been discussed in the critique of Orita (2002). If the classification of responses is unclear then this casts another shadow over the experiment. Unfortunately the authors do not give us a precise figure for how many of the responses were "difficult to determine", as a consequence we cannot judge whether this is a serious flaw or merely a minor irritation. Unlike Wolter (2001) who used retrospective interviews to help categorize responses, in this study there is no such check, the reliability of the classification procedure is therefore questionable.

This brings us to the third problem, the small number of stimulus words used. At first glance we might think that 45 stimulus words ought to be a reasonable sample of a learner's lexicon, and for the results concerning the responses in general this might be so. In fact as the test asked for two responses per item we actually have 90 potential responses for each student in both their L1 and L2. This encourages us to believe the generalisations are based on a reasonably large data set. For the main part of the analysis though, the word class data, the results are divided into nouns, verbs and adjectives. Of the 30 potential responses in each subgroup, for reasons undisclosed, just the first responses were analysed. This means that each subgroup in Table 2.10 is represented by only 15 responses. The seemingly large differences between the groups are not nearly so impressive when we consider the numbers involved. For example, the biggest difference in the paradigmatic L1 data set (43.9% nouns as opposed to 25.1% verbs) is only a difference of eight responses. Other significant differences, such as L2 paradigmatic responses (28.5% nouns as opposed to 15.5% verbs) represent a difference of only six responses. It seems odd in fact that both

the first and second responses were not used in the analysis, if as mentioned "no major differences were observed" then a more convincing set of results (containing double the responses) could have been presented. Alternatively, if it were the first response that they were mainly interested in why not just ask for one response in the first place? This would have given the researchers the opportunity to include more test items and strengthen the results in this way.

The final problem is with the stimulus words selected. These words were selected from Nation's Vocabulary Levels Test, the intention being to control for word frequency. Unfortunately, picking words from this very small pool of items (120) led to a number of unsatisfactory items being included. Specifically, many of the items are liable to give highly stereotypical responses. When we look at native responses to these items (Table 2.11) we can see the links for a dozen of these words are particularly strong, especially the adjectives (over half the adjectives used have strong links to just one word).

Table 2.11 Native norms for 12 stimuli used in Bagger-Nissen & Henriksen (2006)

Stimulus Word !	**EAT primary response** !!	**% of primary response** !!
slow	fast	43
small	large	26 (+2nd response *big* = 46)
bitter	sweet	45
first	last	51
thin	thick	34
fast	slow	52
happy	sad	32
loud	soft	30 (+2nd response *noise* = 52)
pupil	eye	33 (+2nd response *teacher* = 56)
motor	car	72
birth	death	34
bake	cake	40

! stimuli used in Bagger-Nissen & Henriksen (2006)
!! response data from The Edinburgh Associative Thesaurus (Kiss et al.,1973)

Such items are unhelpful as they do not tell us anything about the response characteristics of the test takers, they merely confirm what we already know about the words (i.e. that *first* is strongly associated to *last*). The results might have been quite different if all the stimuli had the potential to elicit a range of responses, the stimulus *meet* for example was a good choice as there are a number of potential words that often associate with it (both syntagmatically and paradigmatically). While we cannot simply assume that stimuli that have high rates of stereotypy for native speakers will also generate stereotypical responses

with learners, it is likely that some of these will. If as suspected, many of the responses were of this type then it would raise serious questions over the validity of the experiment, especially with regard to the responses to adjectives. Whether the Danish equivalents for these words also have strong links to just one other word is unclear. Meara (1983) found that French words also have this tendency and suggests that this is probably the case for many European languages.

As well as the problem that many of the stimulus words are liable to elicit stereotypical responses there are a number of other problems with the stimulus words. When we look at Table 2.11 we can see that two of the stimulus items, *slow* and *fast,* are an adjective pair, *slow* usually elicits *fast* and vice-versa. The use of both of them in the same test is therefore particularly unfortunate. Meeting *slow* early on in the test primes the learner for the stimulus *fast*; such priming makes it even more likely that the response to *fast* will be *slow*. Other words (*dust, savage, solution, hunger, empty*) are problematic from another point of view, they can function in more than one word class so it might be unclear which is being responded to. Homonyms such as *pupil* and *solution* are also potentially problematic as the rater may be unsure of which of the two meanings is being responded to.

Given the many problems, especially with the number and quality of the stimulus items used, it is difficult to place any confidence in the findings of this study. The question of how much of an effect word class has on word association responses therefore remains unsettled.

2.13 Fitzpatrick 2007

2.13.1 Summary

Using a methodology first proposed in Fitzpatrick 2006, Fitzpatrick used free word association responses to create individual profiles for 30 native English speakers. The main aim was to explore an assumption that underpins many of the previous studies that look into how lexicons are organised. The assumption is that native speakers respond to word association tasks in a predictable, homogenous way.

Two research questions were addressed:

> *Do native speakers respond to cue words in predictable, homogenous ways?*
> *Do individual native speakers respond to cue words in a consistent way?*
> (Fitzpatrick, 2007: 323)

In the experiment subjects were given two sets of cue words (100 items per set) a week apart and asked to write the first word that each cue word brought to mind. The cue words were selected from the Academic wordlist (Coxhead, 2000), this was done in order to

avoid many of the words that have strong cue strength (high frequency words and concrete nouns, *bread* for example usually gives *butter*). The responses were classified into three main categories; Meaning based, Position-based and Form-based. A fourth category Erratic was also used for associative responses that could not be categorized. These main groups were further subdivided into ten subcategories. The ten subcategory system was a modified version of the 16-subcategory system trialled in Fitzpatrick 2006 and recently verified in a replication study (Racine, 2012).

The responses were analysed in terms of how much individuals varied from the mean response behaviour of the group and also how much individuals varied between the two word association tasks. High correlations were calculated between the response profiles of individuals' responses to Task 1 and their response profiles to Task 2, of the 30 subjects 22 had correlations of 0.9 or greater. In order to measure how close the individuals' profiles were the Euclidean distance between them was calculated. The Euclidean distance of each subjects' two profiles were found to be significantly closer than any of the other possible combinations (870) of Task 1 profiles and Task 2 profiles. To answer the first research question, the results showed that native speakers couldn't be considered predictable or homogenous as individual responses varied widely from the mean of the group responses. In answer to the second research question, the results showed that individuals responded to the second set of cue words in much the same way as the first, individuals responded consistently.

Fitzpatrick concludes that native speaker groups should not be considered as homogenous although individual response behaviour is internally consistent. She speculates that if analysed from an individual rather than a group perspective, L2 response behaviours may also be consistent.

2.13.2 Critique of Fitzpatrick 2007

Viewed alongside its predecessor (Fitzpatrick, 2006) this paper marks a break from the methodology followed by previous word association studies that had for decades produced inconsistent and conflicting results. It is particularly marked by its use of a cue word list selected on well-founded principles, a more precise categorization system and an analysis of the data from the perspective of the individual as opposed to the group. As well as these positive points there are areas that are not clearly reported, the first is how the 16 classifications used in the 2006 study became ten, and also why in this study a retrospective interview was not undertaken. Another unusual aspect of this study is the use

of Euclidean distance as a measure of the closeness of the profiles. These three areas will be discussed further.

Firstly, let us consider the categorization system. In her 2006 study Fitzpatrick introduces a 16-point categorisation system, far more detailed than the three way classification system (Syntagmatic, Paradigmatic and Phonological) generally used in other word association studies. While it is natural for Fitzpatrick to want to refine this new system, discarding unhelpful categories and tailoring the categorisation system to suit the experiment, this paper does not make clear how or why the 16 sub categories were whittled down to ten.

Table 2.12 Comparison of classification categories used in Fitzpatrick 2006 & 2007

Category	Sub category	Definition	2006 study	2007 study
Meaning-based association	Defining synonym	x means the same as y	Yes	Yes
	Specific synonym	x can mean y in some specific contexts	Yes	Yes
	Hierarchical/Lexical set relationship	x and y are in the same lexical set or are coordinates or have a meronymous or super-ordinate relationship	Yes	Yes
	Quality association	y is a quality of x or x is a quality of y	Yes	No
	Conceptual context association	x and y have some other conceptual link	Yes	Yes
Position–based association	Consecutive xy collocation	y follows x directly (includes compounds)	Yes	Yes
	Consecutive yx collocation	y precedes x directly (includes compounds)	Yes	Yes
	phrasal xy collocation	y follows x in a phrase but with other content word(s) in between	Yes	No
	phrasal yx collocation	y precedes x in a phrase but with other content word(s) in between	Yes	No
	Different word class collocate	y collocates with x + affix	Yes	No
	Other collocation association	y follows/precedes x in a phrase but with other content word(s) in between	No	Yes
Form based association	Derivational affix difference	y is x plus or minus derivational affix	Yes	No
	Inflectional affix difference	y is x plus or minus inflectional affix	Yes	No
	Change of affix	y is x plus and/or minus a prefix or suffix	No	Yes
	Similar form only	y looks or sounds similar to x but has no decipherable link	Yes	Yes
	Similar form association	y is an associate of a word with a similar form to x	Yes	No
Erratic/ Other association	False cognate	y is related to a false cognate of x in the L1	Yes	No
	No link /Blank	y has no decipherable link to x or no response given	Yes	Yes
Total number of subcategories			**16**	**10**

As can be seen in Table 2.12 the two classification systems have the same broad categories and many of the subcategories are the same. In the conclusion to her 2006 paper Fitzpatrick does tell us "some categories, such as those which attracted very few responses or where response behaviour was very similar, might be merged in future studies". It appears that some categories used in the 2006 study were indeed merged into larger categories. We can assume that the *phrasal xy* and *phrasal yx* and *differ*ent *word class collocate* categories were rolled up to form the *other collocations* category. Similarly the *derivational affix difference* and *inflectional affix difference* categories became the *change of affix* category. Another category omitted from the 2007 study is the *false cognate* category; this is understandable as this study only deals with the subjects' L1. The *quality association* category was presumably cut due to the very low number of responses in that category in the 2006 study, such responses were most likely classified as *conceptual context* associations, although this is not made clear. What is also not clear is why the *similar form association* category was omitted. Again we might guess that such responses were subsequently categorized as *similar form.* When the two classification systems are laid side by side (Table 2.12) it is possible to work out how the changes in categories came about. However, given the importance of the categorisation system within this new methodology it seems odd that Fitzpatrick gives no explicit explanation as to how this classification system evolved.

As with the lack of explanation of how the category system was modified there is also a lack of explanation as to why a retrospective interview was not used: the second point of discussion. This is surprising as in her 2006 study Fitzpatrick specifically notes that:

> *...by retaining the interview component of the experiment, we can ensure that very few responses are 'wasted' (only 1% of answers given in our study had to be categorized as erratic) or wrongly categorized.* (Fitzpatrick 2006:144)

By not using a retrospective interview it is likely that some responses were erroneously categorised. If for example we get the response *team* to the cue word *football* we have a problem. Perhaps the subject is thinking along the lines of "it's a team sport rather like rugby" (a Meaning based response) or perhaps the subject is making a collocational association, as in "Swansea City is my favourite football team" (a Position based response). Other kinds of responses that would need clarification would be cue words with responses such as those in Table 2.13. In these cases, without clarification we would not be sure that the responses were due to the similar sound or whether these responses were collocational.

Table 2.13 Difficult to classify responses

Cue word	Response
cook	book
hot	pot
night	light
way	lay
no	go
flower	power

It should be noted that such ambiguous responses are probably not so common if (as in this study) the cue words are well chosen and those doing the classifying come from a similar cultural and linguistic background to the subjects. A lack of a retrospective interview will probably be far more serious in L2 studies where the classifier has to try to second-guess subjects from numerous different linguistic and cultural backgrounds (such as Sökmen 1993). A factor in favour of this study is the number of items, 100 per profile. In a subsequent study (Fitzpatrick, 2009) the interview was again left out of the methodology, as this was an L1/L2 study of Welsh bilinguals we might have expected an interview to have been used. It might be argued that with such a large number of response items, one or two erroneous classifications would be insignificant and it would therefore be unnecessary to go to through the time consuming process of verbally clarifying the thinking behind each response. Were the number of items a lot smaller, say 30 items, one or two erroneous classifications would be more of a problem. Unfortunately, as Fitzpatrick does not give us any information about how many of the responses were ambiguous or even an estimate of the number of potentially erroneous classifications we cannot really say that her decision to cut retrospective interviews from the methodology was justified.

The third point that ought to be mentioned is the use of Euclidean distance as a statistical measure. While it is a valid measure of the proximity of the profiles being analysed it should be noted that it is rare in the field of psycholinguistics, most readers with a more linguistics background (as opposed to psychology) would probably be unfamiliar with it. As the point of statistics is to make sense of the data being dealt with and organise it into a form that is easily comprehensible to the intended audience I feel its inclusion needs further explanation and justification than Fitzpatrick gives.

Despite these negative points, Fitzpatrick's 2007 paper remains a remarkable piece of work; it is well argued and innovative from both a theoretical and a methodological perspective. The analysis of individual response patterns (as opposed to group patterns) in effect turns the standard way of interpreting word association data on its head. The

principled selection of stimulus words and precise categorisation help to breathe new life into the idea that word associations can be used as a window on the mental lexicon. With only 34 citations on Google scholar (as of May 2013) this paper has yet to receive the recognition it deserves although variations on the categorisation system have already been used in recent studies (Fitzpatrick, 2009; Higginbotham, 2010; Fitzpatrick & Izura, 2011; Wang & Zhang, 2012) and the idea of analysing individuals over group data is also gaining wider acceptance (Albrechtsen et al., 2008; Meara, 2011). As the methodology in this paper is fast becoming a standard framework within which to work, it is essential to ensure that it is sound. As Fitzpatrick herself concludes (2007:328):

> *...if we are to use investigative tools such as word association to help us develop a better understanding of vocabulary acquisition, storage and retrieval, it is essential that the assumptions underlying our investigations are well founded and robust.*

Before research can begin on applying this seemingly reliable method of analysing word association data it needs validating against key word association variables. Of the variables identified in classic L1 studies (Deese, 1965; Cramer, 1968), the effects of *word frequency* and *word class* would seem the most salient. While recent studies usually account for these (presumed) key variables when selecting stimulus words to use in word association studies their effect is not normally explicitly analysed. An interesting question to ask would be whether *word frequency* and/or *word class* have an effect on the reliability of the individual response profiles created using Fitzpatrick's framework.

2.14 Zareva 2011

2.14.1 Summary

As argued in the previous review paper, when someone is given a word association stimulus there are two main factors that affect the way that person responds, the first is the word itself, the second is the person making the response. In Fitzpatrick's words (2007) "if the cue word were the only influence, all responses would be identical; if the respondent were the only influence all responses would quite possibly be different". Both L1 studies (Cramer 1968) and L2 studies (Meara 1978) have long shown that while some responses are heavily influenced by the stimulus word (e.g. *knife→fork*) for many words neither of these extremes apply. What actually happens is that responses to some words are stable but with other words there is a lot of variation. Also, it has been shown that some groups of people respond in predictable ways to certain words whereas others don't. The general conclusion is that responses are the result of complex interactions between the various

attributes of the stimulus word (Zareva calls these word-related factors) and the characteristics of the participant (Zareva calls these learner-related factors).

In this study the effect of two word-related characteristics (frequency and word class) and also two learner-related characteristics (proficiency and word familiarity) were investigated, the aim being to "disentangle" these effects. This general aim had three components; the first was to measure the effect of different types of stimulus items (words from different classes and words from different frequency ranges) on the responses. The second was to measure the effects of different kinds of participants (learners of varying proficiency) on responses. The third was to explore the interactions between word-related and learner-related characteristics.

The 108 participants were divided into three groups, 36 native English speakers, 36 advanced English learners and 36 English intermediate learners. The participants were given 36 stimulus items (12 nouns, 12 verbs and 12 adjectives) and asked to write the first three associations that they could think of for these words. The stimulus words were selected from a corpus of 12 million words (Zeno et al., 1995), there were three frequency bands, high, medium and low frequency. As well as the word association test participants were asked to rate their familiarity with each stimulus word on the Vocabulary Knowledge Scale, Wesche & Paribakht (1996), only responses that were rated as being 'known' were included in the analysis. Responses were classified as either paradigmatic or syntagmatic. Following a multivariate analysis of variance the main findings were:

- Proficiency level and lexical category, when combined had a significant effect on paradigmatic associations.
- Lexical class and word frequency, when combined had a significant effect on both paradigmatic and syntagmatic associations.
- Of the lexical classes analysed, nouns and verbs did not affect the proportion of paradigmatic responses but adjectives produced varying proportions at the different proficiency levels. Adjectives generated a mean of 53 (SD 22) paradigmatic responses with the native group, a mean of 36 (SD 25) with the L2 advanced group and a mean of 19 (SD 14) with the intermediate group.
- The higher the frequency of the stimulus words the more responses there were.
- Regardless of proficiency or word frequency participants produced more associations to nouns and adjectives than verbs.

The author concludes that both word-related and learner-related variables have a measurable effect on responses. Particularly, the lexical class of the stimulus word is

argued to have a significant role in determining the response as certain words (nouns and adjectives) connect in richer networks than others (verbs). From a pedagogical perspective, it is suggested that verbs therefore need more study.

2.14.2 Critique of Zareva 2011

In Zareva's thorough review of the literature on word association studies in both L1 and L2 contexts she identifies serious gaps in the research, the lack of studies that explicitly measure how different kinds of stimulus words affect the responses and also how different kinds of respondent affect the responses. Until these effects are more precisely understood the results of word association studies will be open to doubt and couched in caveats that make results difficult to interpret. The current need to account for these variables in word association studies also limits the kind of questions that can be asked. Consequently Zareva's attempts to find out more about the effects of the more salient variables, that have long been assumed to have some kind of effect, are to be welcomed. Unfortunately the validity of the findings is questionable due to a series of methodological problems that will be explained in subsequent paragraphs.

The main problems with this study concern the stimulus items used, it is therefore these that we will concentrate on. The first is that there are only a small number of stimuli (36). This is not a lot when we consider that they are further subdivided into three groups of 12 in the analysis (12 per word class, 12 per frequency range). Given that the number of words in the English language (depending on how you count them) is in the order of hundreds of thousands, twelve items is not nearly enough to make confident claims about how a particular word class or frequency band behaves. The problem with using more items though is that the test becomes impractical to administer. If we were to enlarge the sample size of each word type from 12 to a more representative size, say 100 (as in Fitzpatrick, 2007) to test three word classes then we would need 300 items. The time needed for each participant to give three responses to each of the 300 items would undoubtedly lead to fatigue related reliability problems. Naturally, one might question whether three responses were really necessary, although even if each participant were asked to give just one response the time required would be prohibitive. This leads me to conclude that the experimental design was overly ambitious in attempting to measure the proficiency and frequency effects on three word classes at the same time. It would have been easier to "disentangle" the effect of each word class if they had been attended to in three separate studies. Such single word class studies would probably generate more

reliable data, as more items per word class could be included. It might also be noted that Zareva's misguided preference for using small groups of items (typically 36) of mixed class stimulus words also finds its way into her other studies (Zareva, 2010; Zareva & Wolter, 2012) and consequently casts doubt over the findings in those studies too.

The next point of concern with the stimulus words is with how they were selected, even at a cursory inspection the stimuli seem odd. Would anyone really expect an English learner (even an advanced one) to make responses to words such as *cassava, gambol* or *putative*? Clearly these words were not trialled with learners prior to the main study and were apparently selected purely on the basis of their frequency within a corpus of written English. It ought to be noted that the 12 million word corpus from which the words were selected (Zeno et al. 1995) is an unusual source from which to pick words. As there are much larger corpora, the British National Corpus and the Corpus Of Contemporary American English (Davies, 2008), Zareva's choice is surprising. While corpora are fairly similar in their ranking of high frequency items, corpus size does become an issue when dealing with low frequency items. As low frequency items are an important part of Zareva's study it would have been better to have selected from a larger corpus in order to more accurately identify suitable stimulus items for each of the frequency bands studied. The COCA corpus, which is roughly 30 times larger and also far more up to date than the corpus used by Zareva, would have been a more logical choice. When we look at the frequency of Zareva's stimulus items in the COCA corpus we find that there is actually a great deal of overlap between the three frequency bands. The so called "high frequency" band for example contains items such as *experimentation* and *weaken* which COCA places as lower frequency than items in the so called "mid-frequency" band such as *concede* or *defensive*. Items within the "mid-frequency" band such as *coinage* and *middling* are also of lower frequency than *savor* and *amoral* in the "low frequency" band. In fact I would argue that the "high", "mid" and "low frequency" labels are misleading. Researchers (Nation, 2001) generally regard the most frequent 2000 words as "high frequency" as these words give a high coverage of words within most texts or spoken discourse (around 80%). As only three of the words within Zareva's "high" band are within this top 2000 word range it would have been better to label this band as 'mid-frequency' and the other bands as 'low' and 'very low-frequency'. Due to problems with these overlapping frequency bands, Zareva's conclusions concerning the effect of frequency are not well supported. Again, in order to help with 'disentangling' this complex set of results, it would have been better to have dealt with frequency in a follow up study and in this initial study used a set of stimuli

from just one frequency range that all participants could cope with.

The last problem with the stimulus words is the use of items that are within more than one word class, *hunger* for example can be used as either a verb or a noun. Using items with multi-class functions is not advisable, even when a symbol is given to denote which word class the participants are supposed to respond to (e.g. hunger n.). This questionable methodology is something that also features in Zareva's 2010 word association study. Despite the n. sign (or perhaps because of it), it seems quite likely that some participants might unintentionally respond to the verb form. It is often the case that when one is instructed not to think of something it is precisely this prohibited thought which preoccupies us the most. By cross-referencing the synonyms given in the familiarity measure with the responses to these multi-class words it would be possible to identify such erroneous data, this however is not reported. Whether this was done or not though, I am concerned that so many multi-class words (a third of the stimulus words) should have been included in the first place. As has already been argued, it is often difficult to classify responses; it therefore seems to unnecessarily complicate matters by including such items. As there are many other words available (especially in the frequency ranges that Zareva is working with) which are only used in one word class the decision to include so many words that function in more than one word class was unfortunate.

The poor choice of stimulus words used in this study is also evident when we look at the number of responses per proficiency group. The intermediate learners gave far fewer responses than the advanced or native groups, they could only give 1,124 responses out of a potential 3,888 (36 students x 36 items x 3 potential responses). The advanced and natives could give two or three responses to each stimulus but the intermediates were struggling to give one response per stimulus word. I would interpret this as showing that some of the words used were too hard for these students. As the lowest frequency words were evidently at the very periphery of intermediate learners' lexicons it is likely that a lot of them were guessing. Even though some were able to give an acceptable synonym for these words in the familiarity measure, the general low rate of understanding suggests the data from the intermediate group in the lowest frequency band is unreliable.

Zareva deserves credit for identifying and attempting to answer important questions about the effect of various word-related and participant-related variables on word association responses. Unfortunately the results cannot be accepted due to problems with the methodology, particularly the quantity and quality of the stimulus items used. There is also the general issue of Zareva's unnecessarily complex experimental design. If

proficiency, frequency and each of the word classes had been explored in separate studies (rather than attempting to roll them all into one) the results would have been far easier to interpret.

2.15 Discussion

In this section I will pick up on four issues that were repeatedly raised in the papers reviewed. These persistent issues are: classification, the stimulus words, the group/individual perspective and a general problem with experimental complexity.

2.15.1 Classification

Word associations have been used with learners for over 50 years. One would imagine that in this time, through a process of trial-and-error, a basic categorisation system would have evolved and been agreed upon. As the studies in this review have demonstrated there is still no such consensus on how best to categorise L2 word associations and a variety of competing systems exist. One problem, noted by many researchers (Meara, 1987; Sökmen, 1993; Wolter, 2001; Orita, 2002; Henriksen, 2008) is that clearly assigning a response to one discrete category is sometimes not straightforward. If for example the stimulus *pick* is used and the response is *stick* then the rater has a dilemma. It could be classified as a phonological/orthographical response (the two words have a similar /ɪk/ ending) or it could be classified as a collocational/syntagmatic response (referring to the game *pick-up-sticks*). Even when there are clear category guidelines the researcher is often second-guessing a response by a learner from a different generation and/or cultural background. A variety of measures have been adopted to increase classification accuracy (multiple judges, retrospective self-evaluations, interviews), although as will be explained in subsequent paragraphs, there are drawbacks to all of these.

Early L2 studies (Politzer, 1978; Meara, 1978) followed the system set up to explore L1 lexicons, dividing responses into two main groups, *paradigmatic* and *syntagmatic*. As this two-way distinction not only proved to be difficult to use in the L2 context but was also found to be too broad to be of much value, other categorisation systems were developed. Sökmen (1993) for example used an *affective* category for responses that were given due to a personal experience or emotional response. Unfortunately this particular category was not trialled and due to being poorly defined led to a large number of ambiguous responses being categorized as *affective.* Ultimately it was a rubbish bin for all the responses that didn't fit nicely into the other categories. Two more

recent systems are provided by Fitzpatrick (2006) and Henriksen (2008). While these seem to be improvements on the paradigmatic /syntagmatic distinction they are certainly not the final word on the matter. In recent studies (Fitzpatrick & Izura, 2011; Fitzpatrick et al, 2013) Fitzpatrick felt the need to combine some of the categories in an extended version of the system introduced in Fitzpatrick (2006). In this modified system there was an attempt to deal with the overlapping nature of some responses by creating some "duel-link" categories such as *form and meaning* and *meaning and collocation* to cope with stimulus/responses such as *pencil→pen* and *pen→paper* that she argues belong equally to two categories. It might also be noted that some recent studies (Namei, 2004; Zareva, 2011; Shimotori, 2013) have continued with the traditional *syntagmatic paradigmatic* distinction favoured by L1 studies. As one would expect, over the years the definitions for this two way classification have been made more explicit and clearer guidelines worked out: the classification system has gradually been improved.

Due to the difficulty with second guessing what learners are really thinking when they make particular responses it seems that with whatever taxonomy a researcher adopts there needs to be a method of checking that the rater does indeed understand why a learner made a particular response. Interviews (such as those used in Wolter 2001 and Fitzpatrick 2006) are one solution, these however are time consuming which often means they are left out of large-scale studies. A check that Schmitt (1998a), Wolter (2001) and Orita (2002) employed was to use more than one judge. While on the surface it may seem that another judge will add a level of objectivity to the classifications, this alone does not in my opinion solve the classification problem. As argued earlier, the second judge is quite likely to misclassify an ambiguous response in the same way as the first. Another option which can be used with larger groups, attempted by Henriksen in her 2008 study, is to ask students to undertake a retrospective task in which they review all their responses themselves and describe why specific associations were given. While it seems likely that this measure will allow the researcher to get close in many cases to understanding the response, a word of caution is needed. When self-analysing responses, even if retrospection takes place a few seconds after the response, it may be that respondees are not actually remembering accurately. As associations between many words are probably made subconsciously, students might not be consciously aware of why they make a particular association. Even if they do have the ability to reconstruct their thoughts in this way, then another possibility is that they may edit their reconstructions in some way, perhaps by telling the tester what they think the tester is expecting to hear. With these potential confounding factors in mind

the rater would be advised to weigh up all available information from; a native norms list, a second judge, knowledge of the learners L1 and learning background. When combined with a good deal of common sense, it is likely the dominant cause of most responses will become apparent. Given though that a certain amount of error is unavoidable with this kind of data, there is a strong argument in favour of using as big a set of stimulus words as possible. This issue will be dealt with in the subsequent discussion section.

2.15.2 The quantity and quality of stimulus words

The next issue is with the words that are used as stimuli within the word association tests. The problem is twofold, the number of words used (quantity) and the kind of words used (quality). Firstly, with regards to the quantity of words used (Table 2.14) there is a wide range, from 9 -100 items. Precisely how many items are necessary to obtain reliable data is unclear, although it would seem difficult to place much confidence in studies which at the lower end (Kruse et al., 1987; Zareva, 2011) base their conclusions on responses to a dozen or less items. Studies that have around 100 items are perhaps erring on the side of caution, but as was argued in the previous section, inherent problems with classification mean that longer stimulus lists are preferable. As it is likely that stable results can be obtained with items somewhere between these two extremes it would be useful to know more precisely where this boundary lies. There are unfortunately no studies to my knowledge which explicitly attempt to quantify how many items are necessary to elicit reliable response behaviour.

Table 2.14 The quantity and quality of stimuli used in 11 word association studies

	Number of stimuli	Selection of stimuli
Politzer 1978	20 in each condition	Questionable
Kruse et al. 1987	9	Questionable
Sökmen 1993	50	Questionable !
Söderman 1993b(experiment 1) (experiment 2)	100 64	Questionable ! Questionable
Schmitt 1998a	11	Fit-for-purpose
Wolter 2001	45	Fit-for-purpose
Orita 2002	60	Questionable
Henriksen 2006	15 in each condition	Questionable
Fitzpatrick 2007	100	Fit-for-purpose
Henriksen 2008	24 in each condition	Questionable !
Zareva 2011	12 in each condition	Questionable

! drawn from the Kent-Rosanoff list (1910)

The second issue is the quality of the stimulus words, which is of course linked to

quantity. If the items are capable of accurately measuring the response behaviour then an argument can be made for using less of them (Schmitt, 1998a; Wolter, 2001). If some stimulus items are poorly chosen and so elicit ambiguous responses that could conceivably belong to more than one category, then more items are required so that any poorly performing items do not skew the data. As the reviews have indicated, a number of studies (Table 2.14) suffer from a questionable choice of stimulus items. Many studies used stimulus words from the Kent-Rosanoff list (1910), providing a convenient comparison with L1 associative norms. Some were apparently picked without any real thought (Politzer, 1978), working on the misguided assumption that any word can serve equally well as a stimulus. I would argue that stimulus word lists that can be expected to give the most reliable response data have:

a large number of items;

been selected to suit the purposes of the experiment using principled criteria;

been trialled with a representative sample of learners

The list used in Fitzpatrick (2007) comes closest to fulfilling all these criteria, which along with the stimulus word lists used in Schmitt (1998a) and Wolter (2001), is deemed 'fit-for-purpose'.

One further point that needs to be considered when creating stimulus word lists is how many responses are required. Of the studies reviewed, some studies (Kruse et al., 1987; Schmitt, 1998a; Henriksen, 2006 & 2008; Zareva, 2011) ask students to give two or three responses whereas the other studies ask for single responses. While multiple responses, if well chosen, will give more detailed knowledge about how a word in a learner's lexicon is connected to the rest of the network there are also some associated problems. One is the possibility that by asking for more than one response the second (or third) response might not be a response to the initial stimulus but a response to the initial responses, this is known as chaining. Another is that by asking for multiple responses the experiment limits the number of items that can be tested in a session. It is no coincidence that studies that ask for single responses (Sökmen, 1993; Söderman, 1993b; Orita, 2001; Fitzpatrick, 2007) are also the studies that have the greatest number of stimulus items.

2.15.3 Analysing the data from a group or individual perspective

An important point to come out of both the L1 and L2 studies is the lack of consistency in the findings. Classic L1 studies such as Ervin (1960) and Entwistle (1965) argue that children are characterized by syntagmatic response behaviour with responses becoming

more paradigmatic with increased age/maturity. This phenomenon is known as the "syntagmatic - paradigmatic shift". Studies such as Politzer (1978) and Söderman (1993b) suggest that L2 learners are similar to native children in this respect and present evidence that L2 learner's responses are to some extent syntagmatically dominated. On the other hand, the findings of studies such as Stolz &Tiffany (1972), Wolter (2001), Henriksen (2006), Fitzpatrick (2006) challenge such a shift. These studies mostly analyse the results by membership to a particular group, although there seems to be growing awareness that word association responses are often as idiosyncratic as the individuals that make them. Galton (1883:131) in the first published word association experiment stressed that an association "is the fruit of experience, it must differ greatly in different minds according to their individual experiences". Having begun with a clear recognition of the role individual experiences play on responses, it is curious that for over a hundred years L1 and then L2 research only analysed responses from a group perspective. It is only recently that we see a return to analysing individuals, in the work of Schmitt (1998a) and also to a lesser extent within Wolter (2001), who noted that one of the participants in his pilot test exhibited very idiosyncratic response patterns. Fitzpatrick made this idea explicit in her series of studies (2006, 2007, 2009) that puts forward the concept of "individual profiling". This idea seems to have been taken on board by Albrechtsen et al. (2008) who in their large study of Danish student's vocabulary knowledge build up "vocabulary profiles" for their learners based on a number of separate measures. Although many studies (Zareva, 2011) continue to view word association responses in terms of the groups that participants belong to, the tide seems to be turning in favour of analysing responses from an individual rather than a group perspective.

2.15.4 Complexity of experiments

The brain is still very much a 'black box' in that we cannot look directly into it while it is working, identify particular thoughts and actually see how they interact and connect with other thoughts. Although studies in neurology have shown us which parts of the brain are active when engaged in certain cognitive tasks, they do not inform us of what is really going on within these areas. Mestres-Missé et al. (2010) for example, using MRI technology, demonstrate which parts of the brain become active with particular word classes. Although studies such as these are interesting, understanding where the activity occurs is not particularly helpful in understanding how the lexicon is structured. The best that we can currently hope to do is measure what goes into this black box and make

inferences about what happens in between from what comes out. With a complex cognitive task, such as language production, there are numerous variables that can affect what happens inside this black box; consequently logic would seem to dictate that if we are to correctly interpret what is happening we need to carefully control what goes in. When the input is a word we therefore need to account for all the potential factors that may affect the responses to this word (frequency, word class, emotionality, abstractness, length etc.) as well as all the factors affecting the participant (proficiency in the language, age, background, gender, intelligence etc.). Having accounted for all these variables it is then possible to manipulate one of them and on examining the outcome make an inference about the processes involved. The more variables we try to manipulate at one time the more difficult it becomes to interpret the resulting responses. As has been argued in the reviews, many of the studies have attempted to explore too many variables at one time, therefore making it difficult to understand precisely the causes of the responses. Also, due to the limited amount of responses that any one learner can give in a word association test session, the more times this data is divided up between the variables in question the weaker the support for each claim becomes.

In Kruse et al. (1987) for example we have a study that attempted to measure proficiency based on the stereotypy of the response, the problem is that the stimulus words were a mix of words ranging from very stereotypical to not very stereotypical; there were also two proficiency levels. This means that it is difficult to understand whether a response was caused by the proficiency of the participant or the stereotypy of the stimulus word. In trying to answer the question "which associations are useful to teach?" Sökmen (1993:135) overcomplicated her analysis by making claims about the effect of student background and word class that were not really supported by her data. A far simpler study using learners from just one background and one word class would have strengthened her main findings. Wolter (2001) selected stimulus words from two frequency bands and also from three word classes (nouns, verbs and adjectives). As the precise effect of each word class on responses in largely unknown, it would probably have been wiser to have just used one word class. Follow up studies could have investigated the responses from other word classes. In her 2006 study, Henriksen analysed the effect of word class on responses, as with Wolter (2001) she tried to measure the effect of three word classes at the same time. With just 45 stimulus words, when broken into the three word classes there were only 15 items per condition, this study therefore failed to make any convincing claims about a word class effect. Had Henriksen run three separate experiments for each of the word classes she

could have had 45 items per word class and far more credible data. Of the studies reviewed, the one that really stands out though is Zareva (2011); here she measured the effect on responses of proficiency, word frequency, word-familiarity and three word class, all within one experiment. As there were only 36 responses per student to work with, inevitably she didn't succeed in providing convincing data to support any of the statements that she made.

Another point that compounds the basic problem of trying to measure too many variables within one experiment is that the number of participants is often surprisingly small. Kruse et al. (1987) had 15 learners, Henriksen (2006) had 25 and Wolter (2001) 12 learners and 9 native speakers. With less than 25 participants in each of these experiments, it is difficult to place much confidence in the generalisations that are made about learner response behaviour. Even when the numbers of participants seem quite good, Zareva (2011) had 108 participants, when these are divided (into three proficiency groups in Zareva's study) the numbers become less impressive.

The lesson to be learned from these studies is that it is better to keep the experimental work fairly simple. Looking at just one variable at a time and taking lots of small steps is preferable to aiming for giant leaps forward. It is easier to fit together a series of simple experiments than attempt to divide up the results of a larger and more complicated study. While we should be wary of trying to manipulate too many variables at the same time, this does not mean that experiments should not account for them all. For example, in the studies by Söderman (1993), Sökmen (1993) and Orita (2002) there were no explicit measures of proficiency, the experiments therefore lacked data about an important variable, which puts a question mark over their conclusions. The argument in favour of simplicity ought not to be confused with a lack of control over the main variables.

2.16 Conclusions

With the selected studies I have tried to give a broad view of the last 30 years of research into learner lexicons through the use of word associations as a measurement tool. Within both the strands that have been identified (proficiency and type) some persistent problems have been raised concerning the methodology and analysis of the data collected. It is anticipated that by addressing these problems - adopting a principled approach to stimulus selection, using a clearly defined categorisation system and a reliable method of analysis - the long anticipated potential of word associations to answer questions about the learner's mental lexicon can be realised.

One study in particular that went some way to meeting these criteria was Wolter

(2001). In this study he proposed a depth of word knowledge model, which in a field desperately lacking a decent model within which to explain word association responses, offered a promising theoretical framework. The potential of this model and its upbeat conclusions were unfortunately only weakly supported by the small number of participants and stimulus words. In order to substantiate the claims made in that paper and see if his framework could in fact be used to move the research agenda forward, it was decided to revisit this study. In late 2005, when I began the preparation for this thesis, the approach taken (in what was then a fairly recently published paper) was not only attractive for the reasons given above but was supported by experimental work done in Japan. As I work at a Japanese university I was well placed to replicate this study using a very similar group of learners: an opportunity too good to miss. The following chapter is therefore a very close replication of Wolter's 2001 study.

Chapter Three: A replication study of Wolter 2001

3.1 Introduction

When making an omelette, it is necessary to break some eggs. Similarly, if we are to progress in our understanding of complex cognitive processes, such as language, we ought not to be afraid of cracking open a few studies in order to separate out the useful parts. The eleven critiques of word association studies in the previous chapter demonstrate that when we subject research to detailed scrutiny it is not hard to find fault, and with the benefit of hindsight, give advice on how it might have been done better. The hard part of course is developing new ways of thinking and applying them creatively in order to progress our understanding. It is therefore with this expectation for *progress,* and a sense of respect for those who explore difficult questions in innovative ways that I undertake a replication of a study done by Brent Wolter. A study that is, from various perspectives, at the centre of research currently being done to understand the mental lexicon through the use of word association tests. For the reasons already mentioned in Chapter 2, and also in proceeding paragraphs, Wolter's 2001 paper has had a significant influence on the field: evidenced by 150 citations in Google Scholar as of May 2013.

Firstly, the findings of Wolter (2001) add weight to the idea (Politzer, 1978; Piper & Leicester, 1980; Söderman 1993b; Namei 2004) that the structure of the L1 is not as different from the L2 as some have argued (Lambert & Moore, 1966; Meara, 1978; Channell, 1990). If this is indeed the case then there are implications on how we view the organisation of the learner's mental lexicon. Rather than building models based on the assumption that the L2 lexicon is organised in a different way to the L1 lexicon, perhaps we should consider it as being organised in the same way only smaller in size. Although Wolter suggests that the traditional idea of a syntagmatic-paradigmatic shift is not quite as straight forward as L1 studies (Ervin, 1961) would have us believe, he argues that his results do support the notion of a cognitive shift of sorts, from a phonologically-structured lexicon to a meaning-structured lexicon.

Secondly, this study has also been influential in its well justified methodology: particularly, the classification categories and the choice of stimulus words. The careful description of response categories by Wolter has prompted recent studies (Bagger-Nissen & Henriksen, 2006; Zareva, 2007) to be more explicit in how they classify responses. As can be seen in section 3.2.2, Wolter goes to some lengths to define and exemplify how each type of response ought to be dealt with. These are more precise than the definitions

used in previous studies (Politzer, 1978; Söderman, 1993b) that divide responses according to their word class alone (same word class = paradigmatic; different word class = syntagmatic). Another important element within the methodology is the development of a stimulus list based on well thought out principals rather than drawing from convenient (yet unsuitable) sources such as the Kent-Rosanoff list (1910). Recent MA studies under Wolter's supervision (Sowell, 2006; Racine, 2008; Wharton, 2011) also heed the advice in this paper with regards to basic approach and methodology, although they use word association tests to explore different aspects of the lexicon. Sowell (2006) for example looked at how cultural differences (American vs Arabic) influence word association responses; Wharton (2011) tracked the development of 30 vocabulary items with Japanese university students over a semester.

Thirdly, the influence of this study can be seen in research which follows Wolter's line of thinking by measuring depth of knowledge as a key dimension of the mental lexicon (Namei, 2004; Schoonen & Verhallen, 2008). In her study of Persian-Swedish bilinguals, Namei (2004) looked at the responses of 100 speakers between the ages of six and 22. In support of Wolter's initial hypothesis, the results of Namei (2004) showed "great similarities between the L1 and L2 in terms of the developmental stages of word acquisition". In her analysis of over 30,000 responses, she advances Entwistle's (1966) view that each word in the lexicon develops from unknown to well-known in a predictable way. Words that are barely known have phonological associations, those that are partially known have strong syntactic organisation and well known words are connected to other words on a semantic basis. Namei's study not only concurs with Wolter but offers a hybrid model that combines the main elements from the models offered by Entwistle and Wolter. As can be seen below (Figs 3.1 & 3.2), Namei's "word knowledge continuum" model contains all the elements of Wolter's DIWK. The main difference, taken from the "developmental stages in word association" (Entwistle, 1966:74), is that it gives a sense of progression (an arrow) through the five stages. Another difference is that Namei's 'well known' stage is dominated by paradigmatic and late-syntagmatic responses whereas in Wolter's DIWK this highest stage of knowledge is paradigmatic only.

Fig 3.1 The depth of individual word knowledge model (Wolter, 2001:48)

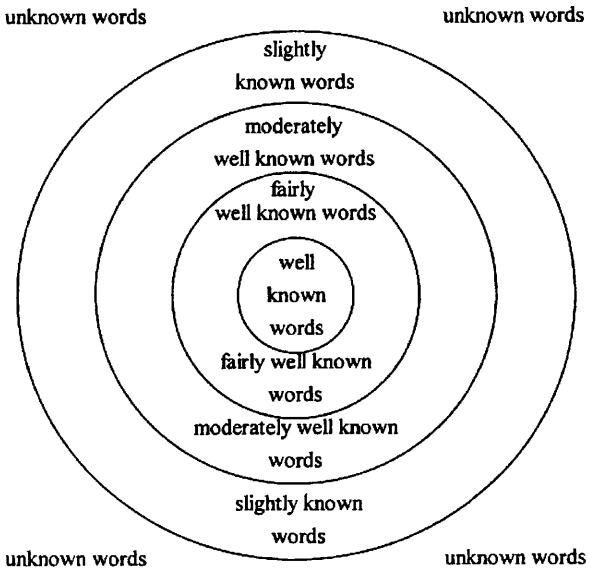
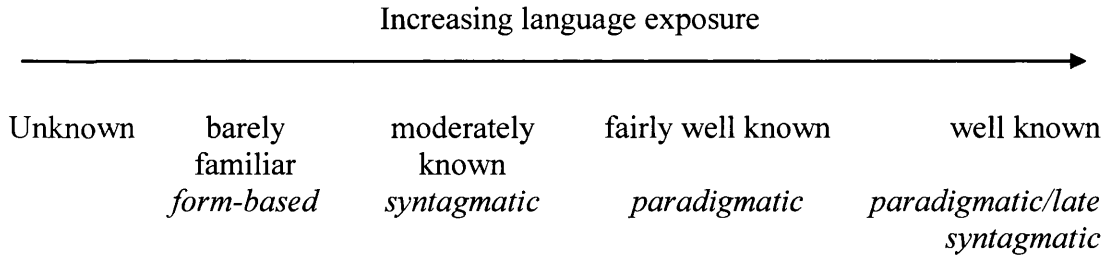


Fig 3.2 The Word Knowledge Continuum, (Namei, 2004:382)



It might also be noted that Wolter (2001) is often cited due to its use of low frequency stimulus words. Most studies investigating word association responses (Meara, 1978; Kruse et al., 1987; Nishiyama, 1996; Namei 2004) have been limited to high frequency nouns. As Wolter's study also investigates responses to low frequency stimuli selected from the three main word classes, it is regularly cited (along with Söderman 1993b) as evidence that WATs can be used with a wide range of stimuli. If broad generalisations are to be made about the mental lexicon, it is necessary to demonstrate that the instruments being used to measure it are not only comprised of a small group of words that do not fully represent the language.

Wolter (2001) has had a large impact on how researchers are currently conceptualising and measuring the mental lexicon. However, as noted in the critique in Chapter 2 the findings are questionable. The small number of participants in the experiments that support the main claims is a particular cause for concern. Another is that the findings from a study that was reviewed in Chapter 2 (Bagger-Nissen & Henriksen, 2006) directly contradict Wolter's. If we are to give credence to the findings of Wolter's 2001 study and the DIWK model that he proposes, then the reliability of the experimental work ought to be verified; a replication study is a logical way to do this. As Cohen et al., (2006) state, "for research to be reliable it must demonstrate that if it were carried out on a similar group of respondents in a similar context then similar results would be found". The following section details an experiment that aims to do exactly this.

3.2 Outline of this study

The main aim of this replication study is to see if Wolter's findings can be recreated using the same materials and similar samples. The main hypotheses are therefore the same as in Wolter (2001:42).

1. *The L2 mental lexicon of a nonnative speaker is structurally similar to the L1 mental lexicon of a native speaker.*
2. *Depth of word knowledge is a key component for determining the degree of integration for the individual words that make up the structure of both the L1 and the L2 mental lexicon.*

In Wolter 2001 there were two main findings. The first was that the initial hypothesis was only partially supported. The patterns of L1 and L2 response behaviour were shown to be similar for the unknown and partially known categories (VKS 1, 2 categories) but not similar for the well-known words (VKS 5 category). The second main finding was that hypothesis 2 was confirmed, "words in the lexicon form connections in a somewhat systematic fashion as they come to be better understood" (p 65). If Wolter's experiment is reliable then this replication should show broadly the same patterns.

3.3 The replication study

As one would expect in a replication study the methodology closely followed the original, the prompt words used for example were the same (Appendices 3.1& 3.2). The NNSs were asked to respond to words from the first prompt word list (PWL1) and then immediately

after asked to go through the list again and rate their level of knowledge for each word. The NSs were also asked to complete the word association test for PWL1, and in addition asked to complete a word association test for lower frequency words (PWL2). The NS group did not rate their word knowledge for the high frequency words but did rate their word knowledge for the lower frequency words. There were 45 items in each prompt word list. Although Wolter initially started with 48 he whittled these down to 45 prior to the analysis as some were unsuitable. Stimulus words such as *loyal* (likely to be confused with *royal*) and *pander* (likely to be confused with *panda*) were omitted. All the word association tests and interviews were done orally, on an individual basis and recorded. To help with classification some of the responses were later discussed with a second judge. Although no time limit was set the word association and word rating tasks took about 20 - 40 minutes per participant to complete. The NNS data for this replication study was collected in January 2007; the NS data was collected during February 2007.

3.3.1 Participants

In the replication study 16 students (NNSs) from Hiroshima Kokusai Gakuin University and nine native adult speakers of English (NSs) were asked to participate in a word association test and depth of word knowledge interview. The participants were nearly identical in numbers to those interviewed in the original study (13 NNSs and 9 NSs). As with the original study, the NNSs were all Japanese university students with genders equally represented and a TOEIC score of over 600. The TOEIC threshold, slightly higher than the 500 level set in Wolter (2001), was to ensure the learners could cope with the stimulus words and interview procedure. The NSs were a more mixed group, college graduates coming from a variety of English speaking countries: there were a wide range of ages.

3.3.2 Classification of data

The same classification system used in Wolter (2001) was adopted. An abridged version of the definitions for each of the four categories is given below:

Paradigmatic

In the same word class as the prompt word with the following three provisos:

a. The word did not show a clear sequential relationship to the prompt word. Such responses were classified as syntagmatic (e.g., *human→error*).

b. The word was not used to make a longer noun phrase (e.g., *discovery→discovery channel*). Responses such as this were also classified as syntagmatic.

c. The word showed a clear connection to the prompt word. Possible, yet distant, connections were determined to be unclassifiable and assigned to the clang-other responses category (e.g., *confine→tolerate*).

Syntagmatic

a. Words from a different word class than the prompt word that demonstrated some kind of semantic or syntactical relevance or relationship to the prompt word.

b. Words from the same word class that demonstrated a sequential or an affective relation to the prompt word, provided that the relation was overtly clear (e.g., *orchestra→conductor, San Francisco→hill*).

Clang-Other

Responses that resembled the prompt word only phonologically (e.g., *genuine→January*), those that were simply a different form of the prompt word (e.g., *concentrate→concentration*) were classified as clang-other responses.

A response that was determined to bear no obvious relation to the prompt word was judged to be unclassifiable. (e.g., *stand→anticipation)*, although in the mind of the participant there may indeed have been some sort of meaningful relationship between the two.

No response

Participants could not reply, or they stated that no word came to mind upon hearing the prompt word.

(adapted from Wolter 2001:52)

As in Wolter (2001), in the interview each word was also rated, based on a slightly adapted form of Wesche & Paribakht's Vocabulary Knowledge Scale (1996). The scale (Table 3.1) was adapted to suit the oral interview format used in this experiment.

Table 3.1 The Vocabulary Knowledge Scale assessment card (adapted from Wolter 2001:54)

The following activity will ask you to assess how familiar you are with the words you have just heard. This time you will be asked to rate each word you hear on how well you know it. For items III and IV you can use either an English synonym (a word in English with the same meaning) or a Japanese translation.
The scale is as follows:
1. I don't remember having heard this word before.
2. I have heard this word before, but I don't know what it means.
3. I have heard this word before, and I think it means _____. (synonym or translation)
4. I know this word. It means _____. (synonym or translation)
5. I can use this word in a sentence:_____ (If you do this section, please do section IV).

As it was assumed all NSs would have a level 5 understanding of the PWL1 words they were not asked to rate themselves on their depth of knowledge for these words. The NSs were expected to rate themselves on the PWL2 words, as these are fairly infrequent words it was expected that many would be either unknown or on the periphery of their lexicons.

3.4 Results

In order to facilitate an easy comparison with the results from the original study (Wolter 2001) and the replication study (GH07), graphs from both studies have been placed together. The initial graph in each pair (Figs 3.3a, 3.4a, 3.5a etc.) are graphs derived from Wolter (2001); the second graph in each pair (Figs 3.3b, 3.4b, 3.5b etc.) are the graphs from the replication study.

3.4.1 Comparisons of general response data

As can be seen by comparing Figs 3.3a and 3.3b the overall percentage of paradigmatic responses by NSs to the higher frequency words was far lower in the replication study (25.2%) than in the original study by Wolter (51.7%). The replication study (Fig 3.3b) also showed that the NNS's gave more paradigmatic responses (22.9%) than syntagmatic responses (14.9%). Looking at the data from this broad perspective the two studies generated quite a different set of responses from the two groups. The only area that seems to show some commonality is that phonological (*clang-other*) responses were far more numerous for NNSs (Wolter, 35.1%; GH07, 38.8%) than the NS clang responses (Wolter, 7.2%; GH07, 10.6%).

Fig 3.3a Percentage of NNS and NS response types for PWL1 (Wolter, 2001)

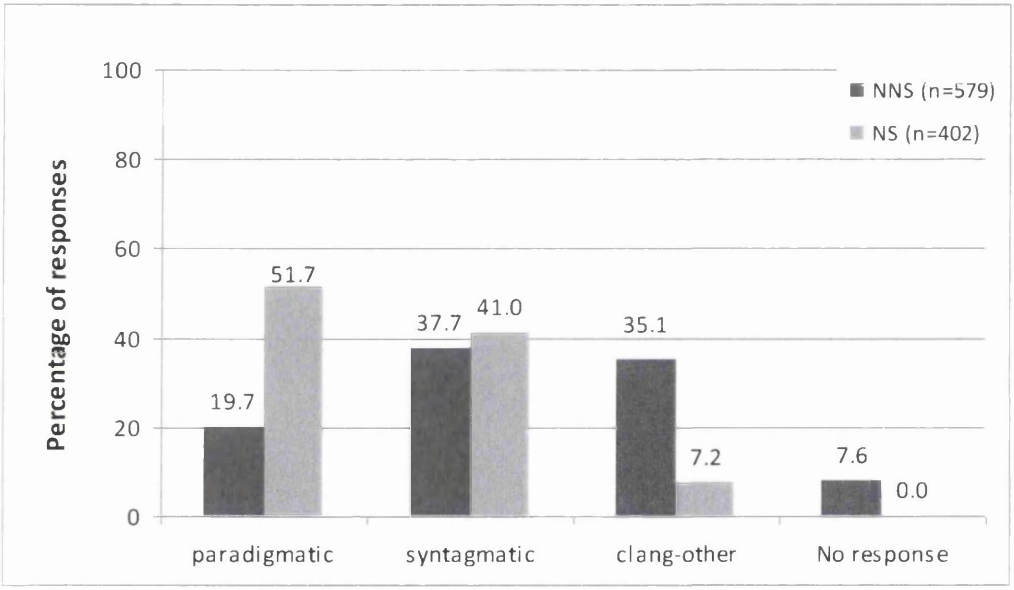
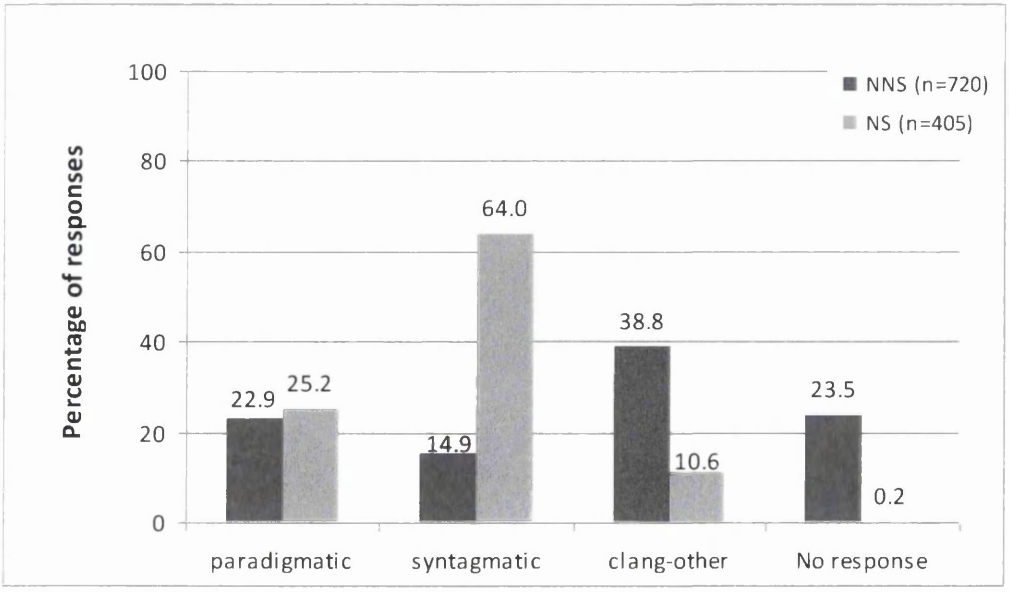


Fig 3.3b Percentage of NNS and NS response types for PWL1 (GH07)



A particularly striking difference between the two studies is that in the replication study syntagmatic responses are dominant for the NS group (64%). The general results of the replication study not only conflict with Wolter's original study but also with L1 studies (Entwistle, 1966) and other studies of L1 and L2 response behaviour (Piper & Leicester, 1980; Söderman, 1993b). A point on which these prior studies agree is that the responses by adult native speakers for common words are mostly paradigmatic. These studies also show that native children (Ervin, 1961) and also non-native learner responses (Söderman, 1993b) are generally syntagmatic. The conflicting findings from the replication echo those found by Bagger-Nissen & Henriksen (2006). The replication findings are incompatible with the idea that increased proficiency in the language is evidenced by a shift from a syntagmatically structured lexicon to a paradigmatically structured one.

In the next set of graphs, Figs 3.4a and 3.4b show how the native groups in each study responded to both the higher (PWL1) and lower frequency (PWL2) stimuli. In Wolter's study (Fig 3.4a) the paradigmatic dominance found with the higher frequency stimuli (51.7%) was still evident with the lower frequency stimuli (38.1%) although this dominance became less pronounced as more phonological responses were generated (27.2%). In contrast, Figure 3.4b shows that in the replication study the dominance of syntagmatic responses elicited from the NS's was not restricted to the high frequency words (PWL1) but also the low frequency words (PWL2). One area where both studies agree is that NSs give a small number of *clang-other* responses (7 - 10%) for the higher frequency words, this proportion rises (20 – 27%) with the lower frequency words.

Fig 3.4a Percentage of NS response types for PWL1 and PWL2 (Wolter, 2001)

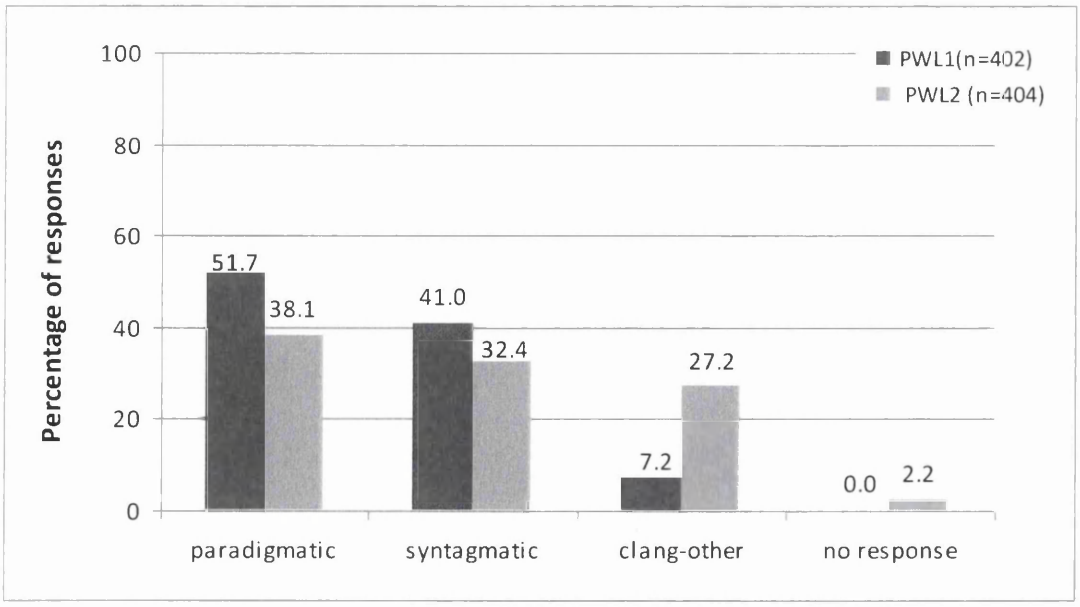
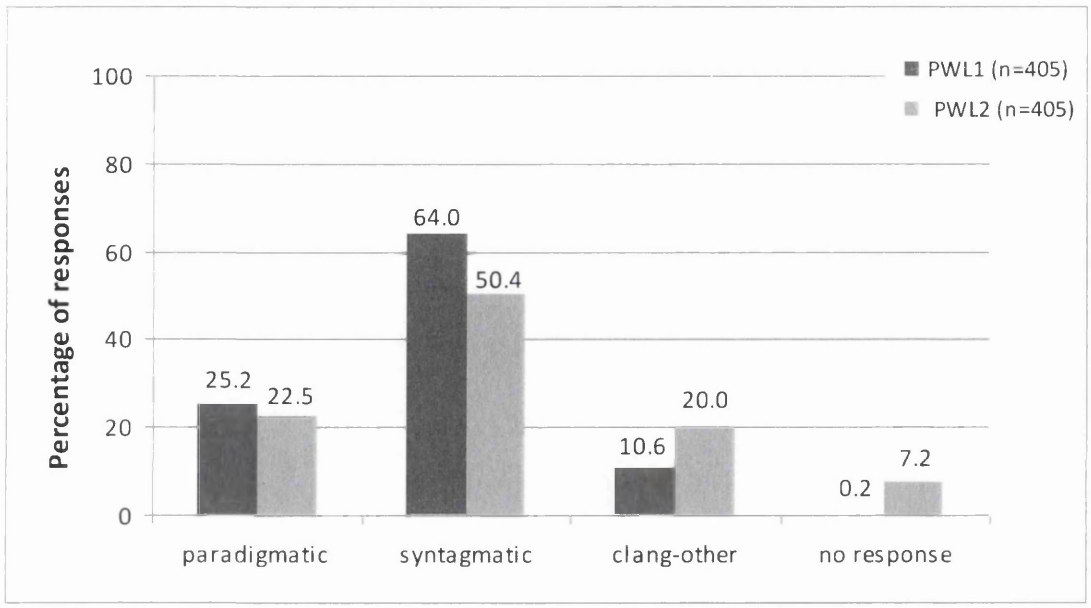


Fig 3.4b Percentage of NS response types for PWL1 and PWL2 (GH07)



In the correlation matrix (Table 3.2) the statistical relationship between the three main data sets (NS responses to PWL1, NS responses to PWL2 and NNS responses to PWL1) from Wolter's 2001 study and the replication (GH07) is shown. The variation between the percentage of responses given in each of the response categories was calculated. The correlation coefficients show that there is no statistical relationship between the NNS data in the two studies and that the NS data only has moderate relationships.

Table 3.2 Correlations between the percentage of responses in Wolter (2001) and the replication study

		Wolter 2001		
		NS PWL1 (n=402)	NS PWL2 (n=402)	NNS PWL1 (n=579)
GH07	NS PWL1 (n=405)	**0.712**		
	NS PWL2 (n=405)		**0.637**	
	NNS PWL1 (n=720)			**0.115**

3.4.2 Comparisons of response data at each level of word familiarity

In the following graphs the general response data in the previous section is broken down into responses given to words at the 5 levels of familiarity identified by the Vocabulary Knowledge Score (VKS). There are five sets of graphs, beginning with the unknown words (VKS1) through to the well-known words (VKS5).

In Figs 3.5a and 3.5b we see the type of responses given for unknown words: rated as VKS1. As with Wolter's original study the paradigmatic and syntagmatic responses were (as one would expect) fairly insignificant, although it is interesting to note that the NSs in the replication seemed better at guessing. Of the NSs' responses for these *unknown words*, 7% were classified as syntagmatic. As they claimed not to know them prior to the test then we can probably judge these responses as lucky guesses, although an alternative explanation will be put forward in the discussion section.

Fig 3.5a Percentage of NNS and NS response types for prompt words that elicited a VKS score of 1 (Wolter, 2001)

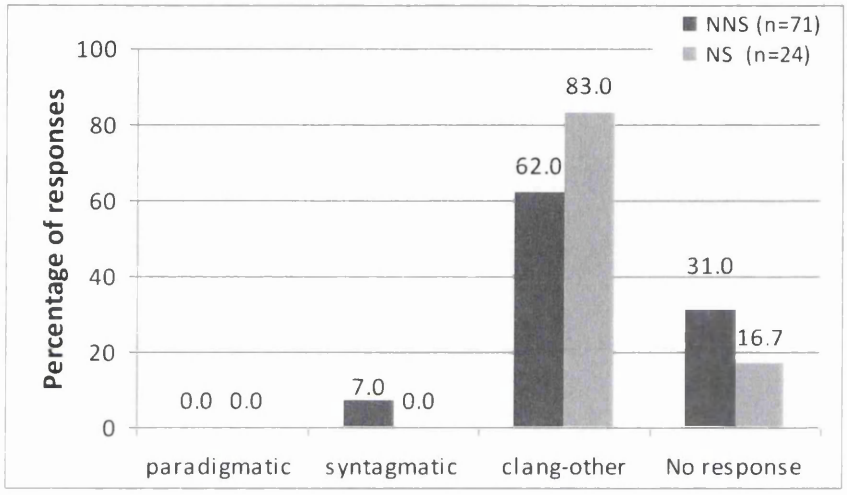
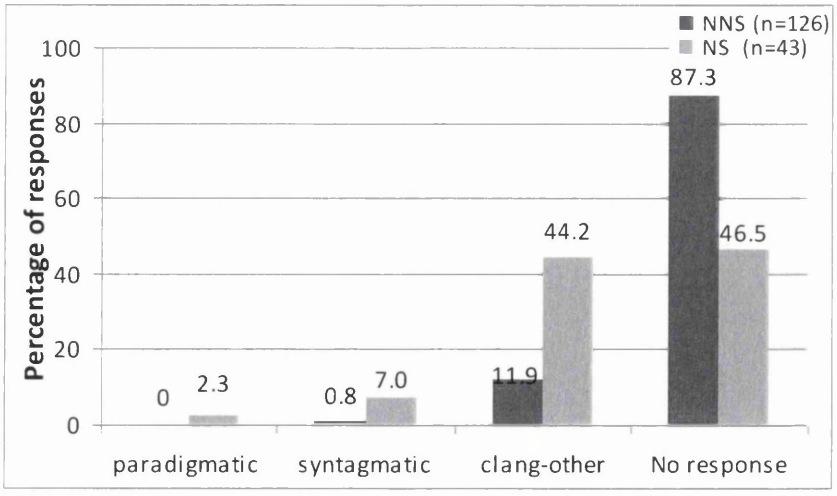


Fig 3.5b Percentage of NNS and NS response types for prompt words that elicited a VKS score of 1 (GH07)



In Figs 3.6a and 3.6b (responses to vaguely known words) we can see that *clang-other* responses and *no responses* dominate: as with the unknown words (Figs 3.5a and 3.5b). In

the replication study (Fig 3.6b) NNSs still gave a lot of *no responses* (57.5%) whereas in Wolter's study there were far fewer *no responses* (13%): most responses were phonological. As shown in Fig 3.6a when a word is vaguely known respondents usually (NS 78%; NNS 65%) give a *clang-other* response. In Fig 3.6b NSs made far more syntagmatic connections (32.3%) than in the replication study (9.4%). A point that both studies agree on is that at this stage of word knowledge paradigmatic responses are still rare.

Fig 3.6a Percentage of NNS and NS response types for prompt words that elicited a VKS score of 2 (Wolter, 2001)

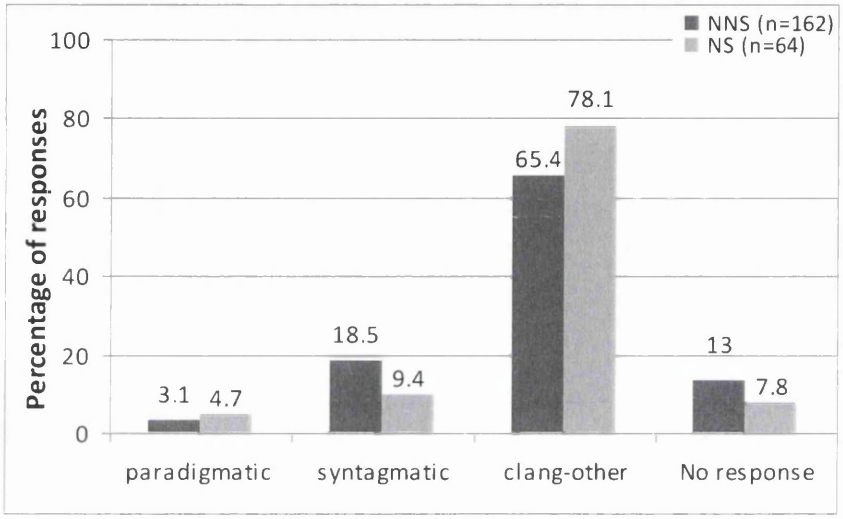
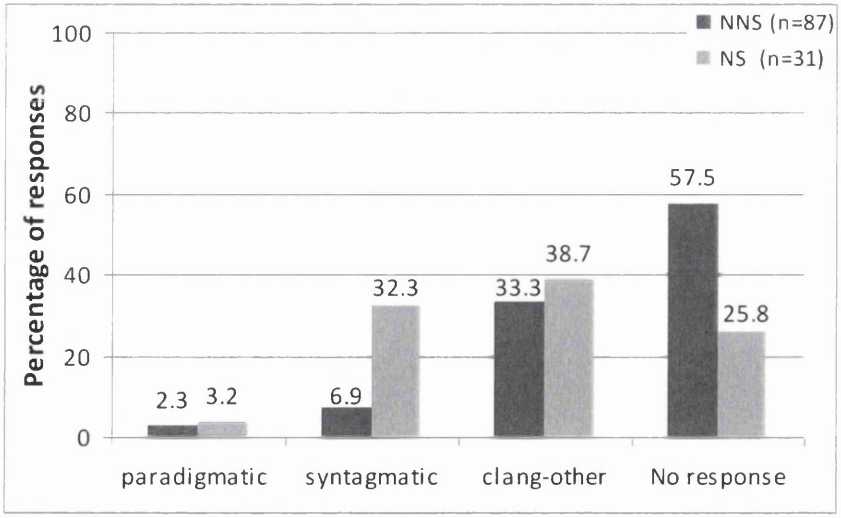


Fig 3.6b Percentage of NNS and NS response types for prompt words that elicited a VKS score of 2 (GH07)



At the VKS3 level of word knowledge the pattern of behaviour between the two studies becomes more disparate. When comparing Figs 3.6b and 3.7b from the replication study, the number of *clang-other* responses increases for both the NSs (38.7% to 48.3%) and NNSs (33.3% to 44.4%). In contrast when comparing Figs 3.6a and 3.7a the high number

of *clang-other* responses drops considerably with a subsequent increase in *syntagmatic* and *paradigmatic* responses. The large jump in the number of *paradigmatic* responses in Wolter's study (NS=37.5%; NNS=16.7%) at VKS 3 is not seen in the replication study.

Fig 3.7a Percentage of NNS and NS response types for prompt words that elicited a VKS score of 3 (Wolter, 2001)

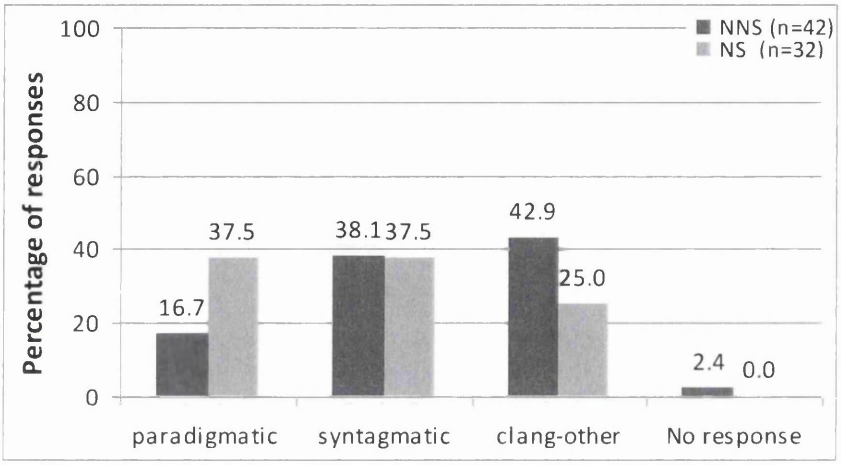
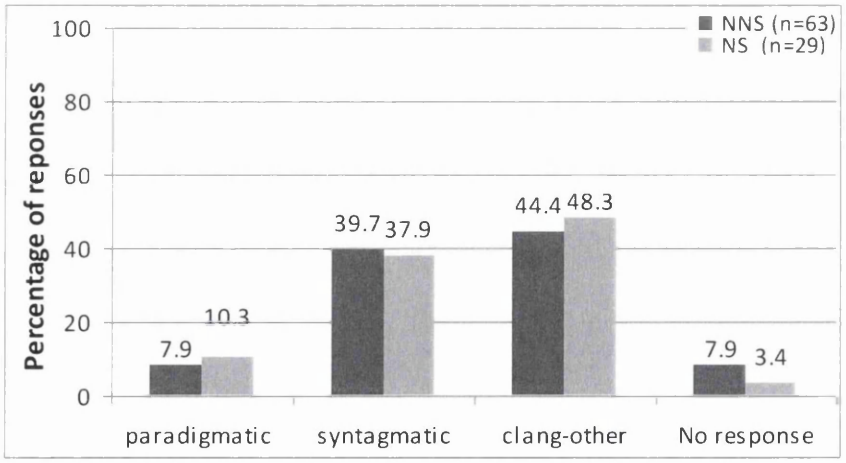


Fig 3.7b Percentage of NNS and NS response types for prompt words that elicited a VKS score of 3 (GH07)



At the VKS 3 level, where participants demonstrated that they did have a basic understanding of the meaning of the words, the original study (Fig 3.7a) showed a rise in both the proportion of *paradigmatic* and *syntagmatic* responses and a corresponding drop in *clang-other* responses. Between Figs 3.6b and 3.7b we see a different pattern emerging. The number of *syntagmatic* responses for NSs and NNSs increased, however the number of *paradigmatic* responses was much lower, and most surprisingly, the number of *clang-other* responses actually increased for both the NSs and NNSs.

Fig 3.8a Percentage of NNS and NS response types for prompt words that elicited a VKS score of 4 (Wolter, 2001)

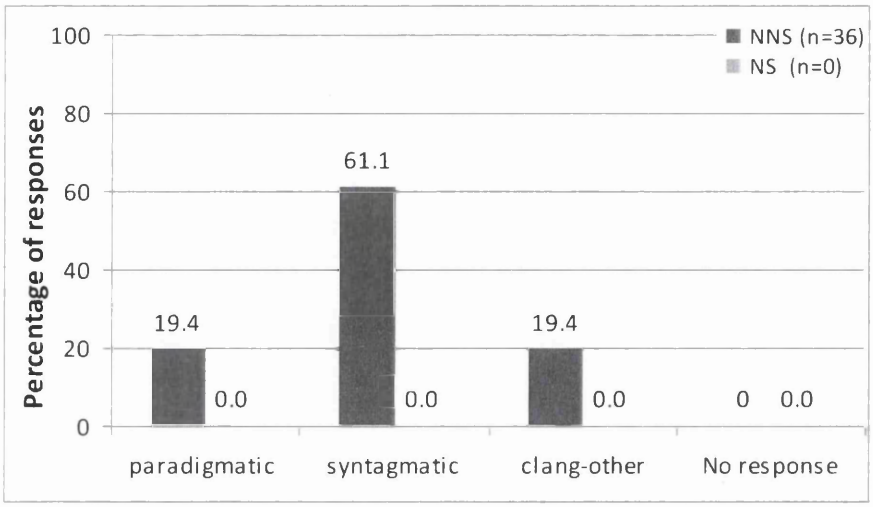
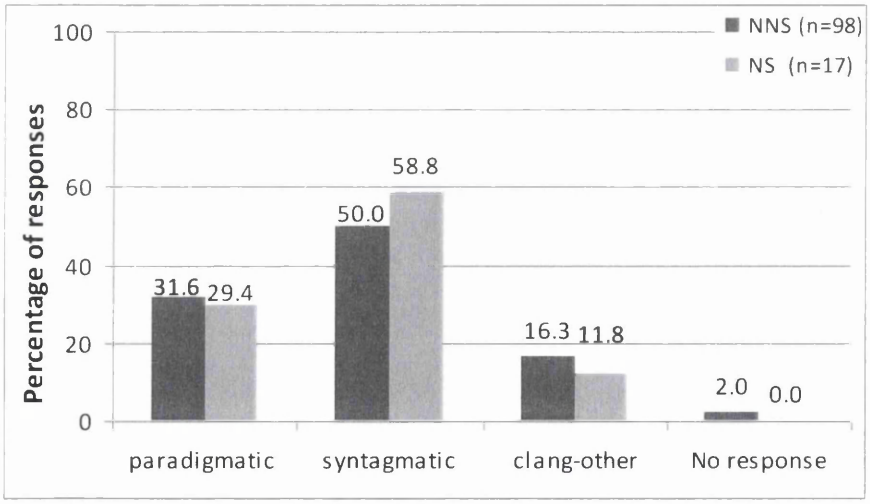


Fig 3.8b Percentage of NNS and NS response types for prompt words that elicited a VKS score of 4 (GH07)



The VKS 4 data (Figs 3.8a and 3.8b) do show a similar pattern with regards to NNS responses, in both studies *syntagmatic* responses dominate with a small rise in *paradigmatic* responses and a drop in *clang-other* responses. As none of the NS responses were judged to be at VKS 4 in Wolter's study the pattern of NS responses observed in the replication data cannot be compared. It might also be noted here that the number of responses judged as VKS4 were also very low (only 17 responses) in the replication study, a point that will be returned to in the discussion section.

Fig 3.9a Percentage of NNS and NS response types for prompt words that elicited a VKS score of 5 (Wolter, 2001)

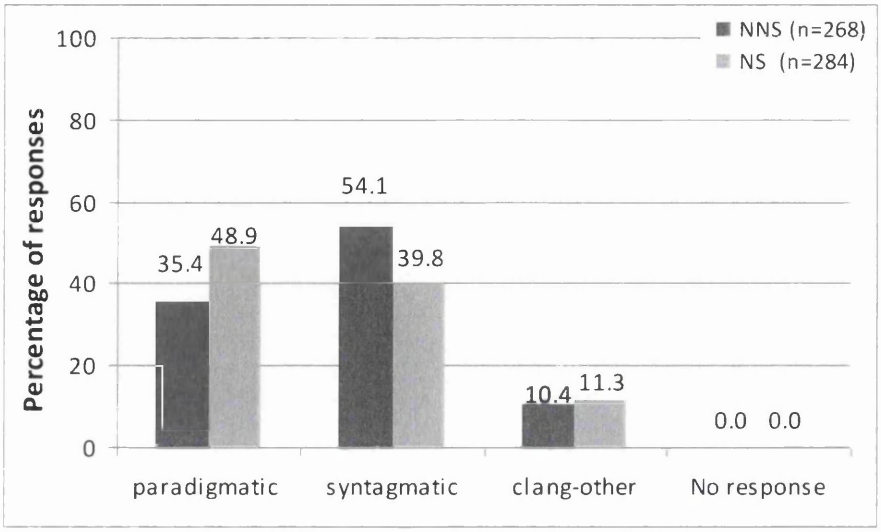
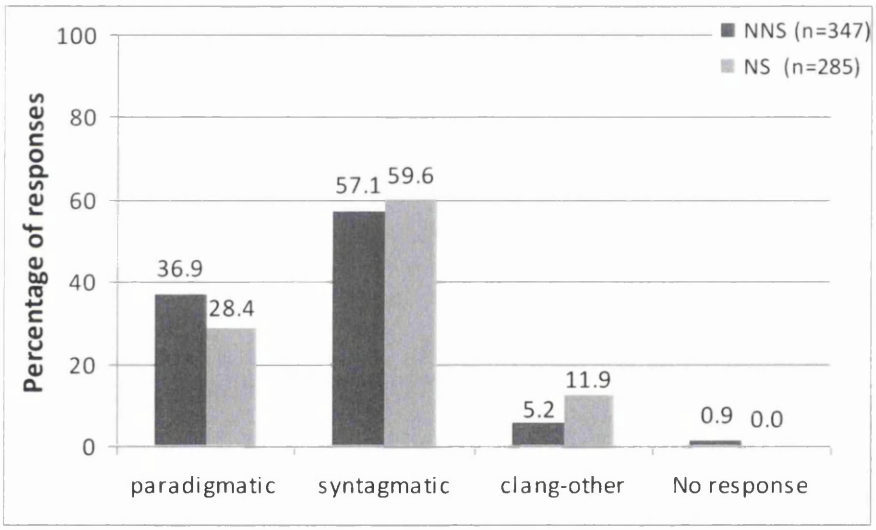


Fig 3.9b Percentage of NNS and NS response types for prompt words that elicited a VKS score of 5 (GH07)



With the highest level of word knowledge (Figs 3.9a and 3.9b) there do seem to be some similarities; in both studies the *clang-other* responses dropped still further with the meaningful responses (*paradigmatic* and *syntagmatic*) dominating. The replication data therefore agrees with Wolter's argument that as word knowledge increases there is a shift from a phonologically-structured lexicon to a meaning-structured lexicon. Where the two studies differ is in the proportions of *syntagmatic* and *paradigmatic* responses each group give at this level. In Wolter's study the NSs gave mainly *paradigmatic* responses (48.9%) whereas in the replication NSs gave mainly *syntagmatic* responses (59.6%). The NNS data does seem to follow a similar pattern in both the studies at this VKS level. In both studies for the NNSs the *syntagmatic* category dominated (Wolter, 54.1%; GH07, 57.1%),

followed by paradigmatic responses (Wolter, 35.4%; GH07, 36.9%). In both studies there were few NNS *clang-other* responses at the VKS 5 level and *no response*s were negligible.

3.4.3 Summary of results

The findings from the two studies do agree in some minor respects, although these are over-shadowed by major disparities, which lead us to question the reliability of the original study. When viewed in series, figures 3.5b - 3.9b show a systematic change (as do Figs 3.5a -3.9a) in the kind of responses that NSs and NNSs make to words along the unknown – known continuum. However, when each level of analysis from the original and replication studies are compared side-by-side, serious discrepancies are evident.

It is important to note that in the original study a Kruskall-Wallis analysis found a significant difference between the five VKS levels for both the NS and NNS data, suggesting that depth of word knowledge was a key indicator of response type. When the data in the replication study (Appendix 3.3. A) were analysed in this way the Kruskall-Wallis values were not significant ($p<0.05$), directly contradicting the original study. It might also be noted that Wolter calculated Mann-Whitney U scores between the NSs and NNSs at each VKS level. With non-significant values at VKS 1 and 2, Wolter argues that there is some support for his initial hypothesis that L1 and L2 lexicons are similar. As with the Kruskall-Wallis values, the U values calculated in the replication study (Appendix 3.3. B) disagree with the values calculated in Wolter's original study.

Both studies agree that phonology (as evidenced by the high number of *clang-other* responses, Figs 3.5a – 3.6b) plays a role when words are unknown/ partially known. Another area of agreement is that '*clang-other*' responses become less numerous as word knowledge increases and responses patterns become dominated by meaningful responses. Unfortunately the similarities between the findings of the two studies are limited, in the crucial question of how they portray the organisation of familiar words in a NS lexicon the two data sets differ. For words that are well known in Wolter's study the NS lexicons are paradigmatically dominated whereas the NS lexicons in the replication are syntagmatically dominated. Given that the two studies used the same materials and methodology and a very similar sample of learners I would have expected the results to be closer. A statistical comparison of the data from the two studies confirmed that the main results were related in some respects but different in others. In Table 3.2 the NNS data from the two studies were unrelated ($r=0.115$) although there were moderate relationships between the NS data ($r=0.637$; $r=0.712$).

3.5 Discussion

The results of this replication are mixed; in this section I will therefore address two main areas. Initially, with an aim of taking something positive from this study, some of the similar findings will be discussed. We cannot however overlook the conflicting findings, and will try to account for these in the second part.

3.5.1 Similar findings

The main area where the replication and original studies agree is that at the lower levels of word knowledge (VKS 1 and 2) phonology plays an important role. This finding is in line with more recent studies (Namei 2004; Fitzpatrick 2006) that found the proportion of '*clang*' ('*similar form only*' in Fitzpatrick's terminology) to be higher with the lower ability groups. There does therefore seem to be support for the outer circle of Wolter's DIWK model (slightly known words) being linked into the lexicon mainly through sound (or form) rather than meaning. This indicates that the first step in word acquisition is learning to deal with what a word sounds and looks like, a step that seems to come before dealing with the meaning of a word. The implication for language learners is that when they come across a new word they ought to listen to it and experiment with how it sounds before they get too involved with figuring out higher level aspects such as: which words it collocates with, words that have similar meanings or what grammatical restrictions it may have.

Another part of Wolter's study that was evident in the replication study was that there were some individuals who responded quite idiosyncratically. In his study, Wolter tells us that there was one individual who persisted in giving an unusually high number of syntagmatic responses. This individual was part of the pilot study and so his responses did not form part of the results. Had this individual's responses been included then given the small sample the results would have been quite different, closer perhaps to the results of the replication study. In the replication study two out of the nine native speakers also seemed to have this 'unusual' syntagmatically dominant characteristic which when taken together with the one in Wolter's study suggest that syntagmatically dominated native speakers may well be common. If it is indeed the case that for well known words native speaker groups are generally paradigmatically dominant but that there is also a subgroup of syntagmatically dominant individuals then the DIWK inner circle of well known words (which Wolter argues is paradigmatically dominated) needs to be rethought. Namei's hybrid model, explained in the introduction, does seem to offer a solution. She argues that

beyond the paradigmatically dominated level there is an even higher level of word knowledge: characterised by paradigmatic and "late-syntagmatic" responses. The data from both the original and replication study do seem to support the "evolution" of word knowledge suggested by Entwistle (1966) and supported by Namei (2004). They hold that as words progress from unknown to known they move through four stages: phonological →syntagmatic →paradigmatic → a mix of late syntagmatic and paradigmatic. The problem I have with this model though is that it is very difficult in practice to tell the difference between syntagmatic and late-syntagmatic responses. Entwistle (1966:128) argues that there is a qualitative difference between the syntagmatic and late-syntagmatic responses. According to her the initial syntagmatic responses are the more "stereotypical" ones such as *bright→morning* or *listen→to me* whereas the late-syntagmatic responses show "semantic enrichment" such a *butterfly→yellow* or *sell→out*. When I look back at my own data I have trouble in deciding on a principled basis precisely which syntagmatic responses seem to have been 'enriched'. This seems a difficult enough task even for the NS data where we might expect more stable responses and have norms lists with which to compare them. With the NNS responses, deciding whether a response is *syntagmatic* or *late-syntagmatic* seems far more subjective. As identifying late-syntagmatic responses with NNSs would probably require a norms list (to sort out the stereotypical syntagmatic responses) based on responses from advanced users of the NNS group in question, this approach would require considerable preparation work and is therefore not an easy option.

A curiosity, that might well have been overlooked were it not in both studies, is that even when words were rated as 'unknown' participants were occasionally able to give valid syntagmatic responses. In Wolter's study (Fig 3.6a) 7% of words that NNSs later claimed not to know elicited syntagmatic responses; in the replication study (Fig 3.6b) 7% of NS responses and 0.8% of NNS responses were syntagmatic. These responses might be dismissed as 'lucky guesses' although it is notable that there were virtually no paradigmatic 'guesses' at VKS 1, suggesting that something else may lie behind such responses. One possibility is that these words have actually been heard or read once or twice as part of a collocation or idiomatic phrase that has yet to be fully unpacked. It is conceivable that when such a word is heard out of context of the rest of the phrase it appears to be unfamiliar. To exemplify this point let us take a fairly infrequent word such as *ulterior*, which is for me a very peripheral part of my own lexicon. This is a word that I would readily associate with *motive,* to give the collocation *ulterior-motive* that I have probably heard in TV police dramas. If presented to me in isolation I would however be

hard pressed to define it or give any real sense of what it meant and perhaps after a moment's reflection actually decide that I didn't know it after all. When I give this word serious consideration I rate my own knowledge of it at VKS level 2. Although if I had just spent 20 minutes concentrating on a list of difficult words, and were pressed for time, I might well think that I had mistaken it for another word such as *interior* or *ultimate* and give it a VKS 1 rating. Although a small anomaly, the 'lucky guesses' in the two studies suggest to me that the VKS is not a particularly precise measure of word knowledge, levels 1 and 2 in particular seem to overlap. Even when this measure was used in a face-to-face format (the interviewer could also use body language, facial expressions and further questioning to help negotiate the level of knowledge assigned to each word) it was often difficult to decide which level of knowledge was most appropriate. I would imagine the written format, without any negotiation of levels with an interviewer, would be less reliable as the judgements would be left entirely to the participants: some of whom would be more cautious in their assessment of their own word knowledge than others. As well as problems between levels 1 and 2 there also seem to be problems between levels 4 and 5. This is evidenced by the lack of any NS responses in Wolter's study at VKS 4 and only 17 NS responses in the replication study. Judging by the larger number of responses at level 5 (Figs 3.10a and 3.10b) it could be argued that VKS 4 is redundant, especially for native speakers who having figured out a synonym for a word (level 4) can then hardly fail to put it into a sentence (level 5). The problem of the VKS not being a particularly accurate measure of word knowledge means that it limits the confidence that can be put into claims about the process of word knowledge development.

3.5.2 Conflicting findings

Having replicated Wolter's study using similar samples, the same word lists and the same methodology it is surprising that the results were in some ways markedly different, the most striking of these being the dominance of syntagmatic responses over paradigmatic responses. Considering such widely disparate results, one can only conclude that the testing methodology is unreliable. Putting aside the obvious problems of sample size that has already been commented on, I believe there are three further problems. These are the interview procedures, categorisation and the assumption that native speakers and non-native speaker groups are homogenous.

The first methodological problem concerns how the interviews were conducted. My concern here is with how much time and encouragement should be given to participants

when they encounter words that they don't know or are unsure of. In a study such as this, many of the words used are on the very periphery of participants' lexicons. Clearly a balance needs to be struck between giving respondents enough time to fully consider whether they know a word or not and keeping the test moving at a brisk enough pace to enable the test to be completed in a reasonable timeframe. As Nation (2007:43) notes in his paper reviewing fundamental issues in testing vocabulary "of all the factors looked at in this paper, the one that troubles me the most is learner attitude." If the testing process is too long and respondents lose interest then it is difficult to place much confidence in the data generated. If an unknown word is dwelt upon for too long then the participant's interest may wane. On the other hand, if the interviewer doesn't give the respondent enough time to fully consider the words which the respondent may have read or heard only once or twice (known in the very vaguest sense and consequently requiring more time to retrieve) then a *no response* will be recorded. The higher proportion of *no responses* in the replication study indicates that differing amounts of time and pressure to 'move on to the next one' were given to interviewees. It turns out that in Wolter's study more encouragement and time was indeed given to participants struggling with an item than in the replication (personal communication). The vagueness of the guidelines in the original paper was a factor, these were as follows:

> *There are no right or wrong answers so try not to take a long time considering your response. Try to respond to every word, even if you don't know the meaning.*
> (Wolter, 2001:51)

The speed at which an interviewer ought to move through the word lists is difficult to judge due to participant motivation and proficiency level varying with each individual, this means that the interview procedure needs to be flexible enough to account for this. That said though, I think the guidelines could be improved in order to ensure that similar amounts of time and encouragement are given to all participants. In the replication study the interviewees' personal schedule seemed to dictate how long they spent considering words they didn't know. Those that only had half an hour or so to spare were more conscious of time, whereas those who were interviewed after work/school (and in no particular hurry) spent more time considering their responses. Guidance on how much time to spend on each item (a maximum of 30 seconds per item?) or even a timeframe within which to complete the whole series of tests and interviews (within 30 minutes?) would I believe improve reliability.

An alternative way of explaining the difference in results obtained from these two studies could be due to the difficulty in accurately categorising responses, the second of

our methodological problems. In the replication study responses were found to vary in difficulty when it came to categorise them. Let us consider some responses to 'cherish' by the students in the replication study:

cherish→love *cherish→children* *cherish→city*

cherish→cherry *cherish→young*

The first example *cherish→love* is difficult to classify. If we think of both the stimulus and response as verbs, then as they are synonymous we would classify them as *paradigmatic*. However, the response may have been made due to an awareness of the marriage vows "to love and to cherish, to have and to hold" which might lead us to class the response as *syntagmatic.* The classification becomes complicated when we realise that the response *love* can be associated in a meaningful (though different) way as either a verb or a noun. In the next example *cherish→cherry* we seem to be on firmer ground when we declare that this is a *clang* response, there is little meaning here and we can probably assume that the link is made due to the similar initial /tʃɜ:/ sound. With *cherish→children* however there is less certainty, there is probably some meaning (parents usually claim to cherish their children) but the /tʃ/ sound might also lead us to think it is phonological. If we accept that *cherish→children* has some kind of meaningful link then do we also accept *cherish→young?* Or do we decide that the link is now too tenuous and mark it as *other*? How about *cherish→city*, can we not also cherish our city as we cherish our children?

Wolter also experienced problems with categorisation, such as how to classify *tolerate* as a response to *confine*. He tackled this issue by using two judges. In the replication study a second judge was used for ambiguous responses, although this did not seem a satisfactory solution. Quite often both judges got stuck on the classification of the same word with no clear resolution. Even if the two judges can come to an agreement and therefore bring some internal reliability to the study, such results might not be comparable to other studies. Raters within one study might, after a little negotiation, agree on some classifications although another group of raters in a different study might agree to classify such responses differently. This problem is exacerbated in L2 studies where the participants often have very different backgrounds to the raters. A middle-aged English man second guessing another middle-aged English man's thinking is quite different to a middle-aged English man trying to second guess the thinking of a young Japanese woman.

Another concern I have with Wolter's classification system is putting the *other* (unclassifiable or erroneous) responses in with the *clang* responses to make the *clang-other*

category. While in L1 studies using familiar stimulus words the *other* responses are usually negligible this is more of an issue with studies using low frequency items or non-natives: more mistakes and misunderstandings are inevitable. In fact the significant role of phonology that both this study (and Wolter's) claim to be a feature at the early stages of word knowledge is probably slightly overstated, as some of the responses in this category are not phonological but just responses that couldn't be classified as anything else. A stronger claim for the role of phonology could be made if it had its own category. A more detailed categorisation system would be preferable, such as the one proposed by Fitzpatrick (2006) that she claims, "provides a more precise insight into the differences between L1 and L2 association patterns" (p121). The 18 sub-categories in this system map directly onto Nation's aspects of word knowledge (2001:27), are well defined and have a discrete category for every potential response. Another positive feature of this system is that the subcategories are organised into three main categories (meaning-based, position-based and form-based) which are similar to the traditional categories (paradigmatic, syntagmatic and clang): some comparison with past studies is therefore still possible.

The studies by Fitzpatrick (2006; 2007) bring us to our third methodological problem, how to deal with idiosyncratic behaviour. As has already been noted, in the replication study two NS's in particular gave mainly syntagmatic responses (86% and 73%), which was surprising. When one of these individuals was questioned later about how his responses he commented "if it was a long word then I jumped to an association before the word had finished". For example, with the word "temporary" he replied "secretary". Before he had heard the whole word he had made an association with part of the word, "temp". Even though he stated his awareness that this was "not the best link with temporary" he still said it because he had been instructed to "respond with the first word that comes into your head". Again, perhaps the methodology could be more specific so that interviewees only respond when they have heard the whole word. Such idiosyncrasies do however seem common, Wilks & Meara (2007) found that testees often used quite different strategies to carry out a task, which were masked by and hardly correlated at all with the group norms. Rather than taking idiosyncratic behaviour to be the exception, their study found that testees had a "surprising degree of individual variation" and questioned the validity of using group norms. Instead of attempting to tweak the current methodology in order to get around the 'exceptions' perhaps a more radical approach which embraces individuality ought to be adopted. The idea of "individual profiling" proposed by Fitzpatrick (2007) rejects the traditional method of analysing responses based on

membership to a particular social group (NS group, NNS group, high IQ group etc.) and argues in favour of analysing responses from the perspective of the individual. She argues her findings justify such an approach as "considerable variation was found in the response preferences, implying that subjects are not homogeneous in their response behaviour. However, individual response behaviour is consistent" (Fitzpatrick, 2007: 319).

3.6 Conclusions

The results of this replication study add little clarity to the already blurred picture we have of how L1 and L2 mental lexicons develop. Despite carefully following Wolter's methodology and using similar samples the main results of the two studies were different in a number of respects. This means that we cannot place confidence in the results of the original study by Wolter or the DIWK model it proposes. Using the same statistical tools to analysis the data there was little evidence to support the DIWK model. There was support in the replication for the *L2 lexicons are similarly structured to L1 lexicons* hypothesis, although as this partially conflicted with the original study little can be claimed. An idea that received support in the both studies was that phonology plays a role with words that are only partially known. A role which appears to decrease as meaningful knowledge of that word increases. Even here though, confident claims cannot be made due to problems with the classification category for phonological responses and a query over the accuracy of the measure of word knowledge depth (VKS). As well as the small number of participants used in this study, that was picked up in Chapter 2, it is suggested that the wide disparity in the results is partly due to problems with the interview process and the categorisation system. These methodological problems are not insoluble, this replication therefore does not lead us to completely abandon the search for a framework within which to consider how the mental lexicon is structured. The research methodology can be improved; adopting a more precise categorisation system, such as the one proposed by Fitzpatrick (2006), appears to be a step in the right direction. There is however a more fundamental challenge to overcome before more research in this area can progress, that is how to deal with idiosyncratic behaviour. As the results of this replication study and recent research suggest, idiosyncratic behaviour seems common and casts doubt over the traditional practice of analysing responses from a group perspective. There seems to be a good argument in favour of a method of data analysis along the lines of what Fitzpatrick (2007: 328) calls "individual profiling". In the following chapter we will therefore attempt to apply Fitzpatrick's approach in order to assess its viability with L2 learners.

Chapter Four: Exploring individual learner profiles through word association tests.

4.1 Introduction

As detailed in Chapter 2, researchers have been using word association tests as a way to understanding the mental lexicon of language learners for over 50 years. Unfortunately this research has not provided much in the way of agreement and is characterised by a series of conflicting findings. Yet more conflicting findings were found in Chapter 3 with a replication of Wolter's 2001 study. Continuing with traditional methods of collecting and analysing word association data is therefore becoming increasingly difficult to justify. In an attempt to find a more productive line of research this study explores a quite different approach: individual profiling.

The development of Fitzpatrick's *individual profiling* approach can be seen in two papers (Fitzpatrick 2006, 2007). Recapping the main points already covered in Chapter 2, in her 2006 paper she introduced an innovative categorisation system and questioned the use of grouped data. This was based on a study of L1 and L2 English users. In her 2007 paper, the categorization system was refined and response data for native English speakers was analysed from an individual rather than a group perspective. Her findings showed that there was a lot of variability between responses from members of the same group although individuals responded in a consistent way. Following Fitzpatrick, this study applies *individual profiling* to Japanese college students. The main question that is being asked at this point is whether this approach can provide consistent response data that will help to shed light on the organisation of the mental lexicon.

The data for this study was data collected in July 2008; a similar report based on the same data can also be found in Higginbotham (2010). Although the views I put forward in that report remain broadly the same, with the benefit of a few years reflection, a slightly different analysis is presented here with some refinements to the argumentation. Without the tight restrictions that journals impose on length, the account given here allows for a deeper discussion into the issues surrounding word associations. The reader also benefits in being able to place this particular study in the context of the series of studies that make up this research project.

4.2 Overview of the study

The present study initially aims to verify Fitzpatrick's (2006 & 2007) claim that NS and NNS groups are not homogenous and that research therefore ought to analyse word association response data from an individual perspective. This study uses Fitzpatrick's 2007 classification system and method of analysis but differs from that study in that it is in a Japanese context and focuses on low ability L2 learners. Another difference is that this study also aims to establish whether individual response characteristics change (or not) when learners are asked to respond to words that differ in terms of their frequency within the language. This study explores the response characteristics of learners to words selected from two word-frequency bands. From the research of Söderman (1993b) and Wolter (2001) it seems reasonable to expect some changes in how learners respond to less frequent words. Frequency has long been argued (Deese, 1965; Cramer, 1968; Stolz & Tiffany, 1972) to be the most likely of the word related variables to have an influence on word association responses. Just how much of an influence frequency has on word associations is largely unknown though, as most word association studies have concentrated on high frequency stimulus words. Not only is frequency the most likely variable to affect responses but given the availability of large modern online corpora (the British National Corpus – BNC and the Corpus of Contemporary American English – COCA; Davies, 2008) it is also one that we can now control to a certain extent.

4.3 Research questions

Alongside the general goal of establishing whether the individual profiling approach can provide consistent response data, this study will explore two specific areas of interest: the issue of group homogeneity and the role of frequency. The two research questions are:

1. Do L2 learners, with a similar background and L2 ability, respond to word association stimuli in a homogenous way?

2. Does the frequency of the stimulus word affect a learner's characteristic response pattern?

4.4 Participants

In this study 60 Japanese college students with similar learning backgrounds and L2 language ability were asked to participate. These students were in two general English classes within the same ability stream, based on a university placement test. The first and second levels of The Vocabulary Levels Test- VLT (Nation, 1990) were used to confirm students' vocabulary ability prior to the word association tests. The students in this study were of a low level (averaging 76.8% at the 1000 word level of the VLT and 62.2% at the 2000 level). It was therefore expected that they would not know all the words in the lists and so a completion threshold (50%) was established. Of the initial 60 students, after language abilities were assessed and the number of responses counted, 10 students' responses were rejected from the analysis; the findings therefore consist of responses by 50 students.

4.5 Materials

4.5.1 Stimulus word lists

Two word lists (see Appendices 4.1 & 4.2) were created from The BNC. The first list (prompt word list 1 – PWL1) was a selection from the 0 – 500 frequency word range, the second list (prompt word list 2 – PWL2) was selected from the 500 – 1000 frequency word range. It was expected that items in PWL2 would be less well known than the items in PWL1, as items in PWL2 were selected from a list of lower frequency words. A completion threshold was therefore established to allow a fair comparison between the students' two profiles. Those students whose erratic and blank scores totalled more than 50% were rejected. It was felt that an individual profile based on a sample of less than 25 words would not truly represent how the person characteristically makes associations between words. Based on VLT scores these frequency ranges were judged suitable for these students. The word lists were piloted with a similar (in terms of age and ability) group and unsuitable words cut from the list leaving two prompt word lists (PWL1 and PWL2) of 50 nouns each. In filtering the word lists the advice of Wolter (2001) was followed. Unsuitable words were those that:

- Strongly associated to just one other word (such as *dog* which would probably give the response of *cat*). Strong associates were identified using the Edinburgh Associative Thesaurus (Kiss et al, 1973), a database of native speaker associative norms (retrieved from http://www.eat.rl.ac.uk/).

- Common collocates of Japanese words (e.g. glass / ガラス).
- Difficult to classify due to belonging to more than one word class.
- Too difficult for respondents.

In the word association test, students were instructed to write the first English word that they thought of when they read the prompt words. The instructions, with an example, were written in Japanese at the top of each prompt word list. These instructions were read aloud to the group with a few minutes given, prior to the test, for students to ask questions about the procedure. This was done to ensure all the learners understood clearly what they had to do, as for them it would be an unusual kind of test.

4.5.2 Classification

The classification system that has generally been used is the broad classification of word association responses into either: paradigmatic, syntagmatic, or clang/phonological responses (Söderman, 1993; Wolter 2001; Bagger-Nissen & Henriksen, 2006). This broad classification system was not used due to problems raised in Chapter 2 and also the replication study (Chapter 3) with classifying ambiguous items. The classification system that was used follows Fitzpatrick 2007 (Appendix 4.3), which subdivides the three main categories of Meaning-based (Paradigmatic), Position-based (Syntagmatic) and Form-based (Clang/phonological) into nine subcategories. These subcategories were: *defining synonym, specific synonym, lexical set/context related, conceptual association, consecutive xy collocation, consecutive yx collocation, other collocation association, change of affix* and *similar form only*.

Immediately after the word association test, students were given partial retrospective interviews to help with classification. Students were only asked about items that on a cursory inspection seemed ambiguous and would therefore be difficult to classify. A full retrospective interview (Wolter, 2001) was not done due to perceived benefits in terms of time (collecting comments on responses while students' thoughts were still fresh) and the realisation that many responses were unproblematic to classify and therefore did not require further explanation.

Having classified the responses, a further step was taken before the main analysis in order to filter out unhelpful stimulus items that the initial screening had failed to identify. The items were analysed in terms of the most frequent response for each item within the group. Those items that generated very strong primary responses within the group were

rejected, as such responses probably relate to the associative strength of the word itself rather than an individual's associative preferences. Two items were rejected from PWL1 (*student, bank*) leaving 48 items available for further analysis. Four items were rejected from PWL2 (*page, blood, hospital, difficulty*) leaving 46 items.

4.6 Results

In this section the following is reported:

- 4.6.1 The completion rates of the prompt word lists.
- 4.6.2 General trends in the group.
- 4.6.3 Individual profiles: example case studies.
- 4.6.4 Profile proximity ranking
- 4.6.5 Analysis of individuals' dominant categories

4.6.1 Completion rates of PWL1 and PWL2

Having given the PWL1 and PWL2 tests to all the students, the papers were initially sorted and those students who answered less than 25 from either of the initial words on the lists were rejected. Exactly how many responses are necessary to get a representative sample is unclear, although it is assumed in this study that over 25 responses is enough to identify an individual's main response characteristics. Of the 60 who took the tests this left 50 papers that had been satisfactorily filled in. As shown in Table 4.1 most of the students made far more responses than the minimum threshold (25), most made between 35 - 45 responses.

The completed lists consisted of 46 words (PWL1) and 48 words (PWL2) that would be used in the analysis, most of these words were known to some extent by most of the students. As PWL2 contained less frequent words it was not surprising to find that average completion rates were lower in this list. The completion rates for PWL2 ranged from between 29 responses (60%) to 48 responses (100%). Of the students accepted for analysis, most knew the words in PWL1 quite well but there were generally two or three unknown words in PWL2. Due to the different number of responses to each list, the analysis is based on the percentage of responses.

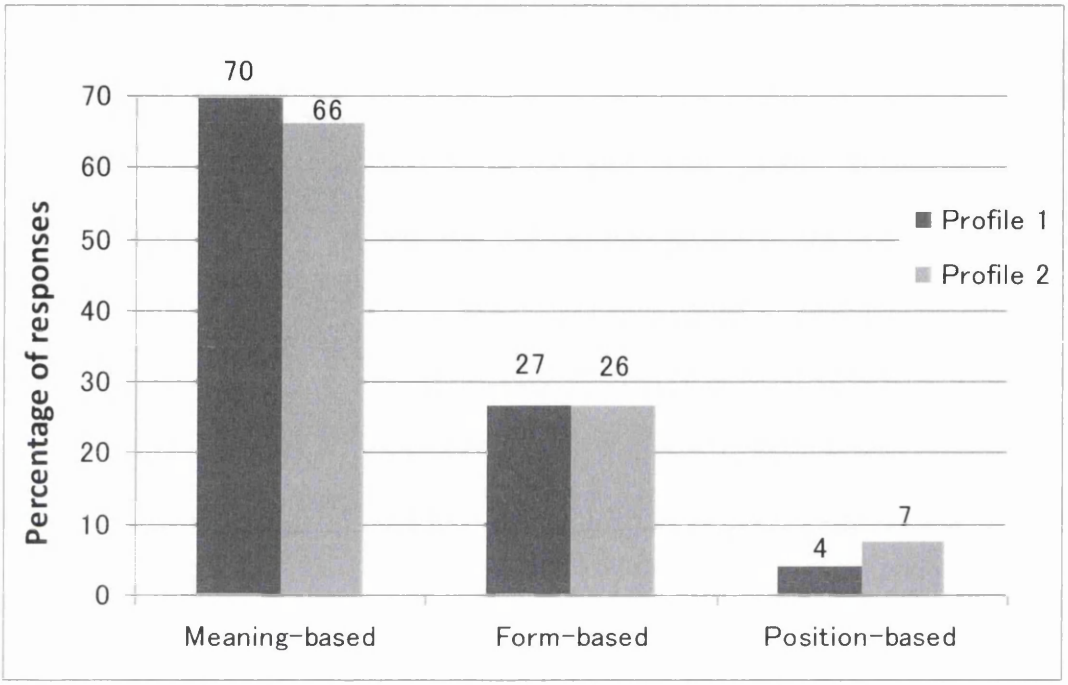
Table 4.1: Prompt word list completion rates for those accepted in the study

PWL1	96%
PWL2	88.3%
Average completion rate	92.2%

4.6.2 General trends in the group.

In order to compare the data from this study with previous studies Fig 4.1 shows how the group as a whole responded to the two word lists in terms of the three main categories. Profile 1 shows the responses to PWL1 and Profile 2 shows the responses to PWL2.

Fig 4.1 Responses in main categories



As we can see from Fig 4.1 the Meaning-based category dominates the responses for both sets of profiles. Although not synonymous, if the Meaning-based and Form-based categories are viewed as broadly overlapping the paradigmatic and syntagmatic categories used in previous studies (Söderman, 1993; Wolter, 2001) some comparisons can be made. One point is that the large number of Meaning-based (paradigmatic) responses for these low ability students does not sit well with the general concept of a shift from syntagmatic to paradigmatic responses as proficiency in an L2 increases. The idea that L2 students generally make syntagmatic (Form-based) associations in the early stages of their language development as put forward by Söderman (1993) is not upheld. This study seems to be more in line with the findings of the replication study and Bagger-Nissen & Henriksen (2006) that challenge the concept of a syntagmatic – paradigmatic shift. Such speculation does not however lead us to any further enlightenment; it merely stirs up an already murky pool. As has been previously argued a more useful line of enquiry would be to view these students not as a homogenous group, but as individuals.

Before going on to consider individuals there is another important point to be made; that is, if the analysis is restricted to the three main categories then a lot of data becomes obscured. An analysis of the first case study (Student 1) in Fig 4.2 demonstrates this.

Fig 4.2 Responses classified by main categories: Student 1

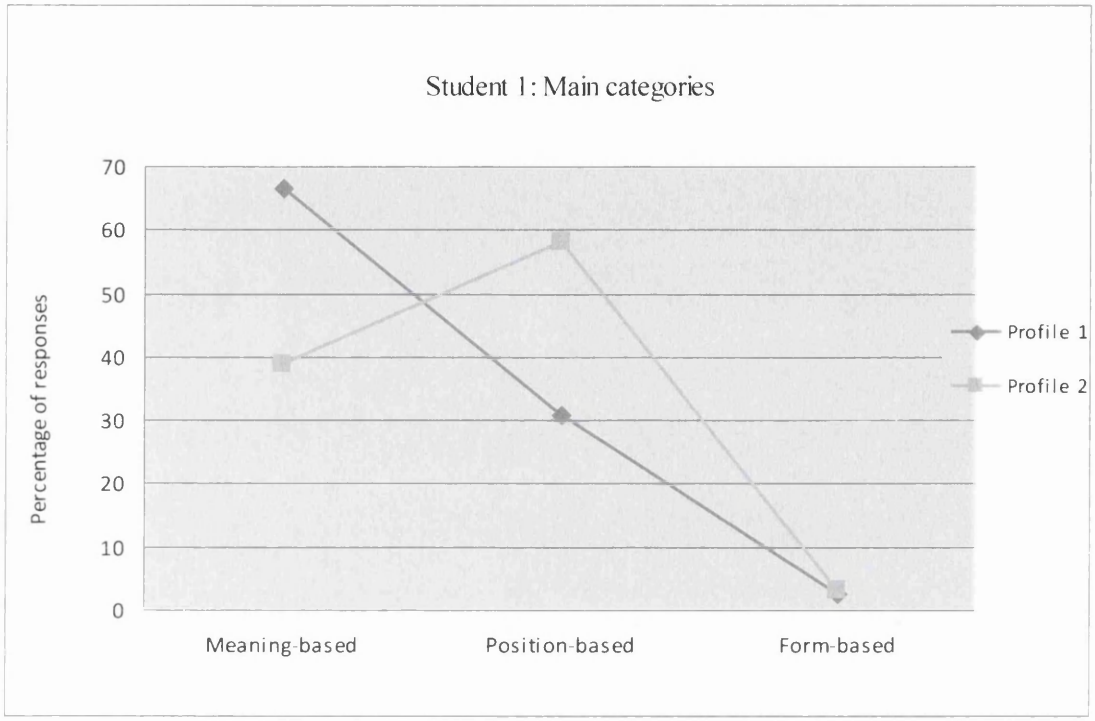
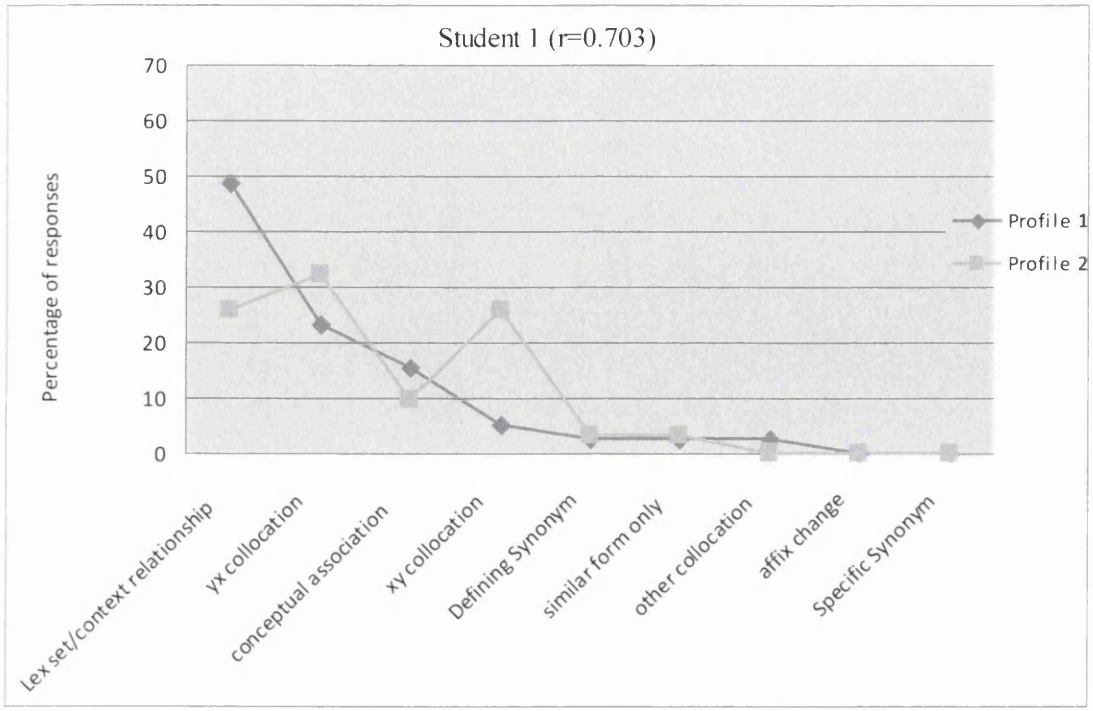


Fig 4.2 seems to indicate that this student made different types of response to the two word lists; the graph shows two very different looking profiles. In Profile 1 there are more Meaning-based responses and in Profile 2 there are more Position-based responses. If however the same student's profiles are viewed from the perspective of the subcategories (Fig 4.3) it can be seen that the two profiles are actually not so different. A Pearson's correlation was calculated between the two profiles to determine the relationship between the type of responses to high and low frequency stimuli. For Student 1 there was a fairly strong positive correlation between responses to PWL1 and PWL2 ($r=0.703$, $N=9$, $p<0.05$).

Fig 4.3 Responses classified by subcategories: Student 1



When one examines the details of the subcategory data (Fig 4.3) the reason for this apparent discrepancy in the results becomes apparent. In both profile 1 & 2 there are two dominant categories; *same lexical set/context related* and *yx collocations* that share most of the responses. As these two subcategories are within different main categories (*same lexical set/context related* responses are Meaning-based and *yx collocations* are Form-based) variation within these two subcategories alone will result in the pattern shown in Fig 4.2. Due to this problem of data becoming misleading when rolled up into a larger category, it is better to view the individual profiles in terms of their subcategories. It is therefore the detailed profiles that will form the basis of analysis for the following learner lexicon case studies. In the graph above it might be noted that the subcategories have been presented in order of dominance; the largest subcategories in Profile 1 are on the left side of the graph with the smaller (and unused) categories to the right. This convention will be maintained throughout this thesis, allowing easier identification of the dominant subcategories in each set of profiles.

4.6.3 Individual profiles: example case studies

In general, when each individuals' responses to high frequency words (Profile 1) was compared with the profile created from responses to less frequent words (Profile 2) the two profiles were found to be highly correlated. The proximity of these profiles was confirmed

through calculating the correlation coefficient between the percentage of responses in each of the subcategories. The correlations showed a range of relationships, from moderate (r = 0.507) to extremely strong relationships (r = 0.990). This was interpreted as meaning that most individuals generated two profiles that were similar in shape.

In order to understand what these individual profiles look like and attempt to establish a threshold value below which profiles should not be considered 'similar', four more individual profiles are examined in detail. The examples chosen give an indication of the range of profiles that were observed within this particular cohort. Some individuals' profiles were statistically very close and others were not so close. Some of the individuals characteristically gave profiles with two dominant subcategories (such as Student 1) where as others had profiles that were overwhelmingly dominated by just one type of response.

Student 2:

The second student was selected as an example of a student with two very close profiles. With a correlation of 0.990 (Fig 4.4) Profile 1 cannot really be seen in the graph as it is hidden behind Profile 2. This student characteristically gives responses that are from the *same lexical set/context relationship* subcategory. In both profiles 44.4% of this student's responses were in this category. Student 2's second most numerous response type was *conceptual associations* (30.6% in Profile 1 and 25.9% in Profile 2). The correlation of 0.990 indicates that this individual gave virtually the same type of responses to both the high frequency and less frequent prompt words.

Fig 4.4 Responses classified by subcategories: Student 2

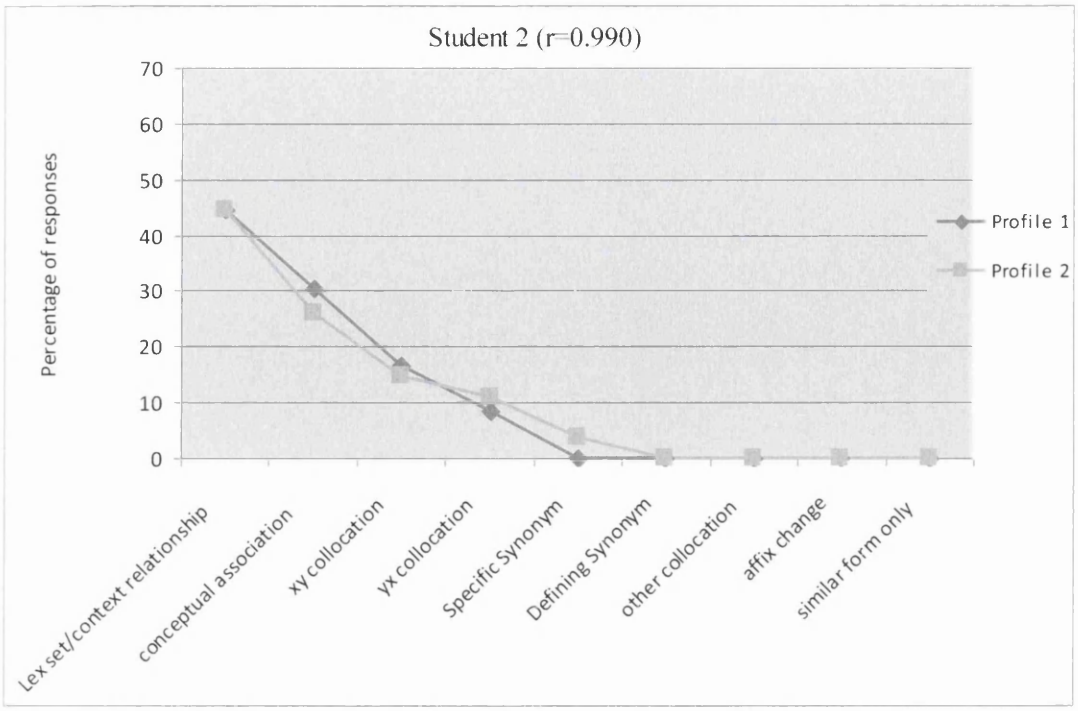
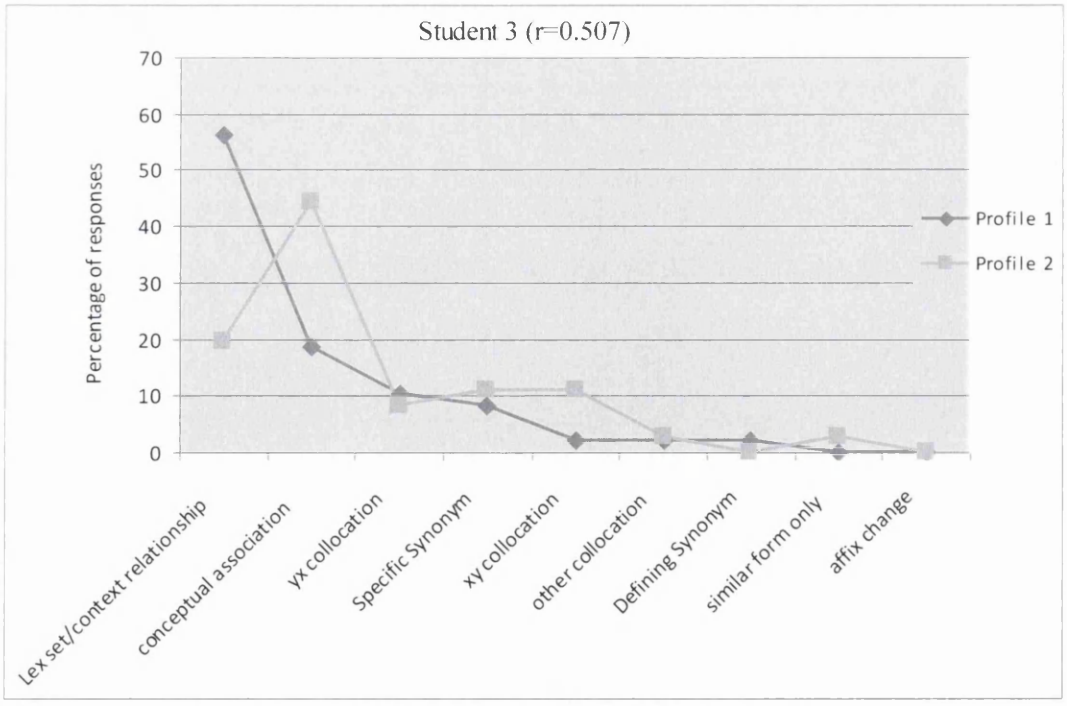


Fig 4.5: Responses classified by subcategories: Student 3



Student 3:

The next example is a student (Fig 4.5) whose profiles showed the weakest correlation of those tested (0.507). When the two profiles are viewed from the subcategory level the two dominant categories are quite different. In Profile 1, Student 3 gave a lot of *lexical*

set/context relationship responses, some *conceptual associations* and also quite a few *specific synonyms*. In Profile 2 *lexical set/context relationship responses* were much less common and *conceptual association* responses dominated instead. In both profiles there were four subcategories with virtually no responses, these account for the slight statistical relationship that the correlation coefficient indicates. While not totally unrelated, I would argue that the divergence between the two dominant sub-categories (accounting for over 60% of all responses) means these two profiles ought to be considered 'dissimilar'.

Student 4:

As can be seen in Fig 4.6, Student 4's two profiles are close. In Profile 1 the two main subcategories are *conceptual associations* and *xy collocation* responses. In Profile 2 these two groups are again dominant, although Student 4 slightly favoured *conceptual association* responses. The main area of difference can be seen in the increase in *similar form* responses; a jump from 6.4% in Profile 1 to 18.4% in Profile 2. The high correlation of 0.864 indicates that this individual's responses to both high frequency and less frequent prompt words are very similar.

Fig 4.6: Responses classified by subcategories: Student 4

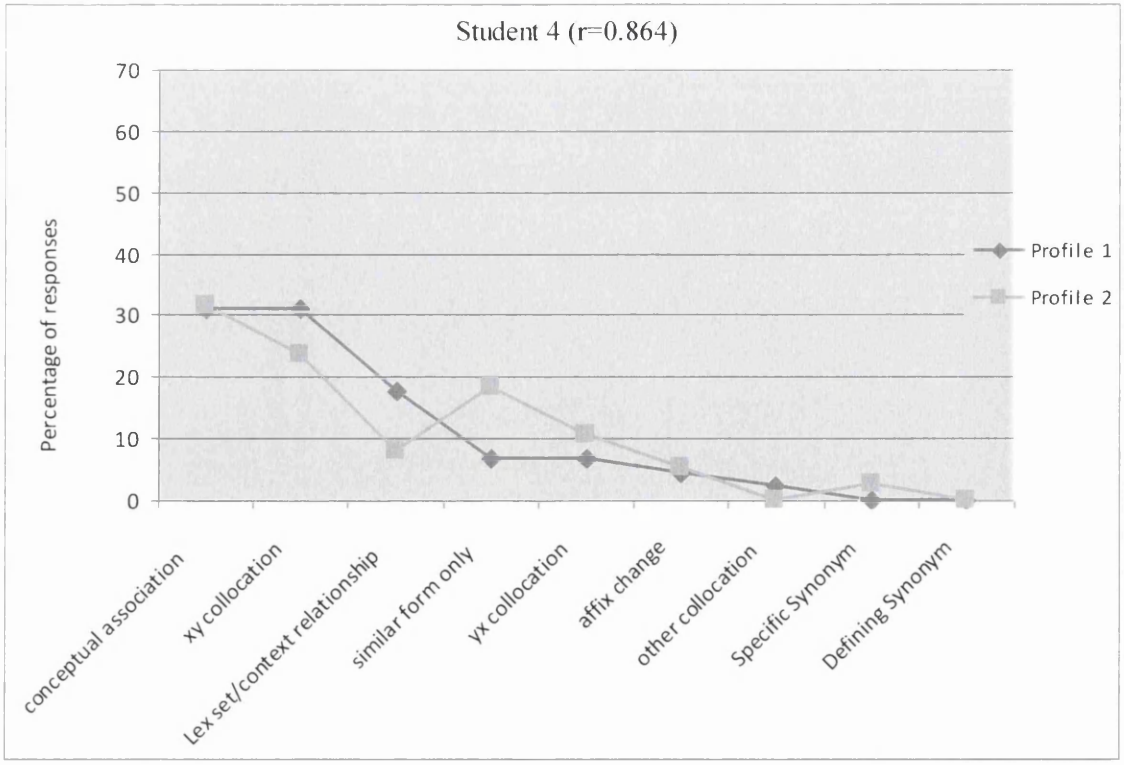
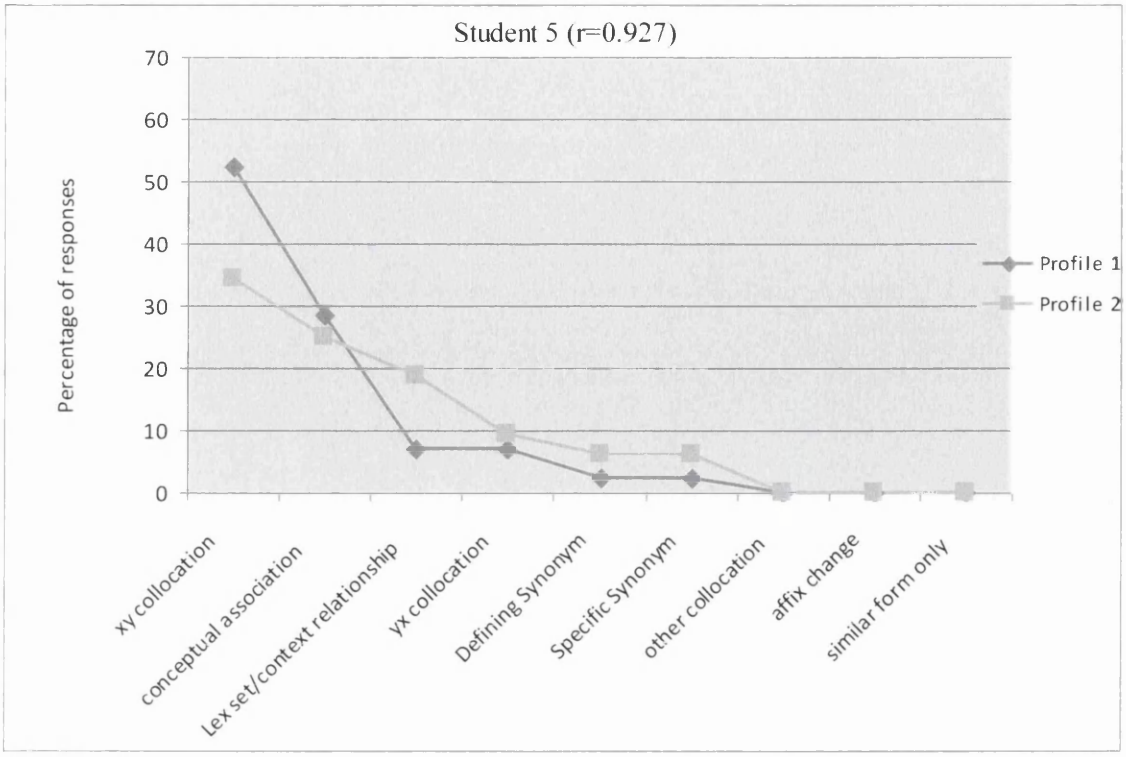


Fig 4.7: Responses classified by subcategories: Student 5



Student 5:

In Student 5's case (Fig 4.7), the *xy collocation* subcategory dominates in both profiles; this learner also favours *conceptual associations*, which was the second largest category in both profiles. This student gave no Form-based categories. The high correlation coefficient of 0.927 indicates that this individual often gives similar types of response to both frequent and less frequent prompt words.

An important point to come out of these case studies is that a wide variety of student profile types were observed. Some (such as Student 2) favour *lexical set/contextual relationship associations*, some (Student 1) favour a mix of *lexical set/contextual relationship and collocations* and others (Student 5) favour a mix of *xy collocations* and *conceptual association* responses. To demonstrate the amount of variation in the group nine profiles were randomly selected from the database and each profile was compared to the other eight profiles using a chi-square test. Of the 36 possible pairs (Appendix 4.4) 30 were found to differ significantly ($p=<0.05$). With 83% of the sampled profiles showing a statistical difference, the idea that Japanese low-level learners respond homogenously can not be accepted. From the high correlations between each individuals' profiles it can be seen that even though there is variation within the group, individuals tend to give similar types of responses to words of different frequency. Many of the individuals had a very high

correlation between the type of responses they gave in the two word association tests. The results of this study therefore support the findings of Fitzpatrick (2007), who also found a variety of response preferences between participants even though the response behaviour of individuals was internally consistent.

4.6.4 Profile proximity rankings

So far the distance between the profiles has been considered in rather vague terms; *very close, close* and *similar*, it would be helpful to define these terms more explicitly. As research in this field is still at an early stage there are no particular guidelines defining what constitutes a *similar* or *dissimilar* profile. To judge the similarity between two profiles Fitzpatrick (2007, 2009) calculated the Euclidean distance. As explained in Appendix 4.5, students' profiles were not compared using this metric but by calculating Pearson's correlation coefficient. A measure, that I would argue, is slightly better at comparing the similarity of profile shapes.

Considering the profiles discussed previously it is proposed that a correlation coefficient of over 0.8 be considered as showing a *very close* match, a correlation of between 0.7 and 0.8 a *close* match, a correlation of between 0.6 – 0.7 as *vaguely similar* and correlations under 0.6 be considered as showing that the profiles are *dissimilar*. In Table 4.2, when these boundaries are applied, the vast majority (72%) of students in this study fall into the *very close* category, with nearly all (90%) being *close* or *very close*.

Table 4.2 Proximity rankings

Correlation between profiles 1 and 2	Definition of profile proximity	Number of students' profiles at each level
> 0.8**	Very close	**36**
0.7 – 0.8**	Close	**9**
0.6 – 0.7*	Vaguely similar	**4**
<0.6	Dissimilar	**1**
Total number of students		50

**p = <0.001, *p = <0.05

The conclusion that can be drawn from Table 4.2 is that subject profiles are internally consistent. When compared statistically, each half of an individual's profile usually correlates strongly with the other half.

4.6.5 Analysis of individuals' dominant categories

With a negative answer to research question 2, most students don't change their response types when asked to respond to stimuli from different frequency bands, the next step was to identify what response was the most characteristic for each individual. As the dominant response usually accounts for less than 40% of individuals' total responses, it was decided to combine the top two responses of each individual in order to obtain a higher coverage of responses. This potentially creates a more complicated picture as by combining each of the nine subcategories to all of the other subcategories we could theoretically make 36 subcategory pairs, although (as is shown by Table 4.3) in practice not all of these are needed. In this study for example there were seven dominant pair categories; this gave a far greater coverage of an individual's response (response coverage averaged 66.19% for the paired categories).

Table 4.3 Dominant Pair Categories

Pair Category	Student responses to both PWL1 and PWL2 combined (n=50)
Lexical set/context relationship + Conceptual association	**37**
Conceptual association + XY Collocation	**4**
Conceptual association + YX Collocation	**3**
Lexical set/context relationship + XY Collocation	**3**
Lexical set/context relationship + Affix change	**1**
Lexical set/context relationship + YX Collocation	**1**
XY Collocation + YX Collocation	**1**
Average coverage of dominant pair categories	**66.19 %**

Many of the students gave responses that were dominated by *lexical set/context relationships* and *conceptual associations* (Table 4.3) in their overall responses to the word association prompts in PWL1 and PWL2. It ought to be noted that the initial subcategory in each pair in the table is not necessarily the dominant subcategory. Within the top category for example some of the students made more *lexical set/context relationship* responses whereas some made more *conceptual associations*.

In order to see if the student's dominant preference pairs changed when they responded to the lower frequency words (RQ2) the students' responses were reanalyzed according to their two most dominant response preferences to both PWL1 and PWL2. It was found that most of the students (78%) did not change their response characteristics. The student who for example answered with a combination of *xy collocations* and *yx collocations* in the first word association test continued to respond with the same combination in the second test. It might also be noted that of the 22% who had a different top pair combination between profile 1 and profile 2, virtually all students had one particular category that was ranked within the top two in both their profiles.

4.7 Discussion

The results show that responses in the group were not homogenous and also that the characteristics of their profiles did not change when learners were presented with either high frequency or lower frequency words. There are however a number of areas of interest that warrant further discussion: the difference in the frequency of the prompt words, the word class of the stimulus items and identifying unhelpful stimuli.

4.7.1 The frequency effect

The study seems to show that frequency has little effect on responses, there is however a problem with this claim. The problem is that the lack of a frequency effect could be due to the two frequency bands being too close.

The first list was based on words taken from the most frequent 500 words in the BNC. The second list was taken from the 500 – 1000 frequency band. These two frequency bands were chosen as the learners in question would have had difficulty in coping with lower frequency words, many of the words even within these 'easy' lists being unknown. The low completion rates for many of the students on the PWL2 (10 were rejected for having completed less than 50%) suggest they would have had even more of a struggle to complete the word association tests had the prompt words for PWL2 been selected from a lower frequency range. The results of the VLT and pilot test indicate that for this cohort using prompt words from the 1500 – 2000 frequency band would have been beyond them. That said though, the two frequency bands are very close (perhaps overlapping) and more convincing claims could have been made about the role of frequency had there been a larger gap between the two frequency bands used. Many would agree with Nation (2001) that all words within the most frequent 2000 might be termed 'high frequency' and so it

was perhaps unreasonable to expect to find a difference between the two bands used in this study. The problem of the close frequency bands is further exacerbated when we consider how words are actually counted. A study by Gardner (2007) highlights the difficulty in programming computers to accurately count words. Problematic word types include:

- morphologically related: is *climber* merely a derivation of *climb*?
- homonyms/polysemes: a *bear* is an animal, *bear* also means to carry.
- multiword units: is *Prime Minister* one word or two?

Given that some types of words are difficult for computers to accurately classify (and therefore count) it would seem that corpora cannot be relied upon for precise frequency rankings. On reflection it would have been better to have used word lists derived from two quite distinct frequency bands, a comparison between prompt words selected from say the 0 -1000 band and 5000 – 6000 band would have allowed for stronger claims. Of course, a similar experiment that used prompt words taken from such diverse frequency bands would require higher ability students than those in this study.

4.7.2 The word class effect

The next factor that could have influenced the results of this study was the decision to only use nouns as prompt words. As shown in the study by Bagger-Nissen & Henriksen (2006) on the effect of word class on word associations by English learners, nouns tend to generate more paradigmatic responses. This phenomenon is also documented in L1 studies such as Deese (1965) and Entwisle (1966). It is therefore not unlikely that the high number of paradigmatic/meaning-based responses (PWL1= 70%; PWL2=66%) in this study (see Fig 4.1) is in part due to the use of nouns as prompt words. Further studies using different word classes are recommended to ascertain how much (if at all) the use of nouns has exaggerated the number of Meaning-based responses.

4.7.3 Filtering unhelpful prompt words

An area that proved complicated was identifying prompt words with strong associates, such words are unhelpful as they do not reveal a persons characteristic behaviour. As noted by Meara (1983) very high frequency nouns have a high proportion of strong associates and so considerable effort was put into weeding them out using a pilot test, pre-test screening of words based on native speaker norms and then finally a post-test screening of item responses.

Although the Edinburgh Associative Thesaurus – EAT (Kiss et al., 1973) was

considered useful in flagging up words that might mask characteristic response types, this database ought to be used with caution. The data was collected in the late 1960's, so when for example we search for the word *help* we find one of the common associates is *Beatles.* I doubt whether university students (British or otherwise) would currently associate the word *help* in this way. The EAT database needs to be considered in its original context, associations made by L1 university students in the UK during the 1960's. As the students in this study are all using English as an L2 at a low level we need to be cautious about assuming that words considered strong associates in an L1 are also strong associates for L2 learners. A word such as *cup* for example would probably be known by nearly all low-level Japanese learners, although they wouldn't associate it with *saucer* (as the EAT data suggests most L1 users would) because they wouldn't be exposed to this word until later on in their studies, if at all. In this study a post-test analysis of the responses was done in order to identify stimulus words that might not be strong associates for native speakers but are actually strong associates for the particular group being studied. The word *bank* for example in PWL1 had to be rejected even though the L1 norms list tells us there are a great variety of words (98) that *bank* is usually associated: *money, account, book, manager, clerk, cheque, overdraft, robbery* etc. Considering the low frequency of some of these potential response words and the difference in banking customs between Japan and England, one can understand why most responses were *money.* Interestingly, some prompt words such as *house* (EAT data suggests this item might strongly associate with *home*) were not problematic.

When using native norm databases such as EAT, one needs to consider not only that each prompt word has a wide variety of possible associations but that these potential associative words are words that students are likely to know. This point is particularly pressing when the respondents only have a relatively small L2 lexicon. It is also important to be mindful of the ever changing sociolinguistic context within which the learners in question will come into contact with the stimuli, as this will also affect responses. Due to the difficulty of identifying the 'unhelpful' stimuli it is recommended that prompt lists be piloted and also made as large as time allows, diluting the effect of any strong associates that do slip through.

4.8 Conclusions

Using prompt word lists from two word frequencies bands, two profiles were created for each of the 50 students analysed in this study. The initial research question received a negative answer, students did not generate similar response profiles and the group could not therefore be considered as homogenous. Even within this group where students were similar in terms of their L1, language learning background, vocabulary size, age and general L2 ability there were a variety of profiles. With a less strictly selected group an even wider variety of profiles might be expected. This confirms the findings of Fitzpatrick (2007) that we should be wary of grouping students and that research that considers students on an individual basis is a more promising line of enquiry. The second research question also received a negative answer, learner response preferences did not change when lower frequency words were used as stimuli. Most (90%) of the individual profiles made from very high frequency words were classified as having either a *very close* or *close* proximity to the profiles made for the less frequent words. These proposed definitions of profile proximities are based on the correlation coefficient between individuals' profiles.

In a further analysis of student's dominant response preferences it was found that when students' top two response characteristics were combined 78% of the students continued to respond using the same combination of responses in the PWL2 as they did in the PWL1. Another point to come out of this study is that Fitzpatrick's classification offers opportunities to look at the profiles in finer detail than the traditional broad categories that tend to obscure a lot of useful information.

Despite these positive findings a question still remains over the frequency bands used in this study, they appear to be too close to allow strong claims to be made on the effect of frequency on responses. Other concerns have also been raised about the possible effect of stimulus word class, and also how to efficiently select prompt words that will yield useful data. These issues all need to be addressed but in line with the argument put forward in Chapter 2, not all at the same time. In such a complex field of research it is better to inch forward, rather than attempt to take huge leaps, each of these areas will therefore be addressed in separate experiments. The next experiment in the series, Chapter 5, will therefore revisit the effect of stimulus frequency.

Chapter Five: Revisiting the effect of word frequency

5.1 Introduction and overview of the study

One of the main findings of the Chapter 4 experiment suggest that individuals' response profiles are not influenced by the frequency of the prompt word. Response profiles to words from the 0 - 500 frequency band correlated strongly to profiles generated from lower frequency words (500 - 1000 frequency band). However, in the discussion of that study, attention was drawn to the closeness of these two frequency bands. It may well be the case that the high correlations between the students' profiles are in part due to the frequency bands being too close to show any difference. Both the 0 - 500 band and the 500 - 1000 band could be lumped together as *high frequency* words. All the stimulus words used in the Chapter 4 experiment (from now on referred to as the *Noun 1* study) might therefore be viewed as 'core items' within students' mental lexicons and so be similar to learners in terms of familiarity. In order to check if the results were indeed due to both prompt word lists being too similar in terms of frequency, it was decided to re-run the experiment using two sets of prompt word lists taken from a more diverse set of frequency bands. In the experiment reported in this chapter (from now on referred to as the *Noun 2* study) the first list contained high frequency stimulus items that were thought to be well-known to students; the second list of stimulus words were of a lower frequency. It was assumed that the second set of words would have been acquired more recently and would therefore be less well-known, peripheral items. The words in the first prompt word list (PWL1) for the Noun 2 study were selected from the most frequent 500 words in the BNC, the second list of prompt words (PWL2) were selected from the 1500 - 2000 frequency band. If frequency does have an effect on the profiles then we would not expect the kind of high correlations between the individuals' profiles that were found in the Noun 1 study.

5.2 Research questions

As the focus of the Noun 2 study is to test the finding from the Noun 1 study, that stimulus frequency has little effect on responses, the main research question is essentially the same. Another general goal within the research project as a whole is to confirm the reliability and, where possible, refine the methodology used. One part of the methodology that was specifically assessed during the data collection phase was the accuracy of interviewer intuitions in correctly categorising ambiguous responses. The second research question recognises that a certain amount of subjectivity is unavoidable in categorising WA

responses and that researchers ought to be aware of the limitations of rater intuition. The two research questions are:

1. Does the frequency of the stimulus word affect a learner's characteristic response pattern?
2. How accurate is rater intuition?

5.3 Participants

Although both the Noun 1 and Noun 2 studies follow the methodology outlined in Fitzpatrick (2007) a key difference between the two studies is the use of lower frequency words in the Noun 2 study. The Noun 1 cohort was of an elementary level with average TOEIC scores of 301.8, achieving an average of 78% on the first level of Nation's VLT (1990). As in the Noun 2 study it was necessary to use a group that had a good productive knowledge of the most frequent 2000 words, a group with a higher ability was needed. Finding a large pool of students of this calibre was consequently more difficult, necessitating the cooperation of another university in the area. Within the class that I was fortunate enough to be given access to, there were two kinds of student:

- Japanese students who had studied abroad (21)
- Foreign students from other Asian countries (9)

This group were of an upper intermediate level with TOEIC scores averaging 802 and scoring an average of 95% on the 2nd level of the VLT. As well as higher English proficiency, participants came from more diverse backgrounds than the Noun 1 study. The 30 learners included both undergraduate and postgraduate students from; Japan (21), Indonesia (4), China (3), Thailand (1), and Cambodia (1). This group was visited on three consecutive weeks in July 2009, during their regular class time, to collect the data. In the first session the whole group took the VLT and then ten students did the WA test and interview; in each of the subsequent sessions ten more students were tested and interviewed.

5.4 Materials

5.4.1 Stimulus word lists

Two word lists were derived from the British National Corpus (BNC). The first list was drawn from the 0 – 500 frequency band, this was the same list used in the Noun 1 experiment. One reason for this was that the results would be more comparable with the Noun 1 experiment, another reason was piloting of the list would be unnecessary. The

second list was selected from the 1500 – 2000 frequency band; these words were trialled with five slightly lower ability (than the main test group) students to help weed out unsuitable stimulus words. Unsuitable words were:

- Strongly associated to just one other word, identified using the Edinburgh Associative Thesaurus (EAT).
- Common collocates of Japanese words.
- Difficult to classify due to belonging to more than one word class.
- Too difficult for respondents.

This left two prompt word lists (PWL1 and PWL2) of 48 nouns each (Appendices 4.1 & 5.1). As students' average score was 95% on the vocabulary test, the prompt word lists were judged as being within the ability of these learners. The prompt word lists were given to the 30 learners and they were instructed to write the first English word that they thought of when they read the prompt word. Instructions were given (written and verbally) in English and Japanese.

5.4.2 Classification

As with the Noun 1 experiment, Fitzpatrick's 2007 classification system was used (Appendix 4.3). Immediately after the word association test, learners were given partial retrospective interviews to help with classification. Learners were only asked about items that on a cursory inspection seemed ambiguous and would therefore be difficult to classify. The interviews were conducted during regular class time on a one-to-one basis; interviews were slotted in between regular class activities run by their teacher. Prior to asking learners about these ambiguous items the interviewer made a guess at how these responses ought to be classified. A note was kept on whether these guesses were correct or not.

5.5 Results

In this section the following is reported:

5.5.1 Completion rates of PWL1 and PWL2.

5.5.2 General trends in the group.

5.5.3 Focusing on individuals

5.5.4 Individual profiles: five case studies.

5.5.5 The accuracy of intuitions.

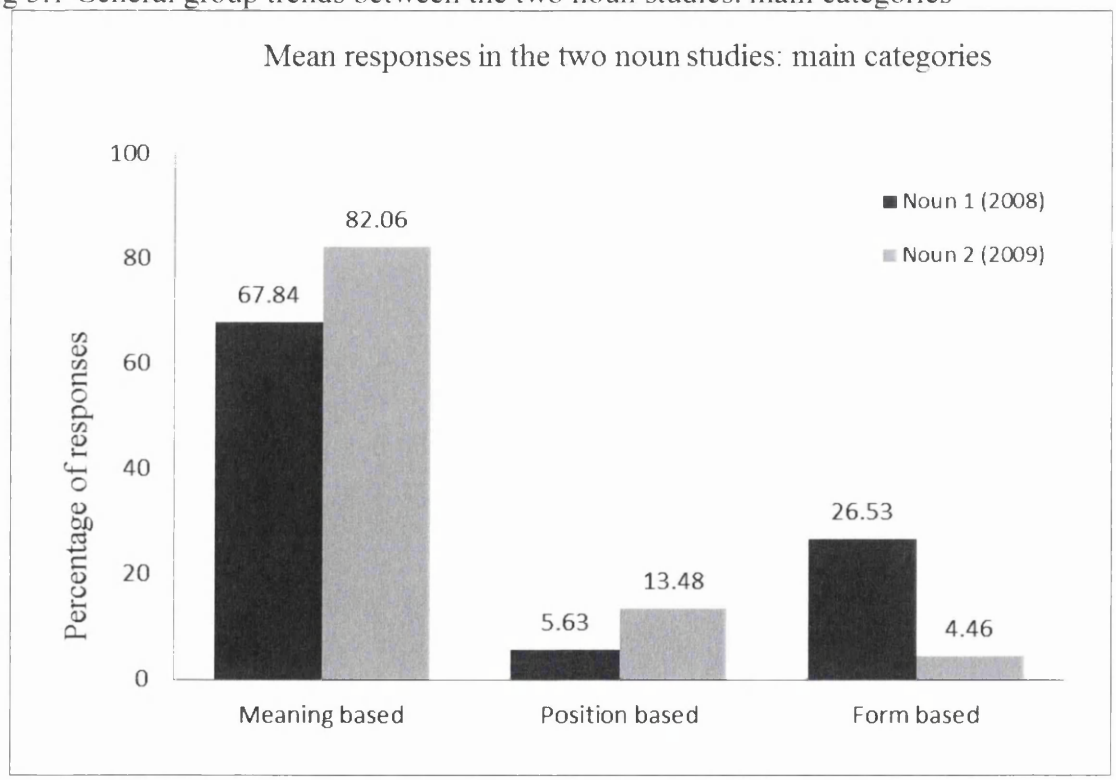
5.5.1 Completion rates of PWL1 and PWL2.

Although it was expected that some learners would not know every item, the 30 individuals all completed enough of the word lists for a satisfactory profile to be produced. As with the Noun 1 study, the threshold for completion was 50% or more for each word list, however with this higher ability group the completion rates for prompt word list 1 (PWL1) were virtually perfect and for PWL2 they were also high (averaging 82%). Most of the learner profiles created were therefore composed from responses to between 45 and 47 stimulus words.

5.5.2 General trends in the group.

In order to compare the data from this study to the Noun 1 and previous studies Fig 5.1 shows the mean percentage of responses to each of the three main classification categories. Although not synonymous it should be noted that Meaning-based is similar to the paradigmatic category used in previous studies (Söderman, 1993; Wolter, 2001), Position-based is similar to the syntagmatic category and many of the phonological or clang responses would be found in the Form-based category.

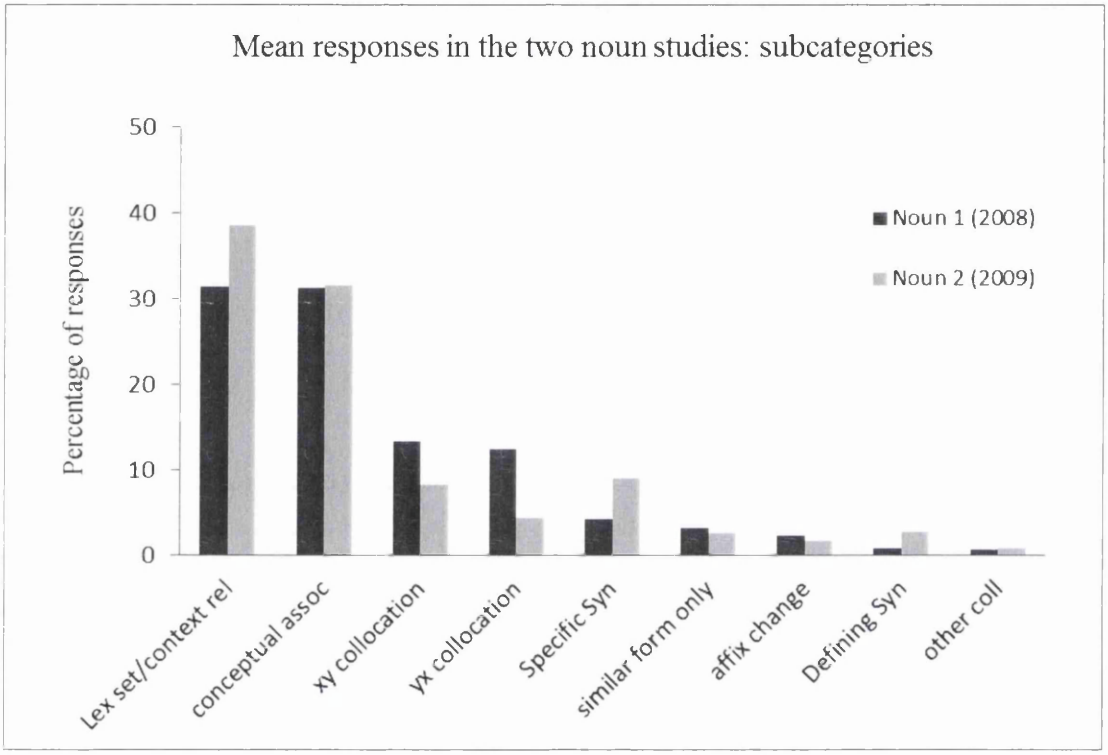
Fig 5.1 General group trends between the two noun studies: main categories



As we can see from Fig 5.1 the Meaning-based category (similar to *paradigmatic* in earlier studies) dominates the responses for both sets of profiles. The general findings in the Noun

2 study are similar to the Noun 1 study. Two areas of difference might be noted, the increase in the percentage of Meaning-based responses in the Noun 2 study and also the drop in Form-based responses. These are probably due to the increased ability level of the Noun 2 group. As argued in the Noun 1 study, it is more useful to view the data from the subcategory perspective (Fig 5.2).

Fig 5.2 General group trends between the two noun studies: subcategories



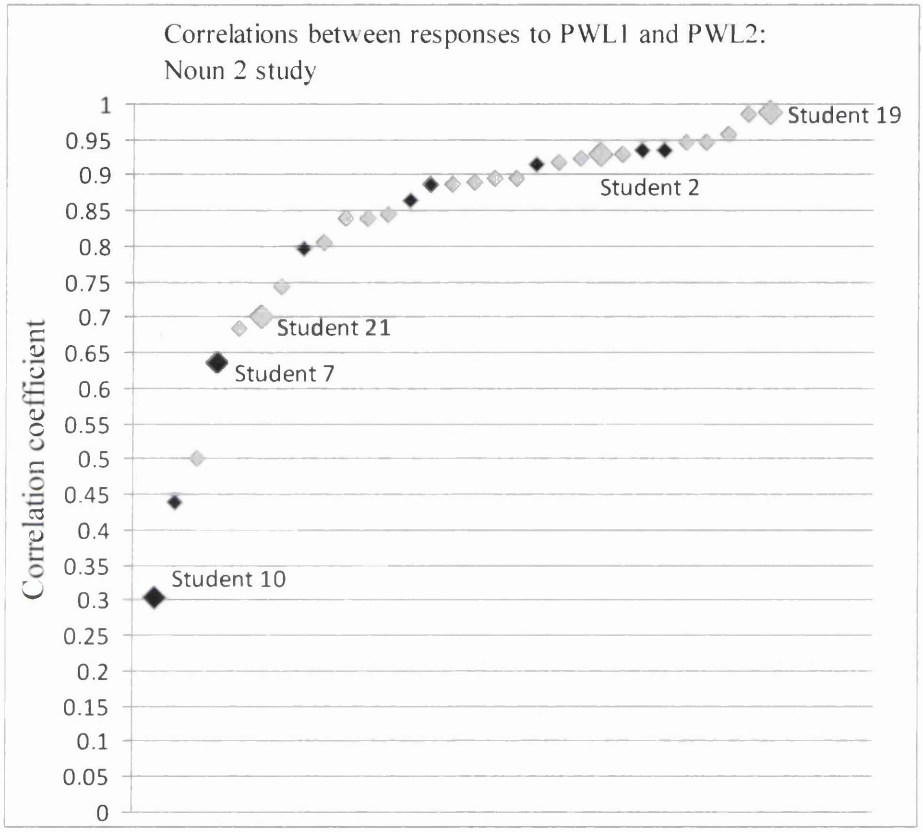
If we look at the more detailed picture (Fig 5.2) we can see that within the Meaning-based categories it is the *same lexical set/context relationship* and *conceptual association* subcategories that dominate. It can also be seen that both the low-ability learners (Noun 1 study) and the upper-intermediate learners (Noun 2 study) responded with very few *defining synonyms*. Although interesting, due to the problems raised in Chapter two with previous (conflicting) studies, the main focus of this study is not with group data but in analysing the data in terms of individuals.

5.5.3 Focusing on individuals

When each individuals' Profile 1 (responses to very high frequency words) is compared with the profile created from their responses to less frequent words (Profile 2) the two profiles were generally found to be similar. The proximity of these profiles was confirmed

through calculating Pearson's correlation coefficient from the percentage of responses in each subcategory. As can be seen in Fig 5.3, most of the learners' correlations were strong, although there were three learners whose profiles were below the 0.6 correlation threshold. The vast majority of individual profile pairs had a strong relationship.

Fig 5.3 Correlations between profiles for 30 students in the Noun 2 study



Note: gray markers indicate Japanese learners,
black markers indicate non-Japanese learners,
larger markers indicate case studies,

In Fig 5.3 students 2, 7, 10, 19 and 21 are indicated with larger markers; these five learners represent the range of correlations that were observed with this cohort. From Student 10 whose profiles had very little relationship, through to Student 19 whose profiles were virtually identical for the two frequency bands. These particular students will be discussed as case studies. As Table 5.1 shows, only a few learners' profiles were below the 0.6 threshold, most (83%) were defined as being *close* or *very close* (r=>0.7). This is similar to findings in the Noun 1 study (90% of learners had *close* or *very close* profiles). As with the Noun 1 study, in order to get a sense of the amount of variation between the learners, nine profiles were randomly selected and each of these profiles compared to the other eight using a series of chi-square tests. As shown in Appendix 5.2, of the 36 possible

pairs 33 were found to be statistically different (p=<0.05). This adds further weight to the claim made in the previous chapter that learner profiles are inhomogeneous.

Table 5.1 Proximity rankings in the Noun 1 and Noun 2 studies.

Correlation	Definition of profile proximity	**Noun 2 study (2009)**	Noun 1 study (2008)
> 0.8**	Very close	**22**	36
0.7 – 0.8**	Close	**3**	9
0.6 – 0.7*	Vaguely similar	**2**	4
<0.6	Dissimilar	**3**	1

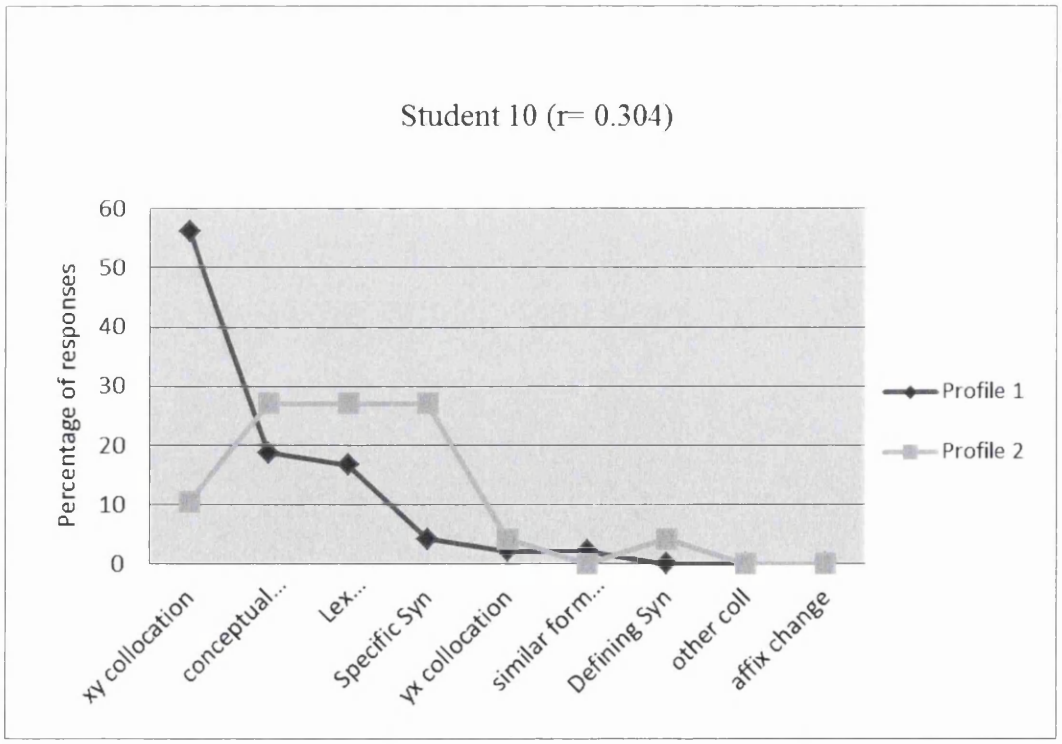
**p = <0.001, *p = <0.05

5.5.4 Individual profiles: five case studies.

In order to understand what these individual profiles look like, the five students indicated in Table 5.1 will be explored in more detail. The case studies were chosen to exemplify the definitions of profile proximity established in the Noun 1 study (Table 5.1).

At one extreme, Student 10 is an example (Fig 5.4) of a learner who gave two sets of responses that were hardly related at all.

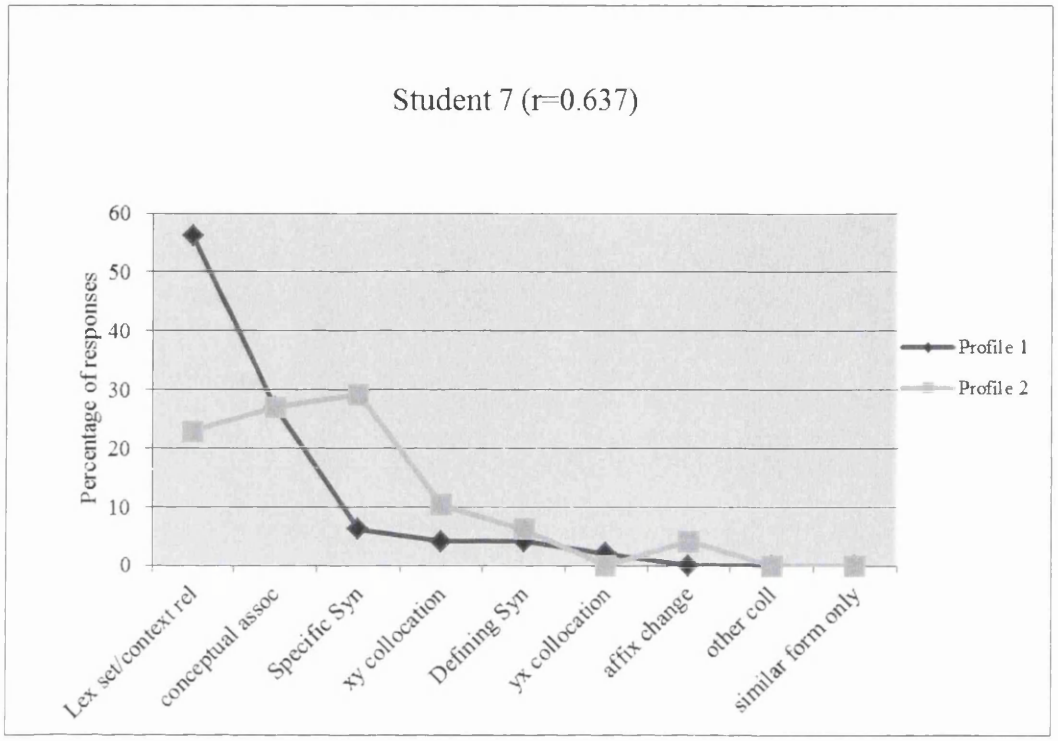
Fig 5.4 *Dissimilar* student profiles



The most striking difference is in the *xy collocation* responses; in Profile 1 it is the top ranking subcategory (56.3%) but then drops to fourth in Profile 2 (10.4%). With a

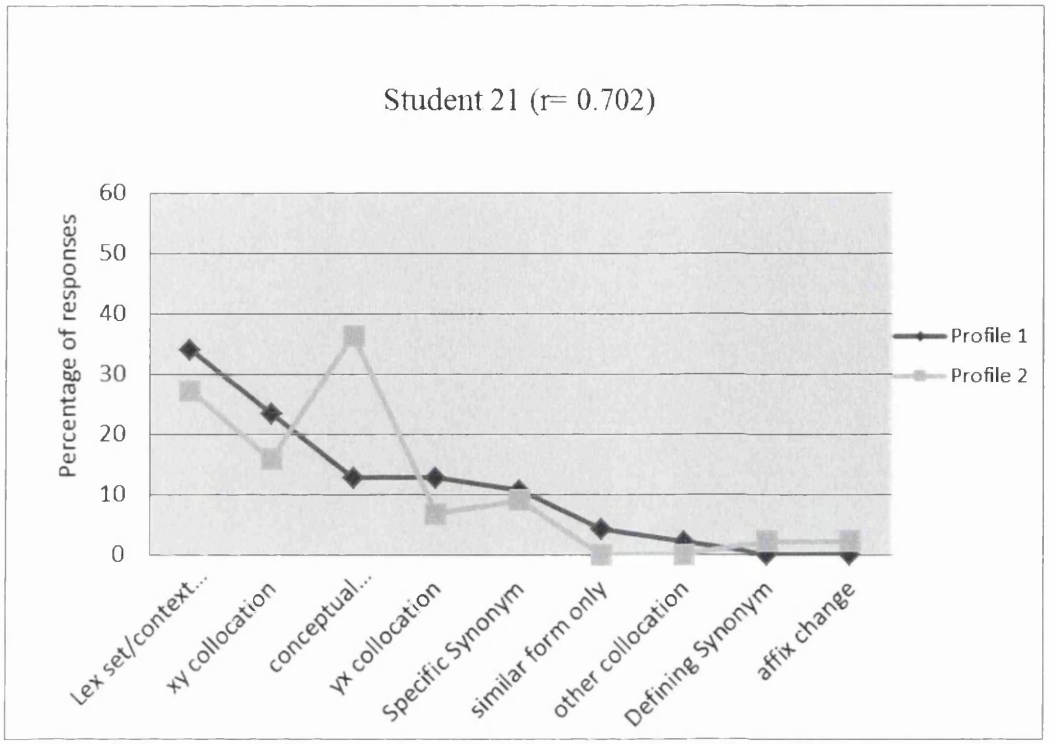
correlation of 0.304 we can classify the profiles generated from this student's responses as *dissimilar*.

Fig 5.5 *Vaguely similar* student profiles



In Fig 5.5 we have an example of two profiles that with a collocation of 0.637 can be classified as *vaguely similar*. In both profiles this student gave a significant proportion of *lexical set/context relationship* responses and also a lot of *conceptual association* responses. There is however wide variation in the responses; the dominance of *lexical set/context relationship* responses in profile 1 (56%) for example drops significantly (to 23%) in profile 2. The percentage of *conceptual association* responses though (around 27%) remain constant and are ranked second in both profiles.

Fig 5.6 *Close* student profiles



In Fig 5.6 we have two profiles that show a *close* relationship, although with a 0.702 correlation this is a borderline case. Apart from one subcategory (*conceptual associations*) the two profiles show little variation.

In the next example (Figs 5.7) we see two profiles that are far more similar than in the previous case studies. In both profiles 40% of responses are *lexical set/contextual relationship* with *conceptual association* and *yx collocations* both ranking 2nd and 3rd in the two profiles. The high correlation coefficient (0.930) confirms that these two profiles are *very close*. These *very close* profiles were typical for most of the students in this study (73%).

Fig 5.7 *Very close* student profiles - a

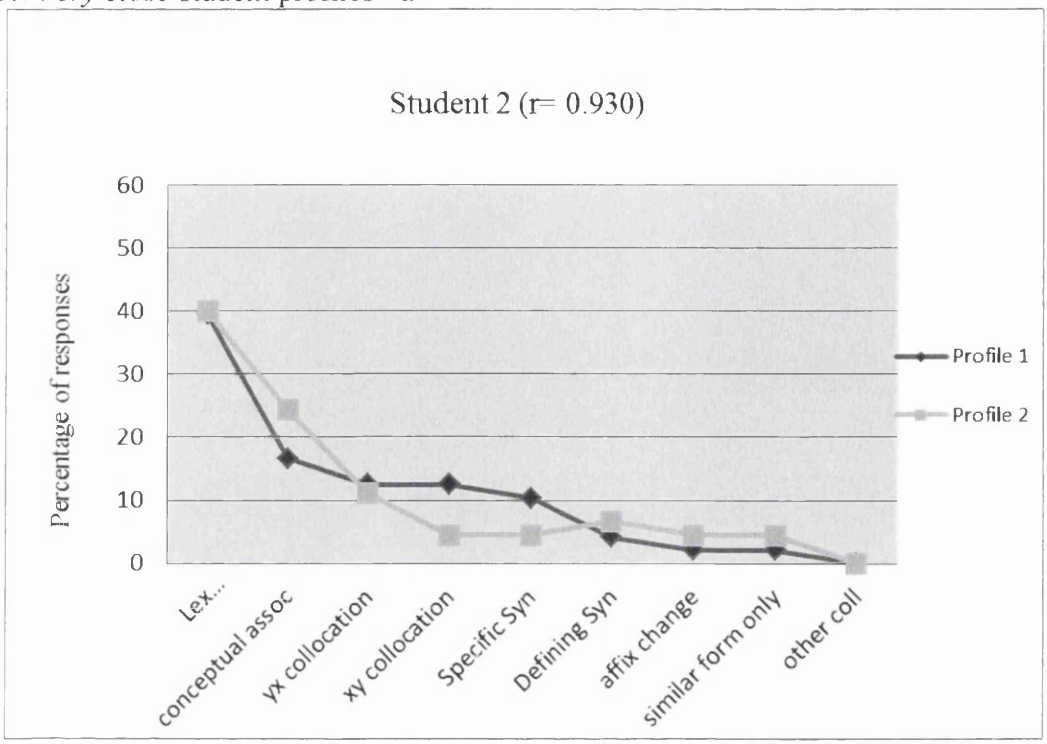
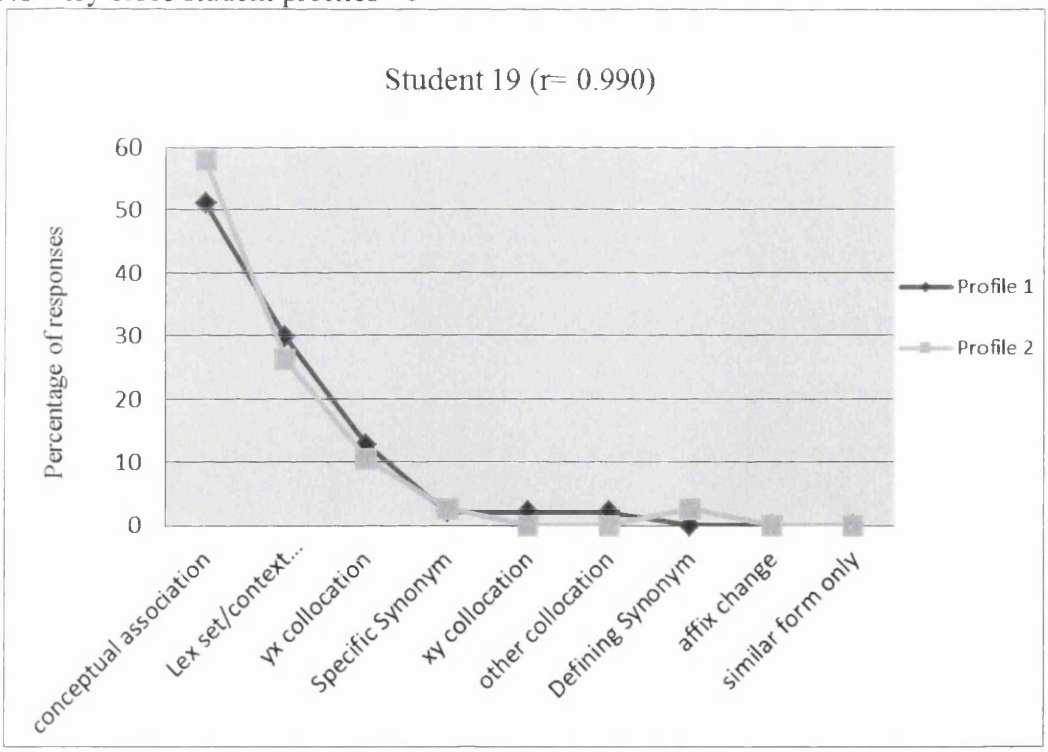


Fig 5.8 Very close student profiles - b



The final case study (Fig 5.8) is another example of two profiles that, with a correlation of 0.99, are defined as *very close*. In both profiles the dominant subcategory (*conceptual associations*) is over 50%, also, within the 2nd and 3rd ranking subcategories there is hardly any variation in the proportion of responses given. This student had the closest profiles in

this particular cohort. It ought to be noted that while the shape of the profiles in Fig 5.7 looks similar to the shape of the profiles in Fig 5.8, as the order of the response type categories (along the x axis) is not the same these two student's profiles are actually quite different.

The main finding is that individuals usually respond to stimulus words in a reliable way. Even when stimuli were selected from more diverse frequency bands than in the Noun 1 study, a frequency effect was not observed. The profile of responses to PWL1 stimuli correlated strongly with the profile of responses to PWL2 stimuli for most students. As with the Noun 1 study the answer to the main research question is negative. We can therefore conclude:

A learner's response profile, generated from high frequency words is usually strongly related to a response profile generated by the same learner to lower frequency words.

Neither the findings of the Noun 1 nor the Noun 2 studies show a frequency effect for individual profiles.

5.5.5 The accuracy of intuitions

To answer the second research question, immediately after learners had completed the two word association tests their papers were checked and a note made of those responses which were potentially problematic; i.e., might belong to more than one subcategory. Following this the interviewer made a guess as to what the learner was probably thinking for these ambiguous responses (20% of all responses were ambiguous) and then asked the learner directly why they made these responses. Typically the interviewer could narrow it down to one of two possibilities, an example from PWL2 was the response *coral* to the prompt word *coal*. If the student had been thinking along the lines of *both are a kind of rock* then it could be classified as Meaning-based. It is also possible that the student was linking them by form. Both these words begin and end with the same letter, so look alike, and both have a similar initial sound, /kɔ/. When asked in the interview phase it became apparent that the learner was associating the two words through the form of the word. As it turned out, the interviewer had in this instance guessed correctly.

As most of the ambiguous responses could be classified as belonging to one of two categories it was not surprising to find that close to half (43.6%) were guessed correctly. Ambiguous responses comprised 20% of all the responses, if the interviews had not been done then there would have been an error rate of around 11%. In this study that would have meant five or six responses per prompt list would have been erroneously categorised. As

this study did verbally confirm all of the ambiguous responses it is argued that in this study the error rate was negligible.

5.6 Discussion

The results of this study confirmed the findings of the Noun 1 study, the characteristics of an individual's profile do not change when learners are prompted with words from different frequency bands. Given that the frequency of the stimulus words does not seem to influence responses we can now turn our attention to other potentially confounding variables. The most likely of these is the word class of the stimulus. In the studies so far all the stimulus words have been nouns, it may well be the case that stimulus words drawn from other word classes behave less consistently. A further area for discussion is how to collect the data as accurately and efficiently as possible, particularly: dealing with ambiguous responses. Finally in this section, due to the learners within this study coming from various countries the effect of learner background on response profiles will also be commented on.

5.6.1 Word class

Other than frequency, the next variable that needs accounting for is the word class of the stimuli. In this, and the previous study the stimulus words were all nouns. As Bagger-Nissen & Henriksen (2006) claim "nouns elicit a higher proportion of paradigmatic responses than verbs and adjectives". This phenomenon is also documented in L1 studies such as Deese (1965) and Entwisle (1966). It is therefore not unlikely that the overall high percentage of paradigmatic/meaning based responses in the two studies (see Fig 5.1) is in part due to the use of nouns as prompt words. There are good reasons to expect different response behaviours from other word classes. One reason is that word classes differ in terms of their size, which means the number of potential same class responses will vary with each word class. Another is that they differ in how they relate to other words (within the same class and to words in other classes) and also how they function in the language. Given these fundamental differences we might therefore expect different response patterns to emerge. The question is:

> *Would the response patterns still show individual consistency if prompt words from word classes other than nouns were used?*

The next logical step is therefore to repeat this experiment with prompt words from a different word class. After nouns, verbs are the most numerous word class, so seem a good

choice. Very small word classes (i.e. pronouns, prepositions, adverbs) might be problematic to research simply due to their limited numbers, which would not allow much of a 'selection'.

As well as repeating the experiment with another word class, another option would be to construct a stimulus list with items from multiple word classes. While possible, there are benefits to looking at just one word class at a time. The first benefit is that a larger number of items from each word class can be explored. As noted in Chapter 2, Bagger-Nissen and Henriksen's 2006 study used stimuli from three word classes although this meant only 15 items per word classes could be included. Due to the problems inherent in classifying this kind of data, and the difficulty in identifying all the unproductive stimulus words, it is preferable to test a large number of items. Responses to just 15 words does not seem enough to make any confident claims about the effect of a particular word class on responses. The second benefit is that through keeping the word classes separate it is easier to interpret the results. As already stated, the word classes are fundamentally different, this may mean the classification system and methodology might also need to be slightly adapted to cope with these differences. For example, the selection criteria for including/excluding words in the prompt word lists would become increasingly complex with each extra word class used. As well as trying to filter out the 'unhelpful' words listed in section 5.4.1, a multi word class stimulus list would need to be checked for any common combinations of the nouns, verbs or adjectives included. If a noun such as *rain* were in the prompt list this would preclude the inclusion of adjectives such as *heavy*, *hard*, *incessant* and perhaps the verb *fall* that would probably prime the item and lead to a collocation response. Such 'primed' responses would not tell us much about the response characteristics of an individual.

5.6.2 Improving the methodology

The second issue that I wish to discuss is how best to balance the practical constraints of time with the collection of accurate data. This is a methodological problem of how to efficiently collect data which is good enough to generate a profile that can reliably show an individual's characteristic responses. Basic questions that need to be considered are:

How many items should we use in a prompt word list?

Is a follow up interview (or partial interview) necessary?

In the studies we have mentioned so far there have been a wide range of prompt word list sizes; Fitzpatrick used 100 items in her studies, Wolter (2001) 50, Bagger-Nissen and

Henriksen (2006) 45. In her 2006 study Fitzpatrick used interviews to help with the classification, Wolter (2001) also used a full interview in which learners were asked about every item. The interview process is time consuming, meaning that in Fitzpatrick's 2007 and 2009 studies it was dropped. In Wolter's 2001 study it should also be noted that there were only a small number of participants, presumably due to time constraints. The issue is that if we are trying to obtain accurate data we need to verify the classification (interviews are a good way to do this), however, if the list of items is long then doing this on a large scale will take too much time. In the Noun 1 study a solution was developed which offered a compromise between time expended and the quality of the data. The solution was a partial retrospective interview: learners were all interviewed but only on ambiguous items and items that were illegible. This turned out to work well as it was possible to test and interview groups of about ten students at a time within a 90-minute class.

In the Noun 2 study this partial interview method was again adopted and in addition a note was made concerning the ambiguous items. The error rate of the author, the sole rater in this study, was calculated at 11%. Of the nine potential categories (excluding the *error* category), the rater could usually narrow it down to one of two possibilities; consequently the thinking behind about half of the ambiguous responses were correctly guessed. One might argue that in a study using a 100 item prompt word list (Fitzpatrick, 2007) an error rate such as this is acceptable, and the time consuming interview can be dispensed with. Even though more time will be needed for participants to complete a 100 item test, this is more than made up for in time saved in not interviewing and generally simplifying the process. There seem to be two basic options:

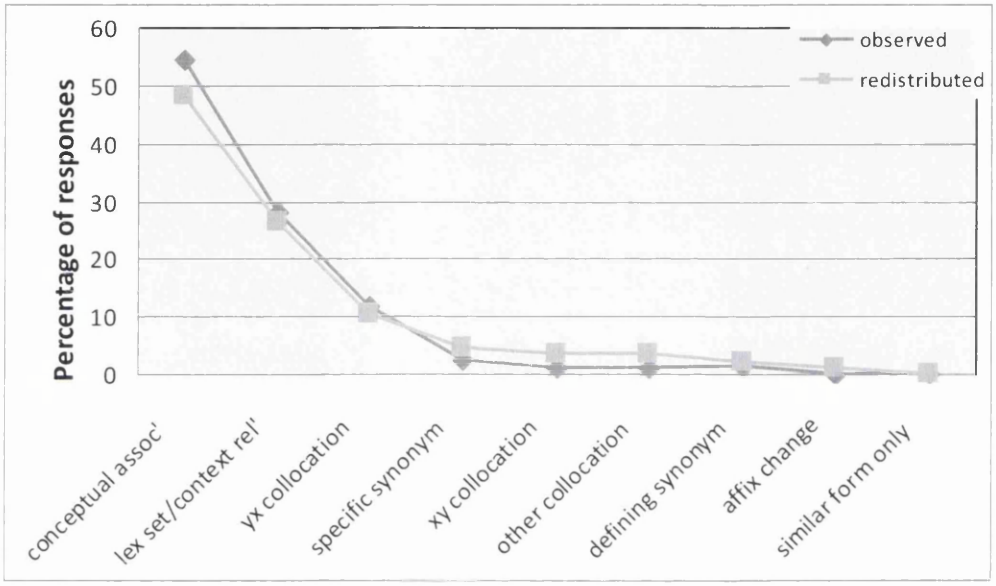
- A word association test with a large number of items (80 -100) without an interview that has an error rate of around 11%.
- A word association test with a smaller number of items (40 - 50) with a partial interview that has a negligible error rate.

For raters with considerable experience in classifying responses (as in this and Fitzpatrick's study) there is not much in it. Both options seem to offer a method for generating data from which a reliable profile can be made, and both put a similar burden on the participants. The intuitions of less experienced raters would probably be less accurate; in such circumstances it would be wiser to incorporate a follow-up interview to limit misclassification.

To see what kind of effect an 11% error rate would cause, the responses from the student with virtually identical profiles (Student 19, Fig 5.8) were re-examined. The first

step in simulating the effect of such a rate of error was to remove 11 % from each of the nine subcategories. The second step was to randomly redistribute these responses to the nine subcategories. As there were 85 valid responses given by this student 11% amounted to nine responses. The original 'observed' response profile and the profile created after randomly redistributing the responses can be seen in Fig 5.9.

Fig 5.9 The effect of randomly redistributing 11% of responses



As can be seen in Fig 5.9 the redistribution did not affect the shape of the profile. The initial three subgroups in both profiles (accounting for 94% of the observed responses and 85% of the redistributed responses) are ranked in the same order. This appears to justify Fitzpatrick's rejection of the interview procedure in her later studies.

5.6.3 Learner background effect

While the effect of learners' backgrounds were not specifically targeted in this study, as the learners' countries of origin were more diverse than in the Noun 1 study it is of interest and something that ought to be commented on. The individual profiling approach would predict that the non-Japanese learners (9) would generate profiles that have similar levels of internal reliability as the Japanese learners (21). This was generally found to be the case; as can be seen in Table 5.1, non-Japanese learners had correlations that represented the full range of the group. Also, as has already been noted, only a handful of the whole group generated profiles that were defined as *dissimilar*. This study therefore does not really support the idea that learner background has an effect on responses. That said though, of

the four Indonesian students, two had the weakest correlations in the whole group (r=0.304, r=0.440) while the other two didn't have particularly strong correlations (r=0.637, r=0.797). Although only a tiny sample, Indonesian students appear to generate slightly less reliable responses than students from other countries. Why students from this country would generate less reliable profiles is unclear as not enough background information was collected for each student. However, differences in the how English is taught in schools, the way vocabulary is presented in local textbooks and the availability of English in the media are all potential factors. Indonesia is also marked in that it doesn't have just one dominant language but a mix of competing 'native' languages, (Malay, Javanese, Sundanese) meaning that many Indonesians are multilingual before they embark on their English studies. I think however that the most likely reason for the increased variation in the Indonesian responses is that the word lists were piloted with Japanese learners and so only checked for possible Japanese/English cognates. If one of the lists contained a number of words that Indonesians have borrowed into their own language(s) it could explain the variation in response patterns. A criticism of this study is that potential prompt words were not sufficiently trialled with learners from all the countries represented in the main experiment. Despite the lower levels of reliability with profiles generated by Indonesian learners, most students generated profiles that were internally reliable when subjected to a split-half analysis.

While the data for the Noun 2 study was being collected, an experiment along parallel research lines was also underway - a further study by Fitzpatrick (2009). In that study she found individual response profiles of Welsh bilinguals to word association tests in their first language to be similar to response profiles in their second language. While her study adds support to the claim that student background has little effect, it ought to be noted that Welsh is not a particularly good language to use for research into L1 and L2 language behaviour. As Fitzpatrick puts it, "the Welsh language has an interesting status in Wales, which sometimes makes it difficult to categorise as a first or second language for an individual speaker" (2009:45). Welsh is a minority language and so has unusual usage patterns, even within the L1 community; some people use it at home but not at work; some use it with friends but not in an academic context; others learned it as children but have rarely used it since. Given the uniqueness of the Welsh language, it is questionable whether we can extend the findings of Fitzpatrick's study to L2 learners in other contexts. It would seem more useful to explore the responses of learners that are representative of language learners in general, and see if the approach she advocates produces complementary

findings. The Japanese, Cambodian, Chinese and Thai students within the Noun 1 and Noun 2 experiments did have clearly definable L1s and L2s and therefore lead us to tentatively conclude that learner background does not appear to have much of an effect on responses. Such an effect cannot however be ruled out by this study alone as the responses by Indonesian students hinted that learners from some countries may give less reliable responses.

5.7 Conclusions

The most important point to come out of the Noun 2 study is the confirmation that the frequency of the prompt word used in a word association test does not have an effect on the individual's response characteristic. The results of this study support research by Fitzpatrick (2007, 2009) that found variation between individuals, even though the response behaviours of those individuals were internally consistent. The degree of proximity between each individuals' response profiles in this study were generally *very close*. Although the findings from the Noun 1 study already pointed in this direction there was a question mark hanging over these results due to the closeness of the two frequency bands used. With a greater difference between the frequency bands in the present study, this question mark has been removed. We can now turn our attention to other unanswered questions from the Noun 1 study, the possibility of a word class effect being the most salient. The following Chapter will therefore explore the responses to stimuli from another word class: verbs.

Two further points also came out of this study; the first is that student background does not seem to be a major issue, if due consideration is given to the learner's L1 when selecting prompt words. If we consider this study (5 nationalities represented) with Fitzpatrick's 2009 study of Welsh bilinguals we have no real reason to believe that background has an effect. The second point, that may help refine the methodology of future free word association studies, is that the intuitions of an experienced rater appear to be quite good when guessing at ambiguous responses. With an error rate of only 11% a good argument can be made for dispensing with the time consuming interview procedure. When simulated, such an error rate had a negligible effect on the profile generated.

Chapter Six: The effect of verb stimuli

6.1 Introduction

The results of the Noun experiments showed that individual response profiles did not change when response profiles from very high frequency prompt words were compared with profiles from lower frequency prompt words. Word frequency does not seem to affect individual profiles. Such findings encourage us to look further at the feasibility of creating individual learner profiles from word association responses and examine the robustness of the methodology outlined by Fitzpatrick (2007) from other angles. The angle that will be explored in this chapter will be the effect that word class has on responses.

In the initial Noun 1 experiment the possibility of word class having some kind of effect was raised as all the prompt words were nouns. Studies such as Deese (1965) and Bagger-Nissen & Henriksen (2006) report a word class effect in L1 and L2 word association studies. We might also note recent studies in neurology, such as Mestres-Missé et al (2010). Such studies, using MRI technology, show us that different areas of the brain become active when subjects process words from different word classes. L1 studies (Gentner, 1982) also claim that there are fundamental differences between how we conceptualise the main word classes and that they are acquired at different stages in language development. We therefore have good reason to believe that the word class of a stimulus in a word association test will have some sort of an effect on a learner's profile. In order to see if this is the case a similar experiment to the Noun experiments was set up, the main difference being that the noun prompt words were replaced with verbs.

An alternative experiment design might have been to test a variety of word classes at the same time, this is the method adopted by Bagger-Nissen & Henriksen (2006). They gave students two 45 item stimulus sets comprising of nouns, verbs and adjectives. The two sets were at different frequency levels, 90 items in total. A major drawback of this experiment was that in the analysis the 45 items in each set were divided into three smaller sets of only 15 items each. As argued in Chapter two, generalisations based on such small samples do not hold much weight. A similar (multi word class) test design which included more items was considered, however increasing the number of items per word class was not really a practical solution. There is a limit to how many items we can reasonably expect students to attend to in one sitting, beyond 100 would seem unreasonable. Were we to adopt a multi word class test design which included say three word classes and 40 items per word class at two different frequency levels then this would mean we would have to

ask students to make responses to 240 items. For such a large number of items test fatigue would become a serious issue and would probably preclude any post-test interviews. For these reasons, and others noted in Chapter 5 concerning the added complexity of such an experimental design, a multi word class test design was rejected.

It is the intention in future experiments to eventually look at all word classes but it seems prudent to deal with them one word class at a time. Which word class to investigate next (after nouns) was therefore not such a vital issue, although as verbs are the second largest word class they seemed the obvious choice. As with nouns, there is a fairly large pool of verbs to select appropriate stimulus words from within each frequency range. Another potential word class that has received attention in previous word association studies is adjectives (Lambert, 1956; Stolz & Tiffany, 1972). The problem with using adjectives as stimulus words is that they often have strong links to just one other word. As both Deese (1965) and Meara (1983) note, high frequency adjectives tend to produce their polar opposites: for example, *black →white* and *soft→hard.* Such stimulus words are unhelpful as they don't tell us anything about an individual's response preferences. Verbs therefore seem to be the least problematic of the other major word classes and a good place to start investigating whether word class has an effect on individual response profiles.

6.2 Outline of the study

The basic methodology in this study (which from here on will be referred to as the Verb study) is similar to the studies reported in Chapters 4 and 5, the difference being that verbs were used as stimulus words instead of nouns. The classification system was the same as the one used in the noun studies (following Fitzpatrick, 2007), a partial retrospective interview was also employed. The data for this study was collected on three consecutive weeks in July 2010 during a regular class. In the first session the second and third levels of Nation's Levels Test (1990) were used to confirm that students' vocabulary levels were high enough to cope with the stimulus words that would be used in the word association test. Following this confirmation that the students' range of vocabulary (28 students) was at an acceptable level, in the second session the prompt word lists (Appendices 6.1 & 6.2) were given to 14 of the 28 students and they were instructed to write the first English word that they thought of when they read the prompt word. These 14 students were then interviewed on a one-to-one basis in between working on a class project (unrelated to this study). In the third session the remaining students took the WA test and were then interviewed. Of the 28 students, 27 were eventually included in the analysis. It was

expected that the use of verbs would produce a different pattern of responses to the nouns but that an individual's response at the different frequency levels would, as in the Noun studies, correlate strongly.

6.3 Research questions

Building on what we have learned in the previous chapters about noun responses, this study asks two questions to see if the claims can also be applied to responses made, by a different group of students, to verbs. A further question (RQ3) was posed in order to help future word associations create WA stimulus lists on a more principled basis.

1. Do learners respond to verbs in a different way to nouns?
2. Do verb stimuli generate individual response profiles that correlate as strongly as noun stimuli?
3. What should a good stimulus list contain?

6.4 Participants

The group of students used in the Verb study were similar to the Noun 1 study in terms of their nationality (all Japanese), their age (young adults) and length of English study (6 - 8 years). The group in the Verb study (27) was however smaller than in the Noun 1 study (50) due to the difficulty in recruiting students with a sufficient ability to cope with the stimulus words used. As with the Noun 2 study a group with TOEIC scores of over 600 was sought, necessitating the cooperation of another university in the Hiroshima area. While the average faculty TOEIC scores that this class was drawn from "averaged just under 800" their teacher did not know the precise score for each individual in the class. The scores on the VLT (Table 6.1) test however tally with this anecdotal figure, clearly this group had a good command of the most frequent 3000 words.

Table 6.1 Mean scores on the vocabulary test

	VLT score (2000 level)	VLT score (3000 level)	Combined VLT score
Mean score (%)	**92.14**	**85.83**	**88.99**
s.d.	2.71	3.30	5.45

6.5 Stimulus word lists

In order to allow greater comparability with the Noun 2 experiment, the words in the prompt word lists were selected from the same frequency bands. The first prompt word list

(PWL1) was taken from the 0 -500 frequency band in the British National Corpus (BNC) using identical selection criteria as previous studies. The second list of prompt words (PWL2) was taken from the 1500 -2000 frequency band. The word lists were piloted and unsuitable words cut from the list leaving two prompt word lists (PWL1 and PWL2) of 50 verbs each. The piloting of the stimulus words was done by a small group (five) of Japanese college students studying at a different university to the main group, they performed similarly on the 2000 and 3000 levels of the VLT.

Unsuitable words were:

- Strongly associated to just one other word (e.g. 42% of associations to *buy* are *sell*, *give →take* 37%, *walk →run* 41%). Strong associates such as these were identified using the Edinburgh Associative Thesaurus (Kiss et al, 1973). Words with > 25% of their primary responses to just one other word were cut.
- Common collocates of Japanese words (e.g. stand/ スタンド, meaning bar/pub).
- modal verbs such as *will* (too strongly linked to other verbs)
- verbs with multiple meanings such as *draw* or *let*.
- Difficult to classify as belongs to more than one word class (e.g. *mean* or *like*)
- Too difficult for respondents (e.g. *derive*).

To ensure students were clear that in this test it was a verb to which they were meant to be responding, 'to' was put before each stimulus word (*to advise, to believe*, etc.). The word lists used in this experiment can be found in Appendix 6.1 and 6.2.

6.6 Results

In this section the following is reported:

6.6.1 Completion rates of PWL1 and PWL2

6.6.2 General trends in the group.

6.6.3 Focusing on individual profiles: four case studies

6.6.1 Completion rates of PWL1 and PWL2

After the test the responses were checked to confirm that within the group there were no prompt words that had proved too difficult for a majority of respondents or were strongly associated with just one other word. Following this, from PWL1 three words were eliminated from the analysis: *to call, to feel, to carry.* In PWL2, four words were eliminated: *to blow, to climb, to vote, to burn.* With the word *to call* for example 90%

answered *telephone/phone*. Although it was expected that some students would not know every item, 27 of the 28 individuals completed enough of the word lists for a satisfactory profile to be generated. One student did not turn over the test paper and so only completed PWL1, this was picked up at the interview stage but she couldn't be persuaded to finish it. As with the previous experiments the threshold for completion was 50% or more for each word list. As can be seen in Table 6.2, of the 27 students analysed, the number of useable responses to PWL1 averaged 95% and the number of usable responses to PWL2 averaged 87%. There were therefore over 40 useable responses per student per list available for analysis, it was assumed that this would be enough to create profiles that reliably showed characteristic response patterns for each individual.

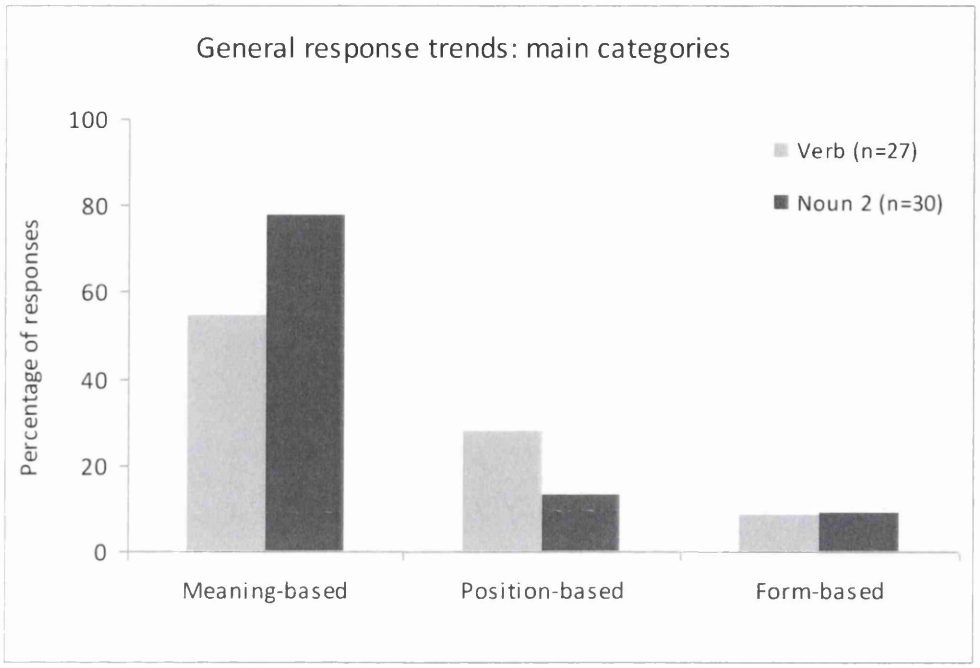
Table 6.2 Usable responses for the Verb study

n=27	PWL1 (Max 47)	PWL2 (Max 46)
Mean	**44.70**	**40.07**
s.d.	3.05	5.67

6.6.2 General response trends

Of the 2508 responses in this study, most (90%) were comprehensible responses that could be clearly classified. As can be seen in Fig 6.1, there were a large percentage of Meaning-based responses, the second largest group was Position-based with the Form-based responses being the smallest group. Bagger-Nissen & Henriksen's 2006 claim that "nouns elicit a higher proportion of paradigmatic responses than verbs and adjectives" is supported by these data. In the Verb study 55% of the responses were Meaning-based as opposed to 78% in the Noun 2 study, a considerable difference. With fewer Meaning-based responses the Verb study (Fig 6.1) generated more Position-based responses than the Noun 2 study. As well as the differences, there are however some broad similarities between the responses in the two studies; Meaning-based responses were both ranked first, Position-based responses second and Form-based responses ranked third - accounting for less than 10% of responses.

Fig 6.1. A comparison of general (main category) response trends in the Verb and Noun 2 studies

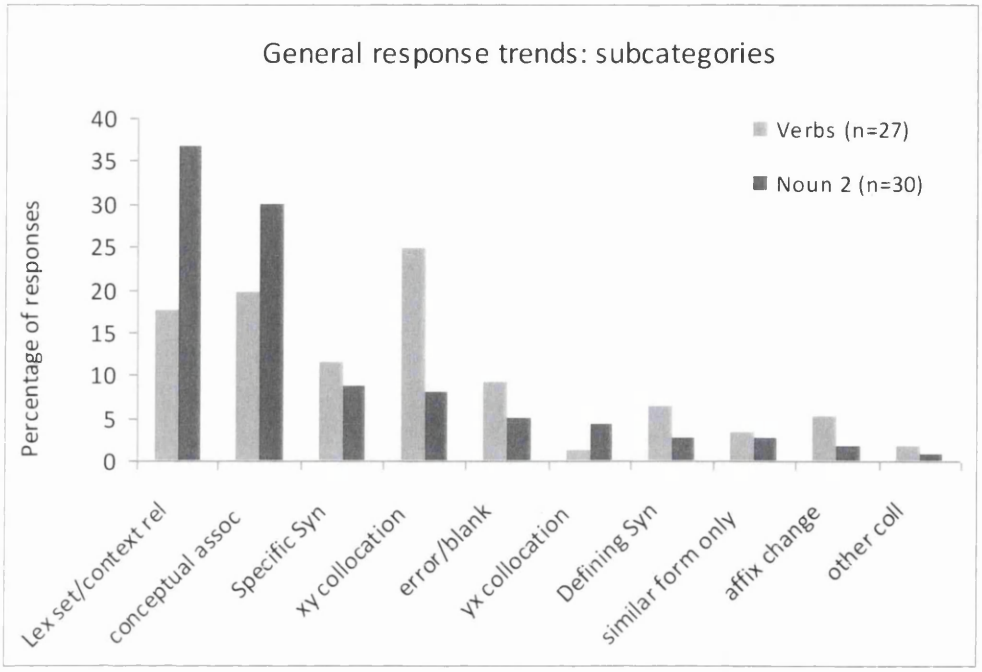


What we can see from Fig 6.2 is that when these broad categories are unpacked there is a difference in response trends. Three subcategories in particular show a marked difference:

same lexical set/context relationship	(19% difference)
xy collocation	(17% difference)
conceptual association	(11% difference)

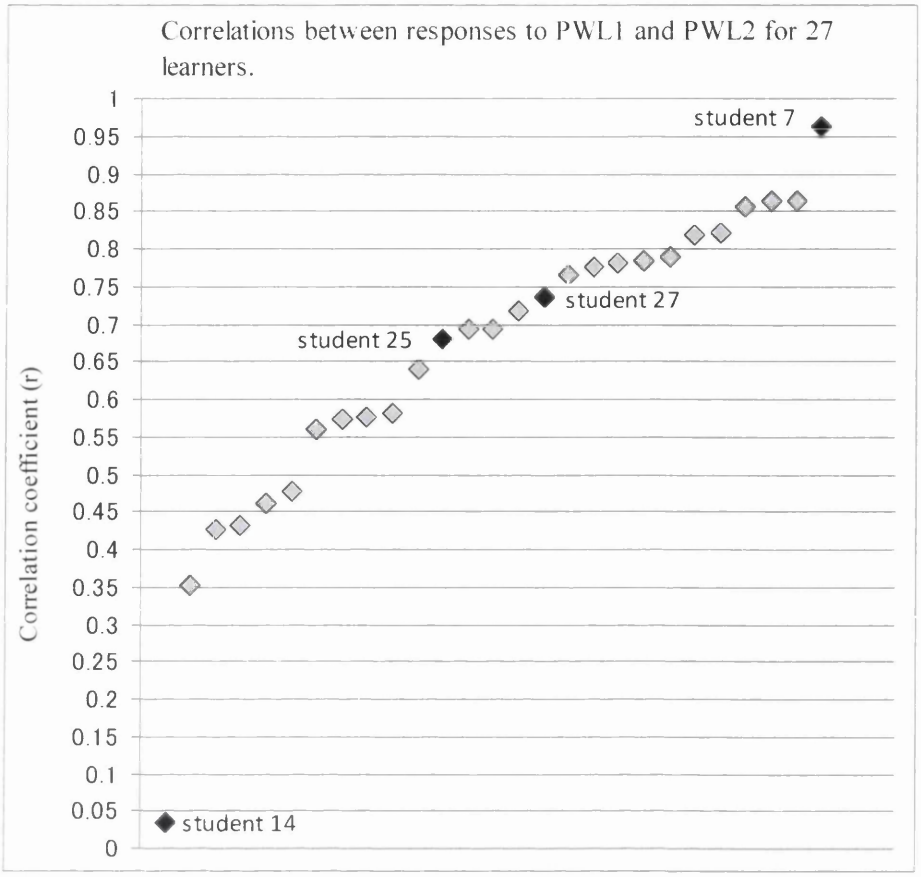
In the Verb study there are far more *xy collocations*, there is also a drop in the number of responses in the two subcategories that were dominant in the Noun studies. In answer to the initial research question *there is a difference in how learners respond to nouns and verbs.* Reasons why this might be so are taken up in the discussion section. It ought to be noted though that so far we have been comparing two different groups of students, even though they are of similar ability we might therefore expect some variation. As argued in previous chapters it is better to look at the data from an individual perspective.

Fig 6.2. A comparison of general (subcategory) response trends in the Verb and Noun 2 studies



As with the Noun studies, each individual's Profile 1 (responses to very high frequency words) was compared with the profile created from their responses to less frequent words (Profile 2). An individual's profiles were usually found to have some relationship. The proximity of these profiles was confirmed through calculating the correlation coefficient between the two profiles. To calculate this, the difference between the percentage of responses in each subcategory were compared. In the scatter-plot diagram (Fig 6.3) there is a wide range of correlation coefficients for participants in the Verb study, from students whose profiles are virtually identical (r=0.96) through to students who have very weak profile correlations (r=0.352). There was also one learner whose responses showed no relationship at all (r=0.035).

Fig 6.3 Correlations between profiles for students in the Verb study



As Fig 6.3 shows, 63% of the correlations (17 out of 27 students) were defined as being *vaguely similar*, *close* or *very close* (r=>0.6). However, a significant number of individuals (10) generated profiles that were below the 0.6 threshold and were therefore classified as *dissimilar*. The learner who had the most dissimilar profiles is considered more fully in the following section along with the three other case studies highlighted in Fig 6.3.

Table 6.3 Proximity rankings for the Verb study

Correlation	Definition of profile proximity	(n=27)
r >0.8**	Very close	6
0.7 – 0.79**	Close	7
0.6 – 0.69*	Vaguely similar	4
<0.6	Dissimilar	10

**p = <0.001, *p = <0.05

The main point to come out of these results is that when two individual profiles are generated from responses to words selected from verbs of different frequency ranges the

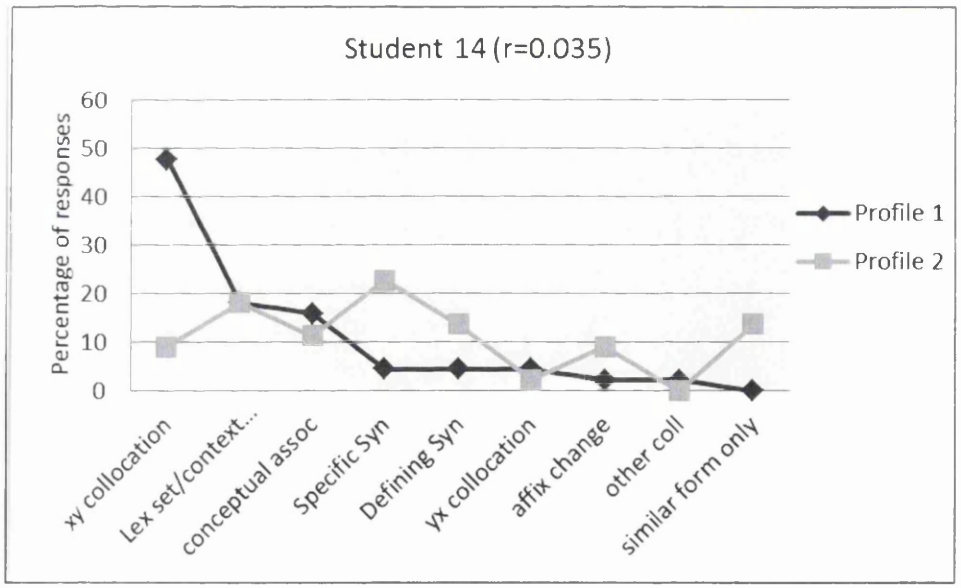
responsses oten produce a similar pattern. In the verb study, 48% of the students had correlattions hat were judged as being *close* or *very close* (r=>0.7*)* with 63% being above the 0.6 thresiold. This was however far less than the Noun 2 study (83% had *close* or *very close* correlaions). Therefore in answer to the second research question: *verb stimuli do not generateprofiles that correlate as strongly as nouns.*

6.6.3 Focusiig on individual profiles: four case studies

In order to understand what these individual profiles look like, four examples will be explored in nore detail. These students were chosen because they represent the four levels of proxiimitydefined in Table 6.3 from the student with the weakest correlation between profiles to the student with the strongest correlations.

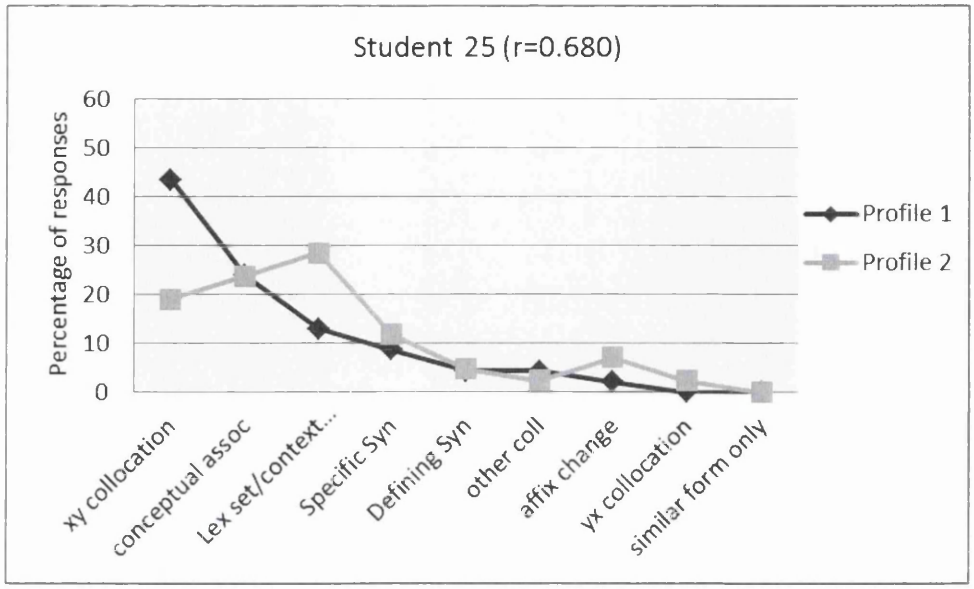
The frst case study (Fig 6.4) is an example of a student whose two profiles show no relatiionship. The dominant *xy collocation* category in profile 1 (47% of responses) drops to 9% n profile 2. As can be seen in Fig 6.3 this was the only student who gave such disparate responses, as such he could be dismissed as an outlier.

Fig 6.4 Dissimilar profiles.



The second case study (Fig 6.5), is however more typical of the students in the Verb study, here we have two profiles which, other than the *xy collocation* category, shows many of the categories match quite well. The second ranking *conceptual association* category for example has around 23% of responses in each profile with a similar percentage of *definitive* and *specific synonyms* generated.

Fig 6.5 Vaguely similar profiles



The next two profiles (Figs 6.6 & 6.7) show profiles that are typical for the top half of the learner profiles (48% of individuals had profiles with correlations > 0.7). In the case of Fig 6.6 we have two profiles that are quite close for most of the categories although vary in one category, again this is the *xy collocation* category. With student 7, in Fig 6.7 we have two profiles that match particularly well for the top two ranking subcategories. The correlation of 0.963 indicates a strong relationship between the profiles. Other than a 10% variation in *lex set/context relationship* responses, very little variation can be observed.

Fig 6.6 Close profiles

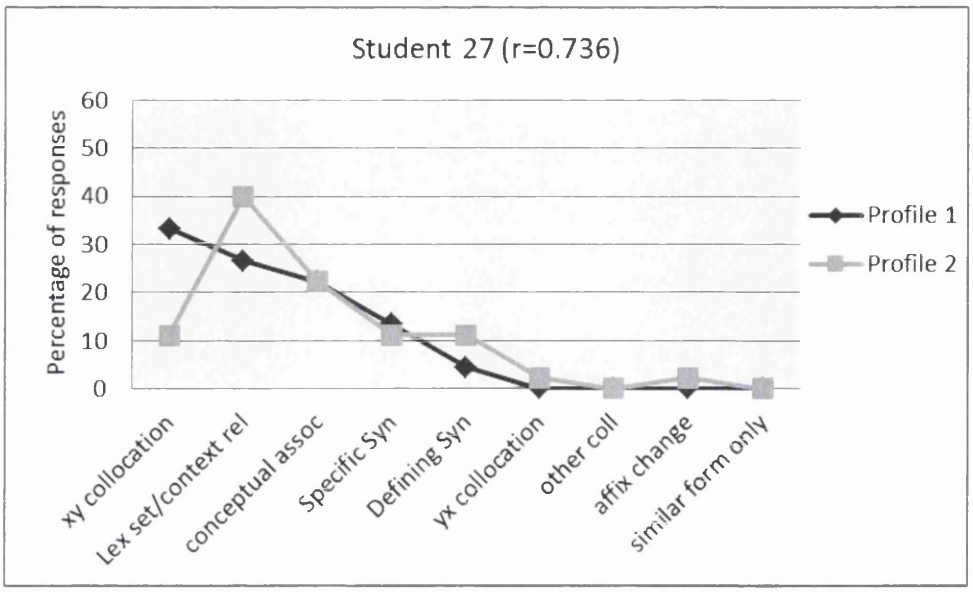
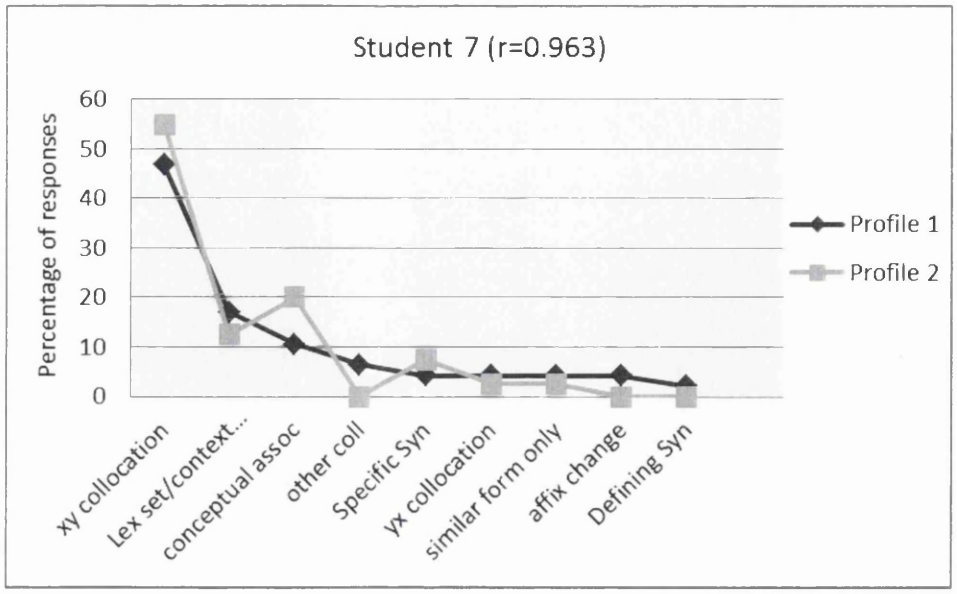


Fig 6.7 Very close profiles



A point that comes out in all the case studies (and was also evident in Fig 6.2) is that with verbs there seem to be a disproportionate number of *xy collocations*, this was a subcategory that did not stand out in either of the Noun studies. We might therefore conclude that there is something about verbs that encourages responses of this type; this point will be returned to in the discussion section.

6.7 Discussion

In this section we will discuss three main areas that come out of the results section. The first is the seemingly disproportionate number of *xy collocations*. It may well be the case that there is something inherent in verbs that makes such a response type likely. On the other hand, this may be indicative of a methodological problem. The second is the 'outlier' student who made responses profiles that were totally unrelated – how can this be accounted for? The third area of interest is in comparing the Noun and Verb studies. Although some comparisons have already been made in the results section this will be explored a little further. Finally the implications of the word class effect, that the findings seem to support, will be considered.

6.7.1 Why so many *xy collocations*?

There are perhaps two ways to look at this question, the first is to ask whether there is something about verbs in general that encourages collocations, the second is to ask if there is a problem in the methodology. As will be argued below, there seems to be a little of both.

When prior word association studies are considered, it is not so surprising that this study elicited a lot of *xy collocations*, such as *to keep→still*. This phenomenon was reported by Deese (1965:106) who found that for verb prompt words 48% of responses were syntagmatic. In Bagger-Nissen & Henriksen (2006) the verbs elicited 59.7% syntagmatic responses in the L1 and 43.6% in the L2. The fact that other studies report this though does not answer the question of why verb stimuli generate a lot of collocations. The reason for this is I think, in part, due to the nature of verbs. In a review of early first language development (six languages - Japanese and English included) Gentner (1982) presents evidence to support the claim that in their first language children learn nouns before verbs. We could argue that nouns are therefore more integrated into the lexicon at an early age. She also argues that "the kinds of things denoted by nouns are different from, and more fundamental ontologically than, the kinds of things denoted by verbs". In the "natural partition hypothesis" Gentner holds that verbs are "less transparent" than nouns and need support from other words to confirm their meaning. In an associative test with a verb stimulus, for example *hold*, it is natural to think of what people usually *hold* in order to supply the context necessary to conceptualise it. Collocational associations such as *to hold→hands* are therefore likely as 'hand' is an easy word to visualise. Synonymous associations such as *hold→carry* or *hold→contain* are less likely as these potential response words are as difficult to conceptualise as the prompt word itself. It is often a noun which provides the imageable context (such as *hand*). Unlike verbs, nouns are more "concrete" and independent; they do not usually need to refer to another word to be visualised, allowing a wider range of responses. As well as the points already made, Bagger-Nissen & Henriksen (2006:402) add that "verbs are more often polysemous and gain additional meaning from words they collocate with, so to know the verb properly the learner needs to know its collocations too".

Another way to view the verb→collocation associative phenomenon is through Wray's (2002) "needs-only analysis" hypothesis. Based on studies of L1 children, Wray argues that a lot of language is initially learned in formulaic units, these units then remain in phrasal form until it is necessary to analyse the parts. If we assume that Gentner's finding, that verbs are learned later than nouns, also applies to L2 learners we might speculate that for learners at an intermediate level more noun collocations will have been analysed than verb collocations. For such learners many of the verb components would not be as integrated as their noun counterparts or have necessitated unpacking of the formulaic unit. Being more recently acquired, verbs are more likely than nouns to remain stored

within the phrases that they were initially learned, they are therefore more likely to generate collocations.

While it seems the nature of verbs has some effect on responses, it should be noted that in this study it is only the *xy collocation* category that shows a large increase with verb stimuli. If the above arguments are correct I would also expect collocations in the opposite direction (*yx collocations*) to increase with verb stimuli. The Verb study does not show this, as can be seen in Fig 6.2 there is in fact a marked drop in this type of response in comparison to the Noun 2 study. This leads to the question of whether there aren't any methodological issues that could have inhibited *yx collocations*. Given that the methodology between the verb and noun studies was the same except for the stimulus words, the most likely source of any problem is the stimulus list itself. One difference between the Noun and Verb prompt lists (compare Appendices 5.1 & 6.1) that may have limited *yx collocations* was the decision to use 'to' in front of each verb. This was done in order to make a clear distinction with any similar sounding or similarly spelt nouns that learners might confuse them with. To ensure that responses were to verbs the prompts were *to believe, to appear* and *to develop* rather than *believe, appear* or *develop*. Giving students the infinitive form of the verb may however have unintentionally affected the number of *yx collocation* type responses. If the prompt word had been *develop* (rather than *to develop*) it is probable that more of the responses would have been *yx collocations* (as in *develop→over, develop→under)*. A further word association study without a 'to' before each prompt verb would be needed to confirm whether this is in fact an issue.

6.7.2 Why did one student make such disparate responses?

In the results section it was noted that Student 14 (Fig 6.3) was exceptional in that he generated two profiles that showed hardly any relationship at all. In Profile 1 for example there is an overwhelming dominance of *xy collocations* whereas in Profile 2 there are few *xy collocations* with responses spread out mainly in Meaning-based and Form-based categories. With an increase in the number of *similar form only* responses we might be tempted to conclude that this student was unfamiliar with a number of the verbs in PWL2, or knew all the words in both sets but had a greater depth of knowledge for those in PWL1. However, as these words all came from the 1500 – 2000 word frequency range and this student scored 95% on Nation's level test (3000 level) it seems unsatisfactory to conclude that this student simply didn't know these words well enough to respond in a characteristic way. A more likely scenario is that he interpreted the task incorrectly and attempted to give

what he thought was the *best* response to each stimulus rather than the *first one* that came into his mind as he was instructed. Another possibility is that he got bored or distracted after the first set of words and so answered uncharacteristically in the latter half of the test. For this student (and one or two other students in the Noun studies) the individual profiles were dissimilar. Whatever the reason, such outliers indicate that the individual profiling method is fallible, and underlines the point made by Milton (2009:143) that a word association test "only works when learners willingly engage with the purpose of the exercise and do not try and maximize their scores". If a word association test is to be used, then careful attention ought to be paid to group selection. It is not enough to merely consider how difficult participants are likely to find the stimulus words, there is also a need for a certain level of enthusiasm on their part and clear instructions on what is expected of them in the tasks.

6.7.3 Comparing the noun and verb studies

In the results section we saw that the responses in the Verb study generated lower correlations between each individuals' profiles than in the Noun study. It would be nice to roll up these correlation coefficients into one number, giving us a simple way to compare the two studies. Unfortunately we cannot simply average correlation values, as they are non-additive. If we wish to bring this data together into one figure to make a neat analysis between the proximities of the profiles in the Noun and Verb studies we need to use a different statistical measure. One way to do this is to calculate the coefficient of determination (R^2), this measure is additive and so can be used to sum up the values for each individual and calculate an average. This was done and the average R^2 value for the Verb study calculated ($R^2 = 0.48$), including the outlier. When the outlier was excluded the figure for the Verb study was $R^2 = 0.50$, the square root of this is gives a moderate correlation value ($r = 0.71$). This was less than the Noun 2 experiment which had an R^2 value of 0.709 ($r = 0.84$), this shows that on average the profiles in the Verb study were less strongly correlated.

Another way to compare the two studies would be to look at the number of profiles that were classified as *very close, close, vaguely similar* or *dissimilar* in both experiments.

Table 6.4 Proximity rankings in the Noun 2 (2008) and Verb (2009) studies

Correlation coefficient (r)	Definition of profile proximity	Noun 2 study (n=30)	Verb study (n=27)
>0.8**	Very close	**22**	**6**
0.7–0.79**	Close	**3**	**7**
0.6–0.69*	Vaguely similar	**2**	**4**
<0.5	Dissimilar	**3**	**10**

**p= <0.001, *p = <0.05

In Table 64 we can see that both studies show that most of the responses generated profiles that were classified as *vaguely similar, close or very close*. The Noun 2 study however has more *very close* profiles which again indicates that Nouns give more reliable responses. A point that has not yet been raised is that the threshold correlation value for distinguishing similar profiles (r=0.6) is quite high for the behavioural sciences. Were the threshold less conservative (r=0.5) fewer profile pairs in the Verb study (five) would be classified as *dissimilar*. A lower threshold could be justified if we consider that for a sample of this size a correlation would only need to be >0.46 to have statistical significance.

6.7.4 Implications of a word class effect

Having examined the two largest word classes it seems that the word class of the stimulus does have some effect on the type of responses. The implication of this is (RQ3), if word association tests are to be used to make inferences about the general organisation of a learner's lexicon then stimulus words from a range of word classes ought to be included. If just one word class is used then it will probably be biased, in the case of verbs this bias appears to be in favour of collocations. Also, in order for subsequent word association studies to be comparable, it is necessary to develop a standard distribution of word classes.

Bagger-Nissen & Henriksen (2006) and Wolter (2001) used an equal number of stimuli from three word classes, while better than single word class lists, this still doesn't really represent the distribution of word classes in the target language. In English, nouns are far more common than verbs and verbs are more common than adjectives, it seems more logical to construct a list of stimulus items that reflects this. In Zareva (2005) a 73 item stimulus list was used that consisted of 55% nouns, 23% verbs, 17% adjectives and 5% adverbs. When this is compared to a count of the number of word 'types' in each word

class within the top 6000 most frequent lemmas in the BNC, the numbers tally quite well (Table 6.5). Considering problems noted by Gardner (2007), corpus based lists currently only give us a rough measure of frequency due to unresolved issues with how some kinds of words are counted. There is also the problem that the percentages change at different frequency levels, the percentage of nouns for example increases at lower frequency levels. Other counts based on different corpora give slightly different values but all agree that in English nouns are the dominant word class. In any 'balanced' stimulus list nouns therefore ought to account for about half the stimuli, with verbs and adjectives warranting lower levels of representation (15 - 20%). In a large word list a case could also be made for including a few (6 - 7%) adverbs and perhaps one or two prepositions.

Table 6.5 Percentage of lemmas in each word class: source BNC

Word class	%
Nouns	51.63
Verbs	20.28
Adjectives	17.79
Adverbs	6.76
Prepositions	1.12
Others	2.43

Zareva's (2005) sampling procedure (words randomly picked from a dictionary) resulted in many items that are, in my opinion, inappropriate for L2 learners (*lackadaisical, putative, cassava, glower* etc.) it did however result in a word list that represents the distribution of English word classes quite well. In this respect Zareva's stimulus list can be viewed in a positive light. While the exact percentage of words in each word class (Table 6.5) ought to be treated with caution, using stimulus word lists that better reflect the target language is supported by this study. Word lists that consist of just one word class appear to give biased responses; this might give a slightly misleading view of an individual's characteristic response behaviour.

6.8 Conclusions

Although student response patterns to verb stimuli were found to be less reliable than response patterns to noun stimuli, the individual profiles generated for most students from verbs were not so unreliable as to be rejected outright. The profiles indicate that verbs are reliable enough to use, particularly if the natural bias of verbs, to generate *xy collocations*, is taken into account when constructing word lists. These findings are not as encouraging

as the Noun studies, although we do not yet know enough about how other word classes behave to make any strong claims about the effect of word class. The next step is therefore to look at another word class, adjectives. From this study we might expect adjectives to be biased towards a particular type of response as well. It would also be useful to know how reliable adjective responses are. The following chapter will therefore explore the effect of adjectives on word association responses in a similar way to the Noun and Verb studies.

Chapter Seven: The effect of adjective stimuli

7.1 Introduction

In previous chapters it has been demonstrated that it is possible to create reliable learner profiles from word association responses to nouns and to a lesser extent with verb stimuli. This chaper explores the word frequency effect a little further with an investigation of adjectives As before, the methodology follows the general format laid out in Fitzpatrick (2007).

The reason for using adjective stimuli in this experiment (from here on referred to as the Adjective study) is that, after nouns and verbs, this is the only major word class that has yet to receive attention. If generalizations are to be made about the effect of word class on individual learner profiles then decent samples from the main word classes are needed. Although ideally it would be nice to test all word classes there are serious difficulties with using the smaller word classes. As the pool of words from which to choose suitable prompt words becomes smaller, it becomes increasingly difficult to identify items that can generate useful responses. As noted in Chapter 6, the adjective word class was not attended to initially as nouns and verbs seemed less problematic. The main concerns were with the small number of adjectives within each frequency band, and also from the nature of adjectives to form particularly strong associations with their opposites. For example, *black* usually generates its opposite *white* and *strong* usually generates *weak*. As documented by Deese (1965) and Meara (1980) these strong associations are particularly characteristic of high frequency adjectives.

The studies in the Noun and Verb studies used a split-half procedure for checking the internal reliability of each individual's response characteristics. Specifically, responses to a set of high frequency words (taken from the 0-500 frequent band) were compared to responses to a set of lower frequency words (1500-2000 band). Unfortunately with adjectives it is not possible to follow the same format as there are not enough adjectives in the higher frequency band which can be used as stimulus words. In the most frequent 500 word band of the British National Corpus (BNC) 33.6% of the items are nouns, 20.2% are verbs, with adjectives accounting for 8.8% - just 44 items. This is already smaller than the 50 item prompt word lists used in the previous studies, and when those with extremely strong primary associations are whittled away there are only about 30 left. A solution to this problem is to widen the frequency range from which the adjectives are chosen. For the Adjective study high frequency items were selected from the 0-1000 frequency band. As

before the 1500 – 2000 band was used to select the lower frequency words.

In order to identify the potentially problematic adjectives online norms lists were accessed (Kiss et al., 1973; Nelson et al., 1998) which give data on the strength of primary associations. As already noted, the validity of these databases for such a task is questionable for two reasons, firstly the Edinburgh Associative Thesaurus (EAT) database is now rather old (collected in the 1960's) and secondly both databases were made using native speaker data. The assumption that the associative norms of native speakers are similar to the associative norms of learners does not necessarily follow. However, native norms lists could be viewed as a 'quick and dirty' way to weed out the words that are unlikely to generate useful data from L2 learners. Precisely how reliable these native norms lists are in identifying unproductive items is an area that I wish to clarify in this study.

7.2 Outline of the study

The basic methodology in the Adjective study is similar to the studies explained in Chapters 4 - 6, the main difference being that adjectives were used as stimulus words. Prior to the word association tests the second and third levels of Nation's Levels Test (Nation, 1990) were used to confirm students' vocabulary level was high enough to cope with the stimulus words that would be used in the word association test. Following this confirmation that the students' range of vocabulary was at an acceptable level, the prompt word lists (Appendixes 7.1 & 7.2) were given to 30 students and they were instructed to write the first English word that they thought of when they read the prompt word. As with the previous experiments a full retrospective interview (Wolter, 2001; Fitzpatrick, 2006) was not done due to perceived benefits in terms of time (collecting comments on responses while students' thoughts were still fresh) and the realisation that many responses did not require further explanation. The classification system that was used follows Fitzpatrick (2007).

The data for this study was collected during the summer of 2011; the pilot tests were done in June and the VLT/word association data in three consecutive weeks in July. In the first week just the VLT was given, in the second week half the class were given a WA test and an interview, in the third week the remaining half were tested and interviewed.

7.3 Participants

Prior to the main word association tests the word lists were checked in two pilot studies totalling 58 (28+30) students - to identify potentially 'unhelpful' items. The participants of the initial pilot study were a group of 28 Japanese university students. To ensure students were of an appropriate level they were given the 2nd and 3rd level of Nation's Levels Test, all students scored over 90% on both parts of the test. In the second pilot test a class (30) of third year Japanese high school students (age 17 – 18) were asked to help further refine the word lists. Although a slightly lower age group they were particularly able, comparable in ability to the group used in the main part of the study.

The students used in the Adjective experiment were similar to the students in the Noun and Verb experiments in terms of: their nationality (mostly Japanese), the size of the group (30), their age (20 – 25), general level of English (intermediate) and length of English study (7 - 8 years). As shown in Table 7.1 the group's mean vocabulary score on the Vocabulary Levels Test (VLT) was slightly weaker than the students in the Verb study. In the Adjective study all students had a good grasp of the 2000 level although some students had not yet mastered the 3000 level. The scores on these tests indicated that students would be familiar with all the words used in the word association test.

Table 7.1 Mean VLT scores for the Adjective study

(n=27)	2000 level	3000 level	combined
Mean score (%)	**91.4**	**74.2**	**82.8**
s.d.	8.2	18.2	12.1

Immediately after the word association test, students were given partial retrospective interviews to help with classification. During this interview period students were engaged in an activity unrelated to the word association test while the interviewer spent five minutes talking to each learner in turn about their responses.

7.4 Research questions

Following the experiments with nouns and verbs, the main concern with this study was whether adjectives could also be used to create reliable learner profiles from word association tests. In order to refine the methodology a little further it was also decided to test how accurate native norms lists were at predicting useful items. Two research questions were posed:

1. Do adjective stimuli generate reliable individual response profiles?
2. When making stimulus lists for learners, can native speaker norm lists be used to identify problematic items?

7.5 The pilot study

Given that prior studies suggest adjectives are a particularly problematic word class, there was a perceived need for a more rigorous pilot study in order to identify suitable prompt items to use with these learners. In the initial pilot study 50 high frequency adjectives (0 - 1000 band) and 50 lower frequency adjectives (1500 – 2000 band) were tested as potential prompt words. As can be seen in Appendix 7.3, 30 of the high frequency adjectives had to be rejected and 18 of the lower frequency words had to be rejected. The adjective *far* for example gave *near* 42.3% of the time and *afraid* gave *scary* 30.4% of the time. These particular prompt words would not tell us very much about the associative characteristics of the individuals. Instead they would tell us what we already know - that the word pairs *far - near* and *afraid – scary* are strongly linked. As in this series of experiments we are trying to look at the associative patterns of individuals, such word pairs need to be avoided whenever possible as they mask an individual's characteristic preferences. Of the initial 100 prompt words 48 had to be rejected for this reason. The fact that so many adjectives had to be rejected was unsatisfactory as there were not enough adjectives left to make two prompt word lists with sufficient items to generate reliable learner profiles. It was therefore decided to run another pilot study with a fresh set of adjectives to try and identify more useful prompt items.

As can be seen in Appendix 7.4 the second pilot was also problematic, of the 52 additional adjectives, 12 items also had to be rejected due to their strength of association to just one other word. Of the high frequency adjectives *able, major* and *total* all generated primary responses to just one other word >40% of the time. The lower frequency adjective *increased* generated the strongest primary response, 57.67% responded with *decrease*. Due

to the problem that many Japanese students have with distinguishing between l and r the adjectives *leading* and *correct* were also cut. With these particular items many students mistakenly responded to the words *reading* and *collect*. The second pilot test meant that a few more adjectives could be added to the initial set of words that the first pilot study identified. In total 40 high frequency and 46 lower frequency adjectives were identified as suitable for PWL1 and PWL2. The lower frequency list was trimmed down further (a few were judged to be difficult) to leave 40 words per list. Although the initial intention was to include 50 items in each word list (as in the noun and verb studies) once the pilot studies had weeded out the unsuitable items it was decided to settle for 40.

7.6 Stimulus word lists

Two word lists were created from the British National Corpus (BNC). The first list was selected from the 0-1000 frequency band, the second list was selected from the 1500-2000 frequency band. As explained in the previous section, potential words were piloted and unsuitable words rejected; this left two prompt word lists (PWL1 and PWL2) of 40 adjectives each. Unsuitable words were:

- Strongly associated to just one other word (e.g. *big* or *black*). Strong associates were initially identified using online databases of native speaker norms. Any words with >25% of their primary response being to just one other word were flagged as problematic (*big* for example gives *small* 29% and *little* 18% of the time)
- Common collocates of Japanese words, such as *single* (used in hotels, as in a *single* room).
- Adjectives with multiple meanings, such as *right*.
- Difficult to classify due to belonging to more than one word class, such as *relative*.

The pilot studies highlighted the problem of selecting suitable prompt words that generated responses from a range of words and were not too strongly linked to just one word. Consequently, prior to analysis a post-hoc check of responses was conducted to identify any 'unhelpful' prompt words that might have slipped through the initial screening. Not surprisingly, despite the pilot tests, there were still some stimuli in each word list that proved to be unsuitable as they had a within group primary response >25%. These words, shown in Table 7.2, were consequently cut from the main analysis. This meant that the 40 item lists had to be trimmed further, 33 items per list were available to create the student profiles.

Table 7.2 Items rejected following within group analysis

PWL1 rejects	primary association	%	PWL2 rejects	primary association	%
national	country	34.6	**entire**	whole	31.6
different	same	29.6	**educational**	school	40.7
significant	important	35.0	**wonderful**	great	37.0
foreign	country	50.0	**ancient**	old	48.0
necessary	need	35.7	**used**	old	32.0
concerned	think	29.6	**odd**	strange	34.8
original	first	33.3	**elderly**	old	29.6

7.7 Results

In this section the following are reported:

7.7.1 Completion rates of PWL1 and PWL2

7.7.2 General trends in the group

7.7.3 The proximity of individual profiles

7.7.1 Completion rates of PWL1 and PWL2

As in previous studies the threshold for including an individual in the analysis was 25 responses. It is difficult to have much confidence in learner profiles generated from fewer responses than this. The initial analysis of the responses indicated that three out of the 30 students fell well short of the required 25 on at least one of their response forms, these three were cut from the analysis. As can be seen in Table 7.3, of the 27 learner profiles that were used in the main analysis, completion rates were generally high, >30 responses per profile with the high frequency adjectives and 25 - 30 responses for the lower frequency adjectives. Two of the learners completed only 24 responses on PWL2. Given that the original level of acceptance was arbitrarily set at 25, it was decided to include these two whose completion rate fell just shy of the threshold.

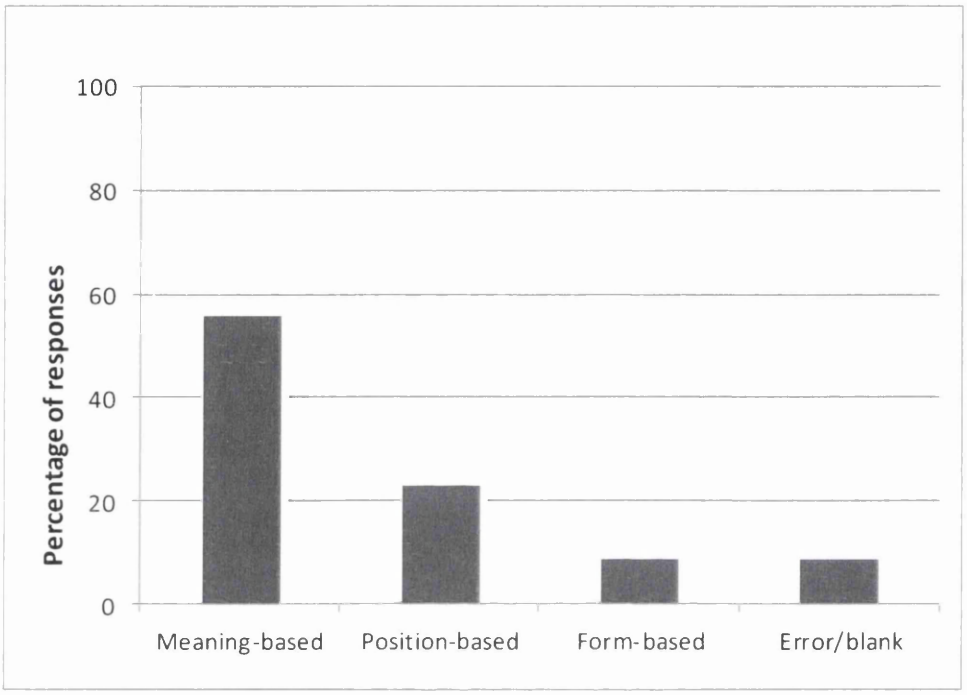
Table 7.3 Completion rates for responses to PWL1 and PWL2

n= 27	PWL1 (max 33)	PWL2 (max 33)
Mean	31.2	27.5
s.d.	1.8	3.5

7.7.2 General trends in the group

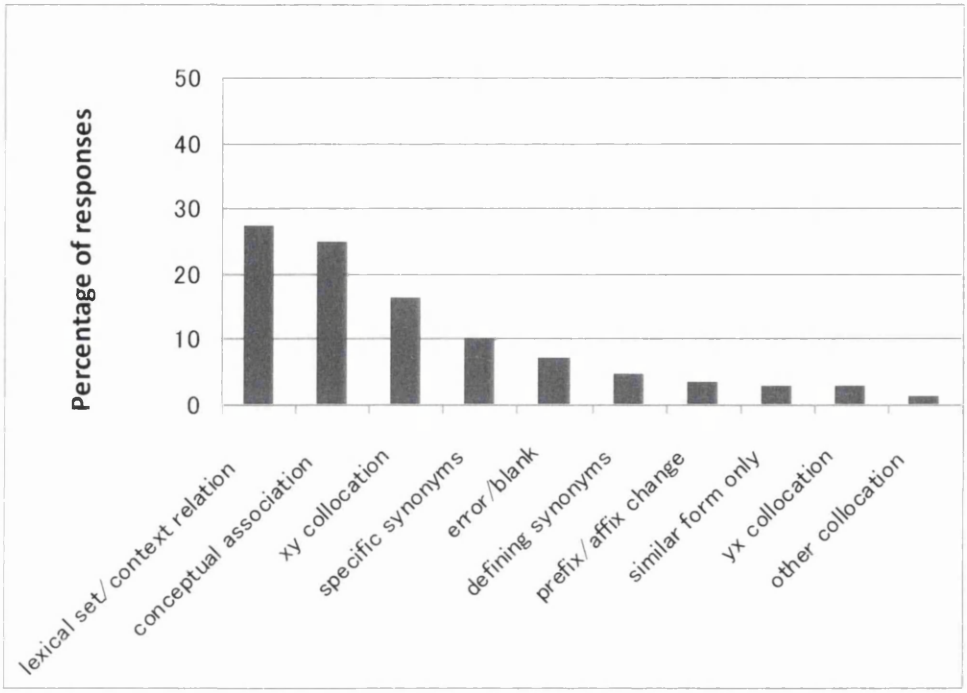
As Fig 7.1 shows, most of the responses to the adjective prompt words were Meaning-based. This trend is in line with the previous studies in this series, the noun and verb studies were also dominated by Meaning-based responses.

Fig 7.1 General trends for adjective responses: main categories



When the broad data shown in Fig 7.1 is broken down into the subcategories (Fig 7.2) we see that of the Meaning-based group, most of the responses are from two subcategories - the *same lexical set/context relationship* category and the *conceptual relationship* category. Of the Position-based categories it is the *xy collocation* category that dominates. In the Adjective study students gave very few Form-based responses.

Fig. 7.2 General trends for adjective responses: subcategories



Unlike the Verb study, where the bias was in favour of *xy collocations*, with adjectives there were a lot of *lexical set/context related* responses, coordinates such as *clear→dark, normal→strange, urban→rural* were common.

7.7.3 The proximity of individual profiles

A Pearson's correlation was calculated for each individual by comparing the percentage of responses in each PWL1 subcategory to the percentage of responses in each PWL2 subcategory. As can be seen in Table 7.4, when a learner's profile 1 (generated from responses to the high frequency adjectives) was compared with his/her profile 2 (generated from responses to lower frequency adjectives) the two profiles were generally found to have a similar pattern. Of the 27 sets of profiles 18 were defined as being *close* or *very close*, that is, they had a correlation coefficient of >0.70. The level of consistency in the responses by individuals, in Table 7.4, shows that these adjectives can be used to generate reliable profiles with Japanese learners. Over 90% of responses profiles were judged to be at least *vaguely similar* with 40.7% of the profiles judged as being *very close*. Only two profiles were defined as *dissimilar*, although it might be noted the correlation coefficients of these two (r=0.565; r=0.550) were only slightly below the 0.6 threshold and were statistically significant at the 95% confidence level.

Table 7.4 Proximity ranking for adjective profiles

Correlation coefficient (r)	Definition of profile proximity	Number of students
> 0.8**	very close	**11**
0.7 – 0.79**	close	**7**
0.6 – 0.69*	vaguely similar	**7**
<0.6	dissimilar	**2**
Total number of students		27

**p = <0.001, *p = <0.05

The results of this study therefore indicate that adjectives can be used to generate reliable response profiles. The answer to the main research question is therefore affirmative.

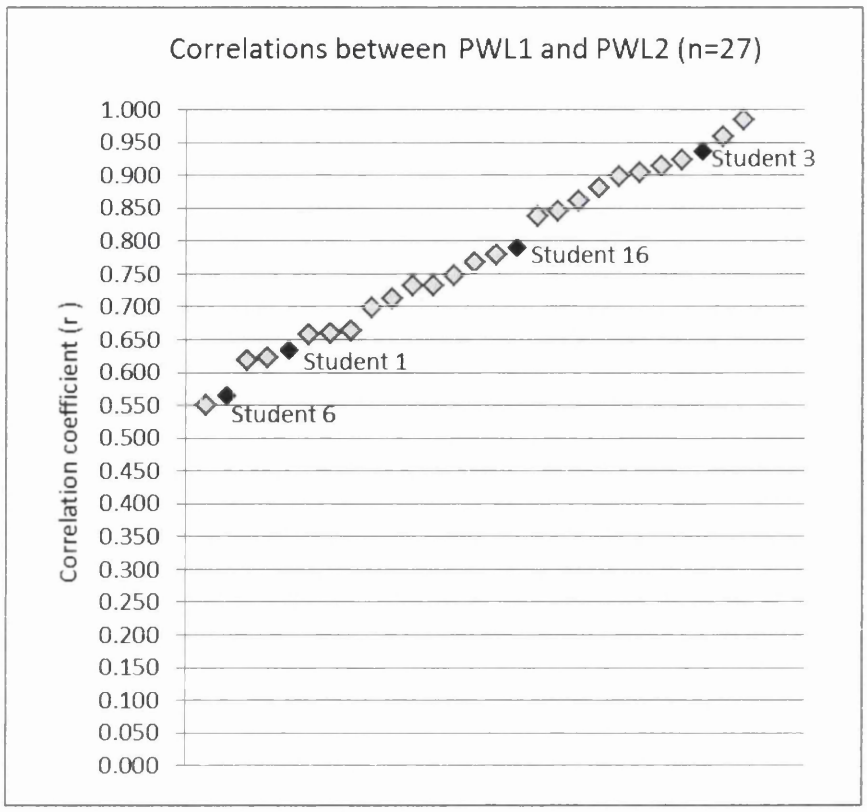
7.8 Discussion

In this section I will initially focus on four case studies that will be used to help identify some general issues and lead into a discussion of some of the problems that are inherent in the study of adjectives. Next, the second research question will be addressed – How useful are native norms lists? Finally there will be a discussion of how typical the adjectives in this study are. Due to the large number of potential stimuli that were rejected following pilot studies an argument could be made that the adjectives finally selected for analysis were not representative of adjectives in general.

7.8.1 Individual profiles: four case studies

The four case studies, indicated in Fig 7.3 were picked to illustrate the spread of profiles that the individuals generated. As noted before only two students fall below the 0.6 threshold. The first case study (Fig 7.4) is an individual classified as having *dissimilar* profiles. The second and third examples (Figs 7.5 & 7.6) are classified as *vaguely similar* and *close* profiles. The fourth example (Fig 7.7) is a student who generated *very close* profiles.

Fig 7.3 Correlations between PWL1 and PWL2 in the Adjective study



As noted in the Noun and Verb studies, the reason for this 'individual' approach was in response to a perceived lack of homogeneity within 'group' data. The examples given here clearly illustrate yet again that, even with citizens of a country renowned for its sense of common identity and conformity to group norms, when it comes to word association responses the patterns of behaviour are markedly different. Two of the case studies characteristically gave a lot of *xy collocations* whereas the other two gave a lot of *specific synonym* responses.

Fig 7.4 *Dissimilar* profiles

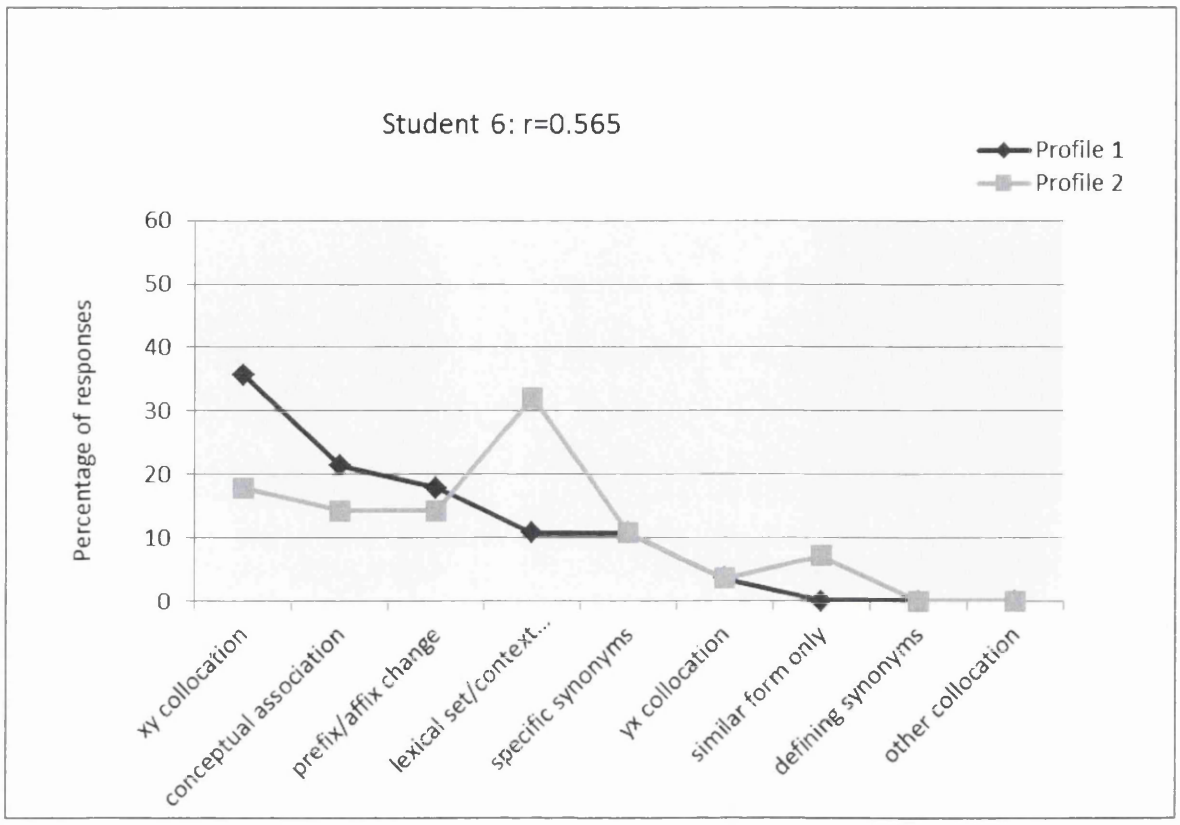


Fig 7.5 *Vaguely Similar* Profiles

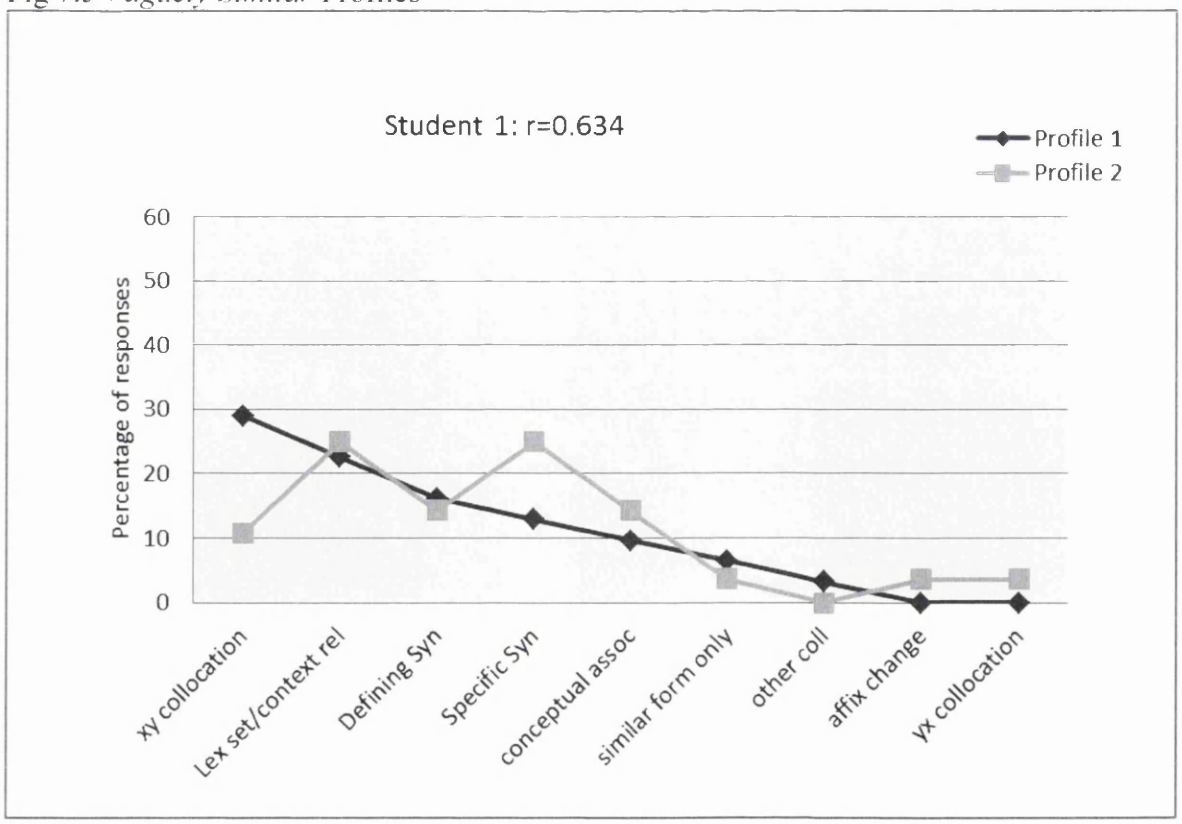


Fig 7.6 *Close* profiles

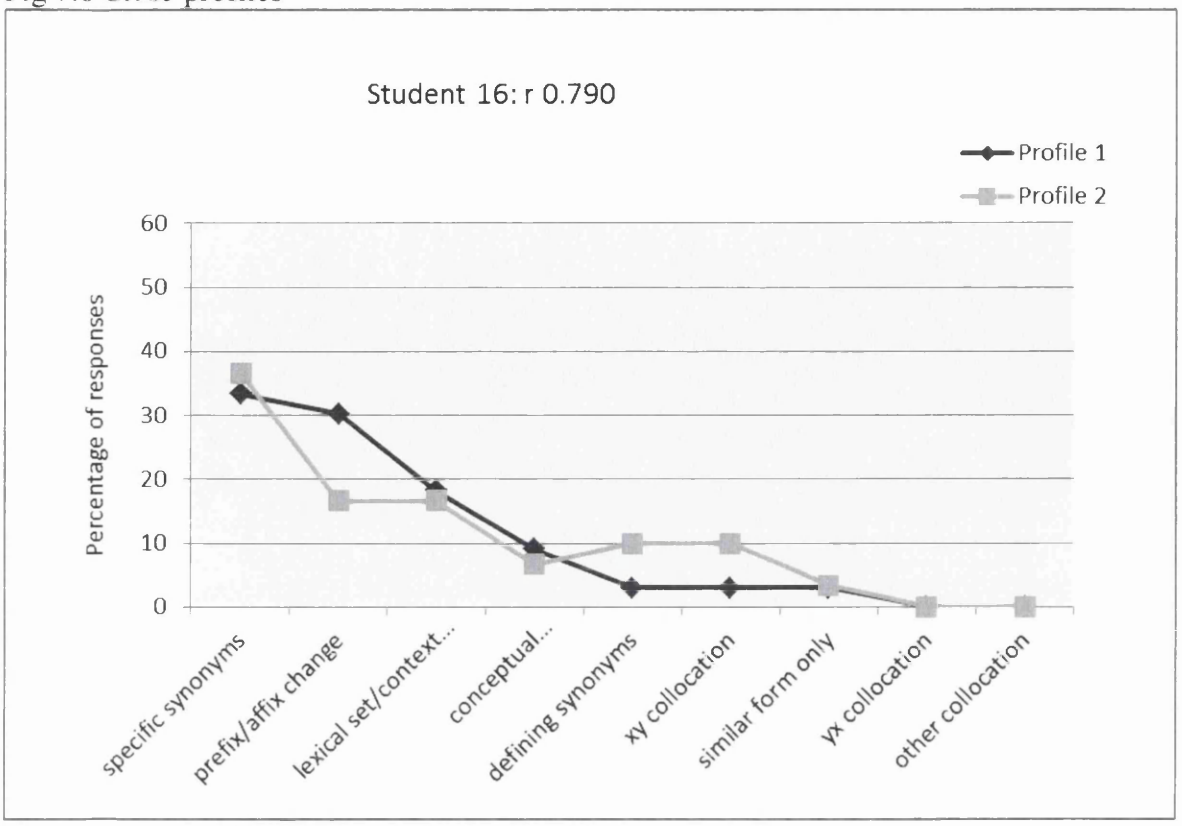
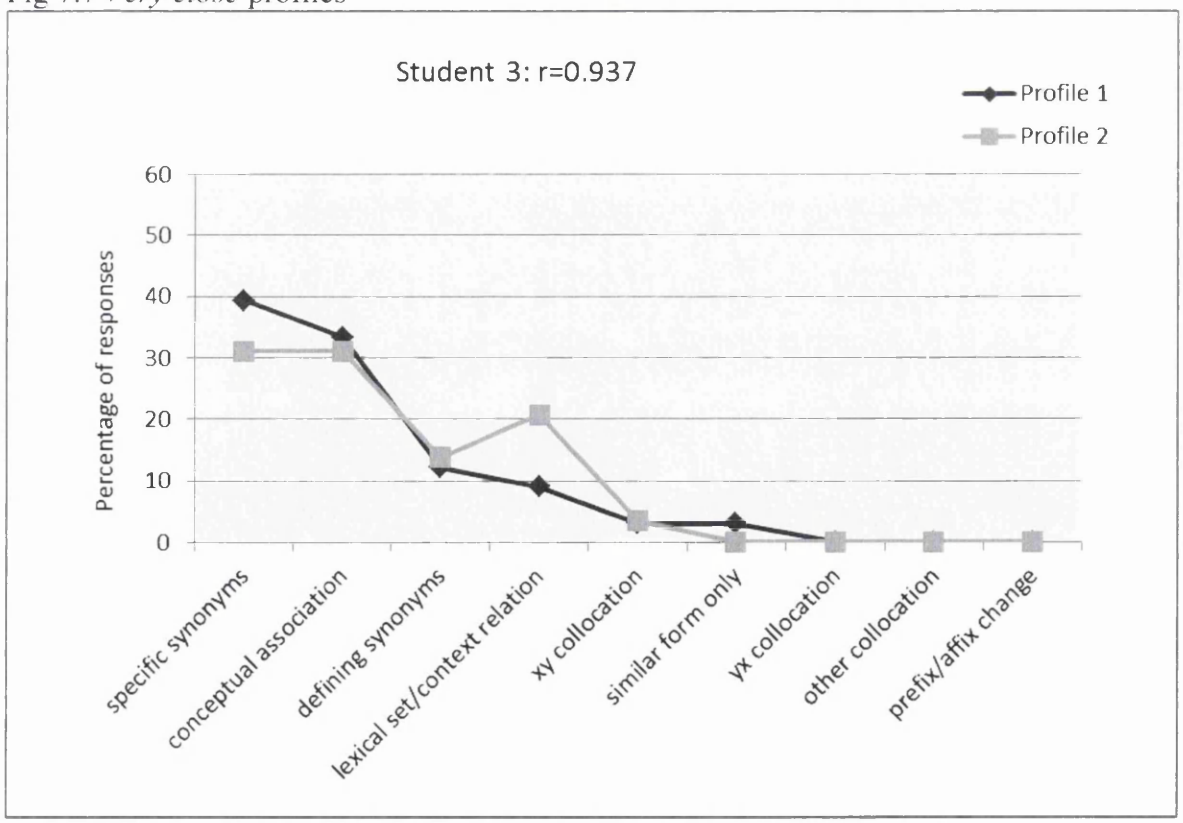


Fig 7.7 *Very close* profiles



When comparing Figs 7.4 – 7.7 it is clear that they each have quite different profiles. If we set aside the dissimilar profile and examine the students who gave profiles that had collocations >0.6, their dominant categories are:

Student 1: xy collocations, same lexical set, synonyms

Student 16: specific synonyms, prefix/affix changes, same lexical set

Student 3: specific synonyms, conceptual associations

These results strengthen Fitzpatrick's (2007) argument that we should be analysing word association data from an individual rather than a group perspective. The variety of profiles generated demonstrates that Japanese learners cannot be considered a homogenous group.

Another trend that is evident from these case studies is the lack of *yx collocations* or *other collocations*. If we look back to the graph of general trends (Fig7.2) we can see that the lack of *yx collocations* or *other collocations* responses is not limited to these four students, very few *yx collocations* were given by any of the learners. As these are second language learners we might have expected fewer *other collocations* as idioms and proverbs are unlikely to be produced until an advanced level of proficiency is attained. But what about *yx collocations* - why were there so few in this category? The lack of *yx collocations* was also noted in the Verb study, a possible reason given there was that the use of the infinitive form (*to believe, to hold* etc.) may have been a factor. As the *yx collocations* are also lacking in the Adjective study we can perhaps discount this reason, this phenomenon is not restricted to one word class. In the Adjective study Student 1 for example gave quite a lot of *xy collocations* in both profiles. So, why is it that this student made so few *yx collocations*? There are various possibilities. In Fitzpatrick (2006) the non-native speakers only gave a few of these responses, perhaps indicating that awareness of these kinds of collocations is only picked up on by advanced learners. Another possibility is that when the prompt words were selected there was no conscious attempt to ensure that all the words had an equal chance of being responded to in the nine possible ways. It might be the case that the particular prompt words used here were simply unlikely to generate *yx collocations*. When we look at the prompt lists (Appendices 7.1 & 7.2) however it is possible to find numerous examples of stimulus words that have a very likely *yx collocation*. From PWL1 for example we might expect: *special→today's* and *public→general*. In PWL2 there are: *married→happily*, *official→public* and *careful→be*. Based on a recent study by Shin & Nation (2012) that listed the top 100 collocations found in the BNC, I would have expected the PWL1 stimulus *sure* to generate the collocation *make* (*make sure* is ranked 55th) or *not*

(*not sure* is ranked 89th). These two responses were not however made by any of the students in the Adjective study, the top four responses to *sure* were in fact: *OK, certain, right, yes*. While the reasons remain unclear, I think the most likely possibility for the lack of *yx collocations* is the format in which the responses were collected. As can be seen in Appendixes 7.1 & 7.2 when students read the prompt word they were asked to write a response in the space to the right of the word. It is probable that such a format encourages respondents to make right to left collocations, such as *public →enemy* rather than left to right collocations such as *public →general.* If the space to write responses were on the left or perhaps under the word we might see an increase in *yx collocations.* An oral testing procedure could also be used to see if the written format in general had an effect. Further testing using variations on the format for response collection are recommended to check this.

As the number of *yx collocations* and *other collocations* are particularly low in general perhaps it would be better not to include them in our calculations as they make the correlation figures seem stronger than they really are. In the above examples (Figs 7.4 – 7.7) the *yx collocation* and *other collocations* subcategories are all near zero, consequently they correlate nearly perfectly and so inflate the overall correlation values. When these two categories are removed the correlations between profiles 1 and 2 are not quite as strong as they first appear.

Student 6:	r = 0.565	drops to r = 0.437
Student 1:	r = 0.634	drops to r = 0.493
Student 16:	r = 0.882	drops to r = 0.842
Student 3:	r = 0.937	drops to r = 0.924

Calculating coefficients by ignoring inconvenient categories is not however a satisfactory solution, and it therefore seems better to look at the data in a different way. Another way would be to simply consider the dominant response subcategory. Is the dominant subcategory in profile 1 also the dominant subcategory in profile 2? In the Adjective study the answer is 'yes' for 70.4% of the learners. For example, with Student 16 we can see that *specific synonym* responses were the dominant form of response with both sets of stimulus words, in both profiles they account for over 30% of responses in each profile. The characteristic response for this particular learner could be said to be words that are nearly synonymous with the stimulus words. When for example this student was given the high frequency stimulus *similar* he responded with *same,* which in some specific situations could be used interchangeably. When he was given a lower frequency stimulus *quiet* he

again characteristically responded with a near synonym *silent*. Looked at in this way, there seems to be further confirmation of the initial research question: *individuals respond to adjective stimuli in a reliable way*.

7.8.2 The value of native norms lists

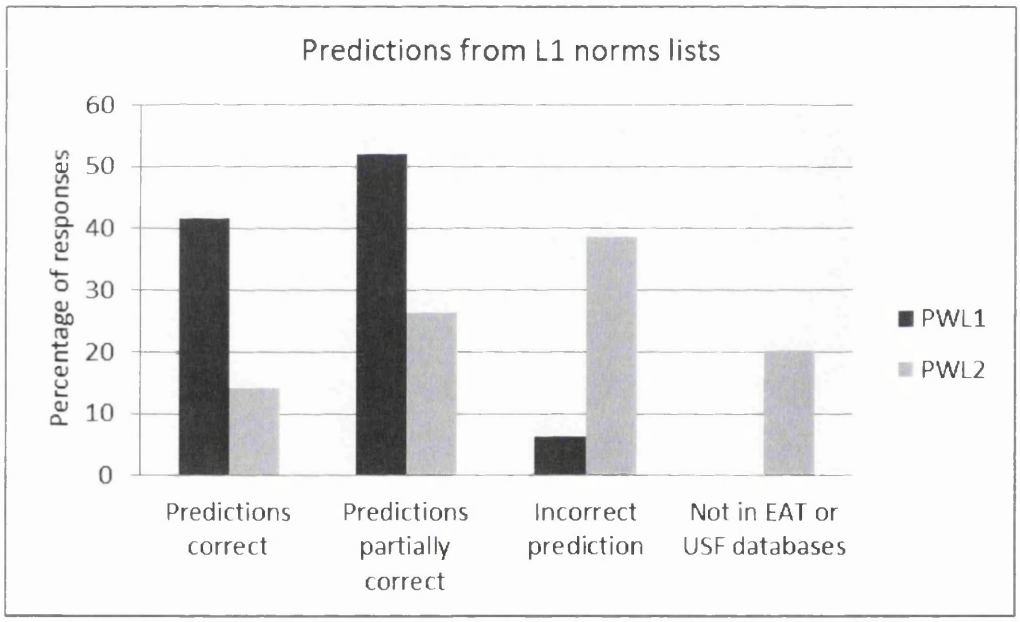
Prior to the pilot studies, online databases of native speaker associations were used to help identify potentially problematic prompt words. The idea that native norms would provide a simple method for filtering out the more extreme items was based on the tenuous assumption that native associative norms are similar to learner associative norms. To test this assumption, responses given in this study were compared with two norms lists, The Edinburgh Associative Thesaurus (Kiss et al, 1973) and the University of South Florida (Nelson et al, 1998) both of which have databases that can be searched online. The main concern with these databases is that they are comprised of associations made by people living in quite different cultural and linguistic environments to the learners in this study. While we may guess that some native associations will indeed match learner associations it is unwise to assume that all native speaker associations match all learner associations. It would be helpful to be able to say with a little more precision how far we can rely on native speaker norm lists to predict L2 learner responses. In other words we are asking: How well can a native speaker norms list predict the usefulness of stimulus words that are to be used with L2 learners? To try to answer this question the primary responses from the EAT and/or USF databases were compared to the responses in the Adjective study. Responses were grouped into four categories (Table 7.5).

Table 7.5 Criteria for classifying the native-norms list predictions

Prediction correct	The norms lists predicted whether the prompt word would have a primary response greater or less than 25% and also predicted the exact primary association.
Prediction partially correct	The norms list predicted which side of the 25% threshold the prompt word would be, but incorrectly predicted the primary response word. Or The norms list did not predict which side of the 25% threshold the prompt word would be, but did correctly predict the primary response word.
Incorrect prediction	The norms lists failed to predict whether or not the response would be greater or less than the 25% threshold, and also failed to predict the primary response.
No data	The adjective was not listed in either the EAT or USF databases.

As can be seen in Figure 7.8, the high frequency adjectives (PWL1) were actually fairly predictable, with the norms list accurately predicting over 40% of the responses and being useful in identifying prompt words that would/would not give strong primary responses (a further 52%). An example is the word *possible*, the EAT data shows that for this word 38% of native speakers responded with *impossible*, this was similar to the finding in this study that 55% of learners responded with *impossible*. The lower frequency adjectives were less predictable, an example is the word *equal* which was not flagged up by the norms list because the strongest response was only 15% (to the word *same* in EAT), in this study the strongest response was also to *same* but at 33.3% had to be rejected from the main analysis. Another problem with the lower frequency adjectives is that many are not listed in either of the native speaker databases (20% of those used in this study).

Fig 7.8 Predictions from L1 norms lists



Given that the high frequency adjectives are the most troublesome, these findings are encouraging, L1 word association norms accurately predicted many of the problematic high frequency items. Beyond the top 1000 most frequent word range the databases currently available seem to be of little help in predicting learner responses. In answer to the second research question it can be concluded that: *native speaker norms lists are useful as a rough guide to predicting problematic adjective stimuli drawn from the most frequent 1000 range but are less useful for lower frequency adjectives.*

As this study has shown, using native norms lists alone is insufficient when

selecting which adjectives to use in a word association test. Piloting of potential adjective stimulus words is also necessary in compiling the stimulus word lists. With the verb and noun stimulus word lists it was enough to pilot the lists once, this led to two or three words being omitted. The fact that the initial pilot test led to nearly half the potential stimulus words being dismissed highlights the care that is needed when selecting adjective stimuli, particularly high frequency adjectives. As also noted in previous L1 studies, the tendency of high frequency adjectives to be strongly matched to their polar opposite, *hot-cold* and *hard-soft*, makes this word class particularly challenging to use as a means of identifying an individual's characteristic response patterns. As L1 norms lists can be viewed as a 'rough guide' in identifying unproductive stimulus items for L2 learners, it is argued that using them in conjunction with a pilot study will result in lists than can generate useful responses. As each L2 is likely to differ with regards to which stimuli are productive, even if age and language ability are accounted for, it would be unwise to use the words in this study with a group of non-Japanese learners without first piloting them. Unless a researcher has the time to carefully pilot adjective stimulus words it would be better to stick to noun and verb prompt words.

7.8.3 How typical were the adjectives in this study?

Due to the large number of adjectives cut following the pilot studies, one concern was that the final list of adjectives selected for the experiment might be odd or somehow untypical of adjectives in general. A useful way to think about whether words are typical of the category that they are assigned to is to use prototype theory. As Aitchison (1992) explains, when we try to classify words into discrete categories it is often difficult as some items exhibit more of the characteristics of the category than others. She exemplifies this with a study of 200 London school children on classifying items, one of the categories she asked students to classify was birds. She found that when asked how bird-like various birds were the students viewed *blackbirds* and *robins* as being particularly good prototypes (they have feathers, lay eggs, nest, can fly, etc.) but that birds such as *penguins* and *peacocks* were not (they are quite big, don't fly, etc.). In much the same way, adjectives can be viewed along a cline from prototypical adjectives, such as *beautiful,* to atypical adjectives, such as *utter*. As a measure of adjective prototypicality I will use the four main attributes suggested by Greenbaum & Quirk (1998:129). An adjective can:

-freely occur in an attributive position, e.g. A *beautiful* painting.

-freely occur in a predicative position, e.g. The painting is *beautiful.*

-be pre-modified by an intensifier, e.g. The painting is very *beautiful*.

-take comparative and superlative forms by either adding er/est or by being preceded by more/most, e.g. That's the most *beautiful* painting I've ever seen.

To Greenbaum & Quirk's list I will also add:

-form an adverb through adding –ly e.g. She paints *beautifully*.

By these five criteria we can view an adjective such as *beautiful* as being one of the most prototypical of adjectives as it fulfils all of the above conditons. An adjective such as *utter* though only fulfils two of the five criteria and can therefore be said to be atypical.

When compared against these criteria most of the 66 adjectives analysed in this study (Appendices 7.1 & 7.2) could be viewed as prototypical adjectives, in PWL1 for example 81% had all five characteristics with 91% having at least four. Of the PWL2 adjectives 56% had all five characteristics with 79% having at least four. The only atypical adjective in PWL1 was *previous*. In PWL2 there were two aypical adjectives *chief* and *overall*. Of these *overall* was the most atypical; although it can be used in an attributive position, it does not possess any of the other adjective characteristics.

In short, the adjectives that were used in this study were not strongly associated to just one other word (therefore could potentially give a variety of response types) and were also, on the whole, fairly typical examples of adjectives.

7.9 Summary

The main findings are that the adjective stimuli used in this experiment generated reliable learner profiles. Unfortunately they were much harder to work with than other word classes, necessitating multiple trials of prompt words and post-hoc tests in order to identify suitable stimuli that could provide meaningful responses to analyse. The tendency of some adjectives to strongly associate with their polar opposite, and therefore mask characteristic responses, means adjective stimuli need to be selected with care.

Another interesting finding is that native norms lists can be helpful in sorting the high frequency adjectives that are likely to be useful items from the problematic ones. Native norms lists can therefore serve as a coarse filter to separate productive stimuli from unproductive stimuli. When used as an initial step in the preparation of stimulus lists, it is argued that they allow pilot tests to more efficiently sift through and pick out the stimuli that have the best potential to generate useful responses for the L2 group being tested. While this study did not specifically examine the usefulness of norms lists in aiding the selection process of stimuli from other classes it seems likely that this is also the case.

7.10 Conclusions

It has been shown that it is possible to generate learner profiles that have a high degree of internal reliability using nouns (Chapters 4 & 5) and to a lesser extent verbs (Chapter 6). With the finding that it is also possible to generate profiles that reliably show a learners response characteristics using adjective stimuli, we can conclude that word class does not seem to have much of an effect on the reliability of the response profiles. That is not to say that word class doesn't have any effect, as it does, each word class seems to be biased toward a particular type of response. However a question hanging over the findings raised in Chapter 4, that the reliability of responses might vary with stimuli from different word classes, has now been removed. In this respect some progress has been made.

Given that adjectives, particularly high frequency adjectives, require careful selection, piloting and post-test checks a good argument can however be made for leaving them out of stimulus lists altogether. A researcher who does not have the time to screen adjectives carefully would be advised to use the less problematic nouns and verbs. For research within a Japanese context using university aged subjects the stimulus adjectives that have been used in this study (Appendixes 7.1 & 7.2) could of course be used. With subjects from different backgrounds though, it would be sensible to pilot all adjectives prior to use.

As the other word classes get progressively smaller it would seem fruitless to continue in the vein of previous chapters and try to verify yet more word classes, based on studies of the three largest word classes we can assume that word class has a minimal effect on test reliability and leave it there. Rather than continue with another word class a slightly different tack will be attempted, reanalysing one particular subject. One reason to do this is that most of the students in the Noun 2, Verb and Adjective experiments were drawn from different cohorts, limiting the strength of any claims based on comparisons between these studies. As has already been argued, the value of such group data is also questionable due to the lack of response homogeneity within the groups. The student that will be examined in the subsequent chapter is however unique in that she participated in all three experiments. A detailed case study approach, spanning all the word class studies, therefore allows for more valid comparisons of responses between word classes than an analysis based on comparisons between the group data alone. The group data suggests that a typical Japanese subject would mainly respond to the nouns and adjectives with a mix of *same lexical set/context related* and *conceptual associations* responses and respond to

verbs with *xy collocations*. On the other hand, due to the considerable within-group variation in response patterns we may find that this individual responds in quite a different way to each word class. Another reason to take a case study approach is that with this particular student there was also the opportunity to conduct further tests. Retesting this subject on the same items (after a considerable time gap/gain in proficiency) could give insights into how characteristic preferences change over time. Based on Fitzpatrick's 2007 and 2009 findings it would seem likely that over time (as a person's proficiency increases) their responses will more closely resemble their L1 characteristic responses. Although we might expect the individual to retain a similar level of internal reliability, some kind of change in characteristic responses over time would therefore be expected.

Chapter Eight: And then there was one

8.1 Introduction

In previous chapters I have shown that it is possible to create reliable learner profiles from word association responses. In short, the classification system and "individual profiling" approach laid out by Fitzpatrick (2007) has held up surprisingly well to analysis from a variety of perspectives. Having analysed the responses of 134 learners, I can state with some confidence that this approach is robust; it has been demonstrated that it can cope with stimulus items chosen from different frequency ranges and also words from different word classes. Having removed many of the question marks that surrounded initial attempts to collect word associations and analyse them in this way, we can begin to move forward in applying these findings. Before we explore how such findings might be used though, this chapter will focus on one L2 learner. As well as looking in greater detail at how this learner performed in each of the word association tests so far this individual was retested on the Noun 2, Verb and Adjective stimuli, allowing us to see how responses change (or not) over time. It is expected that the retest data will not only confirm the reliability of the individual profiling approach but also allow additional insights into how the learner lexicon is structured. With only one learner to consider this study also takes advantage of an opportunity to trial a methodology that would require considerable effort in large group studies – concurrent think aloud.

One of the problems with the analysis of group data is that the idiosyncratic and/or erroneous responses (and L2 learners by definition are apt to make many of these) are often misinterpreted or ignored. With a single subject study however there is a greater opportunity to dig deeper into the cognitive processes underlying these kinds of responses. Another problem is that, due to the considerable within-group variation observed in the studies reported in previous chapters, no strong claims can be made from comparisons between the different groups. Although it is unusual within behavioural science for experiments to limit themselves to just one person there is a precedent within psycholinguistics for this kind of study (Galton, 1883; Ebbinghaus, 1885; Churchill, 2007; Meara, 2011). In the introduction to a collection of single-case studies Meara (1995: iii) comments that "vocabulary acquisition is a lot more varied and individualistic than we sometimes pretend. Details like this tend to get lost in large scale studies".

As well as focusing on just one individual the other important point about this experiment is that in retesting the student using the same instruments, response patterns can be observed over time. In a longitudinal study using word associations as one of four measures of lexical growth, Schmitt (1998a) reported how three students made associations to 11 words at six month intervals over the period of a year. In that study he argued the need for measuring lexical development over a considerable time period due to the "incremental" way in which words are acquired. There are many aspects of word knowledge (Nation, 2001:27, details 16) that need to be acquired before a word can be said to be fully known. Any study tracking lexical development therefore requires considerable time in order for the learner to meet the word in various contexts and so acquire these aspects. One criticism of Schmitt's study though is that given the words chosen (*abandon, brood, circulate, dedicate, illuminate, launch, plot, spur, suspend, trace*), which are all fairly infrequent items, intervals of six months do not seem sufficient to allow the amount of exposure necessary for incidental learning to have a measurable impact. This criticism might also be applied to a more recent L2 study of lexical organisation (Crossley et al., 2009) that also investigated the network development of a limited number of learners (6) over a year. In contrast, in the current experiment there is a much longer period between each testing session (between one and three years). Also, the stimulus words (Appendices 4.1 – 7.2) are of a higher frequency than those in Schmitt's study, a proportionately higher amount of exposure (and thus acquisition) can therefore be expected.

Added to this, general language proficiency tests (TOEIC) and more specific vocabulary tests (Vocabulary Levels Tests) that the participant has taken over the years allow us to gain some insight into the effect of proficiency on response patterns. Although we might expect some change in response patterns, it is not clear quite what these changes might be. There are two main possibilities that the literature predicts. On the one hand, L1 studies (Ervin 1961, Entwistle 1966) and also some L2 studies (Politzer, 1978; Söderman 1993) would lead us to expect the participant to generate more "child-like" responses with words that are less well understood (the lower frequency items) and give more native-like responses to the more frequent items. These studies would predict that the more times the learner meets each item the less attention she will pay to the word's formal aspects, giving more attention to semantic aspects. As well as fewer Form-based responses, as proficiency increases such studies would also predict fewer syntagmatic (Position-based) responses and more

paradigmatic (Meaning-based) responses. In the second set of tests more Meaning-based associations (such as synonyms) might be expected. For example, if the stimulus word *dust* were only partly understood at the time of the initial test a response such as *must* might be generated (similar form) or *pan* (a collocation) which after one or two years might develop into a more 'mature' meaning-based response such as *rubbish* (a near synonym).

On the other hand, more recent studies (Wolter, 2001; Bagger-Nissen & Henriksen, 2006; Fitzpatrick, 2006, 2007, 2009) question such a syntagmatic – paradigmatic shift, meaning that a different kind of change in response type might be expected – or perhaps no change at all. Rather than a 'shift' towards native-likeness Fitzpatrick suggests that we can expect an individual to move closer to their L1 characteristic. Following this line of thinking we would therefore predict that a learner will make a similar proportion of responses in each response category irrespective of their proficiency level. Coming back to the *dust* example, with increased knowledge of this word an initial response like *pan* (a collocation) might remain the same or perhaps be replaced by a response such as *bin* (another collocation). Such a response (different but of the same type) would suggest that there had been some development in knowledge for this word even though the response type remains unchanged.

8.2 The participant

The individual used in this study, from here on referred to as M, was selected for three reasons. Firstly M was fairly representative of other students in previous chapters in terms of her length of study, language ability and background. Secondly, having been part of the previous three studies (Chapters 5, 6, 7) there were already considerable data available on her word association response characteristics and language ability with which to compare any retests. Finally, as a fairly keen and self-motivated learner, in the interval between testing she engaged in considerable explicit language study (including informal exposure to the target language through reading and general conversation). It is not unreasonable to expect that most of the words used in the initial study were incidentally met many times in different contexts between the initial tests and the retests. We can therefore assume that the various aspects of word knowledge (Nation, 2007:27) for each of the words tested would have been more fully acquired than in the initial tests. By the second test, the assumption is that the depth of knowledge for most of the stimulus words will have improved.

Concerning M's general language ability, the measures we have for her are a TOEIC score of 801 in 2006 which improved to 880 in 2010, based on this she can be said to have been of an intermediate level at the start of the project and had begun to move to a higher-intermediate level midway through. These scores tally well with the observation that M is a keen, self-motivated student steadily progressing in her language studies. As well as her general language ability, she was also repeatedly tested on her vocabulary knowledge using an improved version (Schmitt, 2000) of the Vocabulary Levels Test originally developed by Nation (1990).

Table 8.1 VLT scores before and after the word association tests.

	2000 level	3000 level	5000 level	10,000 level	academic
2006	**93.3%**	**53%**	**30%**	----------	**56.7%**
2012	**100%**	**96.7%**	**90%**	**63%**	**93%**

(30 items per level)

As can be seen in Table 8.1, prior to the initial word association tests M had a good command of the 2000 word level, some knowledge of the 3000 and academic levels and a partial knowledge of the 5000 level. It might be noted that she gave up after the 5,000 level of the test, so no data are available on her ability with the 10,000 level. This contrasts with the recent test data (2012) in which she demonstrated mastery at the 2000 level and a high level of proficiency at the 3000 and academic levels, she also showed good coverage of the 5000 level. Not surprisingly though (as her studies were solely within Japan) she has still not mastered the 10,000 word level. Added to this, we might also note that prior to all the initial word association tests (2009, 2010 and 2011) all participants were required to take a 60 item 2000/3000 level test to ensure they had sufficient ability to cope with the stimulus words used in the tests. In these tests M scored highly throughout, in 2009 for example she averaged 98% and in 2010 and 2011 scored perfectly.

8.3 Outline of the study

The basic methodology in this study (following Fitzpatrick, 2007) is similar to the studies explained in Chapters 4 - 7, there are however two important differences. The first is that the three retests were given to M over the period of a week within August

2012. The initial word association tests were done at yearly intervals (noun stimuli 2009, verb stimuli 2010 and adjective stimuli 2011). Consequently, the gap between the two noun tests is three years, two years for verbs and one for adjectives. The other main difference is that unlike the initial tests that were completed silently, the participant was trained in the concurrent think-aloud technique and encouraged to verbalise all conscious thoughts. During each retest a recording was made on a digital voice recorder of these verbalisations. It has been demonstrated that the think-aloud procedure (Van Den Haak et al., 2003; Leow & Morgan-Short, 2004; Albrechtson et al., 2008), while increasing task time, does not react with cognitive tasks and is likely to generate useful qualitative data about what the respondent is thinking while making responses. It is hoped that such verbalisations can resolve an issue that many researchers (Meara 1983, Wolter 2001, Henrikson 2008) have commented on, the difficulty in correctly categorising all word association responses. Both the think-aloud procedure and the retrospective interview were used in the retests. The intention being to compare them in terms of the quality of the data obtained and the effort needed to implement such checks. To summarise:

- Over the period of a week, three word lists were given to M; these were identical to the lists given in chapters 5 – 7 (Appendices 4.1, 5.2, 6.1, 6.2, 7.1 & 7.1). The words that were deemed unsuitable in those initial studies, and therefore not used in the analyses, were not used in this study either. This accounts for the unequal number of items in each study.
- In the noun list there were 96 items, half of which were high frequency (0-500 frequency band) with the other half comprising of lower frequency words (1500-2000 frequency band).
- In the verb list there were 94 items, half of which were high frequency (0-500 frequency band) with the other half comprising of lower frequency words (1500-2000 frequency band).
- In the adjective list there were 66 items, half of which were high frequency (0-1000 frequency band) with the other half comprising of lower frequency words (1500-2000 frequency band).

As before, immediately after the word association test a retrospective interview was administered, to help with classification. In previous chapters individuals were only asked about items that on a cursory inspection seemed ambiguous. In the retests there was sufficient time to allow all words to be verified. Although present during the

think-aloud practice sessions, the rater left the room while the main tests were being done so that the subsequent retrospective interviews were not influenced by the think-aloud verbalisations. The analysis of the think-aloud data was also left until last in order to keep this addition to the methodology as separate as possible and allow for comparisons with the data obtained from the retrospective interviews.

8.4 Research questions

The main question to be addressed is whether when digging deeper into M's responses there is evidence to support the claim, indicated in previous chapters, that individual response behaviour is consistent. With word association retest data and also data available on M's language/vocabulary proficiency at various stages, the opportunity was taken to look at how responses change with increased proficiency. In addition to these research questions, the *think-aloud* procedure is explored as a potentially useful addition to the current methodology.

Specific questions that this study addresses are:

1. Do M's general response characteristics change over time?
2. Are the profiles generated from M as reliable in the retests as in the initial tests?
3. Do M's responses to specific words change over time?
4. Is the think-aloud procedure a useful addition to the methodology?

8.5 Results

In this section the following are reported:

8.5.1 General response trends

8.5.2 Comparing profiles in the initial and follow up word association tests

8.5.3 Changes in responses to specific words

8.5.4 Think aloud data

8.5.1 General response trends.

In both sets of experiments there were 510 word association responses in total, half of these were generated from high frequency words and half from lower frequency responses. The response data, which was collected over a period of four years, consisted of responses within three major word classes; nouns, verbs and adjectives. There were initially 192 responses to noun stimuli, 186 responses to verb stimuli and

132 responses to adjective stimuli. In order to make the size of each word class data set equal, 60 response items were randomly selected from each of the word class data sets to use in the analysis. Given that in each of the separate experiments M generated a reasonably large sample of responses, this initial step in standardizing the data was not thought to compromise reliability. As shown in Table 8.2, after standardization 360 responses in total were used in the analysis, with 30 responses per sub-test.

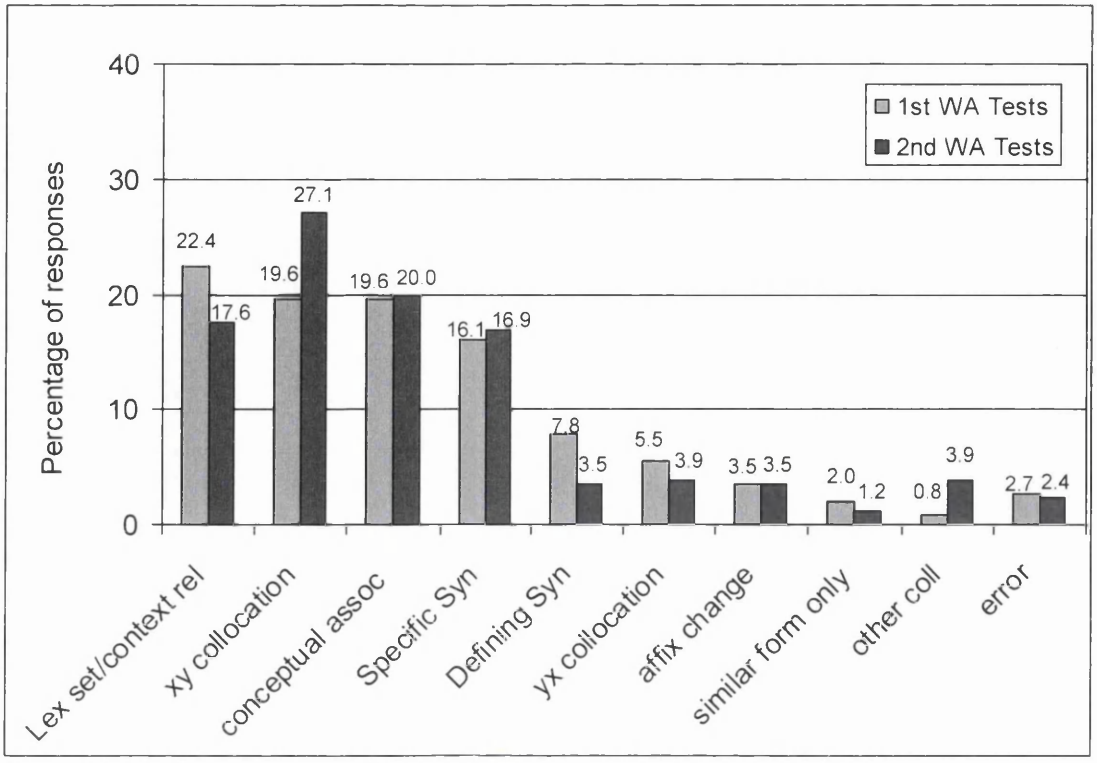
Table 8.2 The number of responses sampled from each sub-test

Word association test stimuli	Initial Test (test year)	Re-test (2012)
High frequency nouns	30 (2009)	30
Low frequency nouns	30 (2009)	30
High frequency verbs	30 (2010)	30
Low frequency verbs	30 (2010)	30
High frequency adjectives	30 (2011)	30
Low frequency adjectives	30 (2011)	30
Total number of responses	180	180
Combined total	360	

In Fig 8.1 the percentage of responses in each category in the initial word association tests can be seen to be very similar to the percentage of responses in the retests. This graph shows the combined responses (180 in each set) to the noun, verb and adjective stimuli. When the percentage of responses from the first set of tests were compared statistically with the percentage of responses from the second set of tests they were found to correlate very highly. A Pearson's correlation coefficient ($r = 0.93$) indicates a strong relationship between the responses given in the initial tests and the retests.

In both the initial tests (2009 – 2011) and the subsequent retests (2012) there was no overwhelmingly dominant group, with responses being spread between four groups: *specific synonym responses, lexical set/context related, conceptual associations* and *xy collocations*. M characteristically responds with Meaning-based associations, although *xy collocations* (Position-based) responses also feature. In both sets there were also a limited number of *defining synonym* responses.

Fig 8.1 Combined responses to the initial word association tests and the retests.



What is especially striking about these responses is that there were very few Form-based responses in either profile. Based on L2 studies (Söderman, 1993; Wolter 2001; Namei 2004) we might have expected M to make more Form-based responses in the initial tests, reasons for this pattern not emerging will be taken up in the discussion section.

If we unpack these data a little we can see that the general similarities still hold when they are broken down by word class (see Figs 8.2 - 8.4). In each category the initial test and retests for each word class correlate very highly. The noun stimuli for example (Fig 8.2) generate a large number of Meaning-based responses in both the 2009 test and also the 2012 test, the two profiles have a correlation coefficient of 0.78. In the other word classes the correlation between the two tests remains high although verbs and adjectives tend to generate more collocations. Of the three word classes analysed, the profiles created for the adjectives had the lowest correlation value (r= 0.70), which I would argue still shows quite a strong relationship between the profiles.

Fig 8.2 Responses to noun stimuli

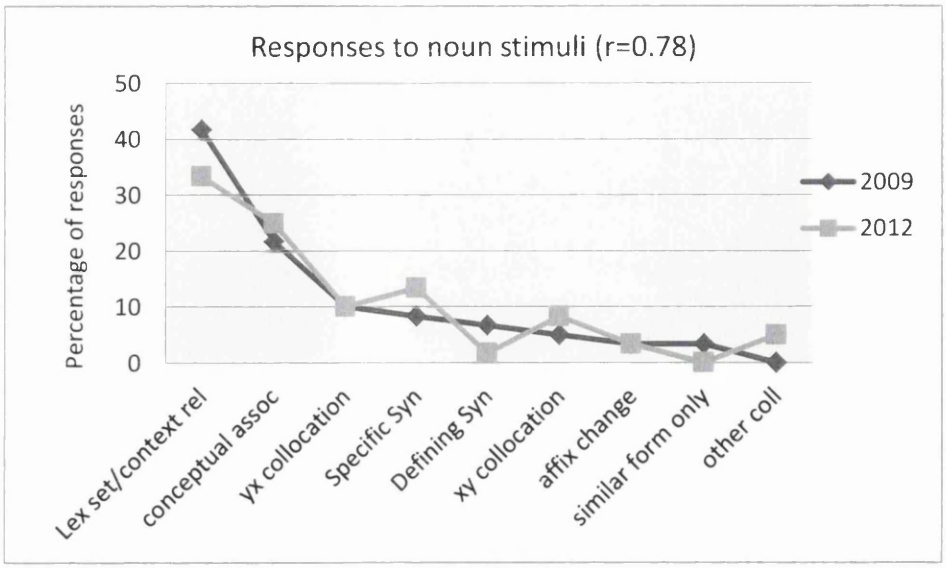
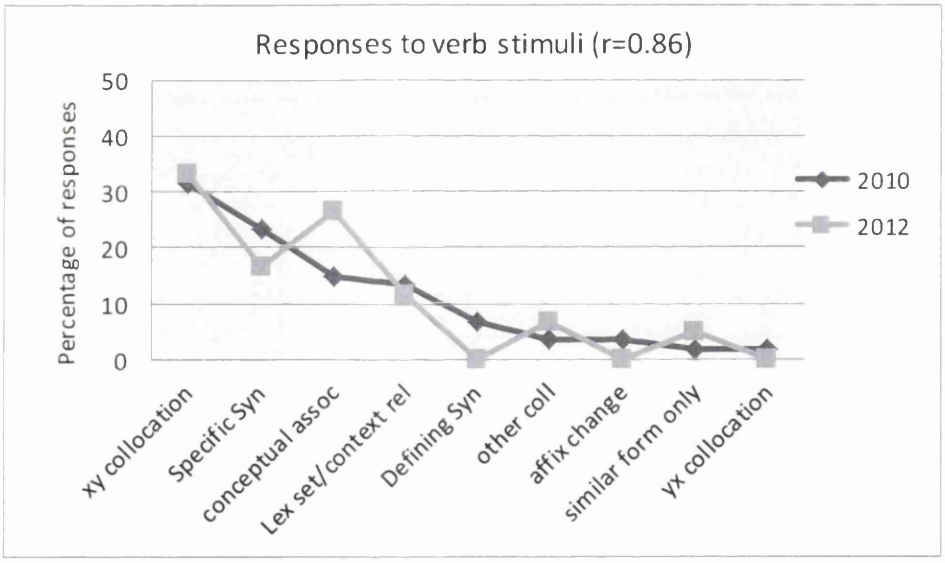
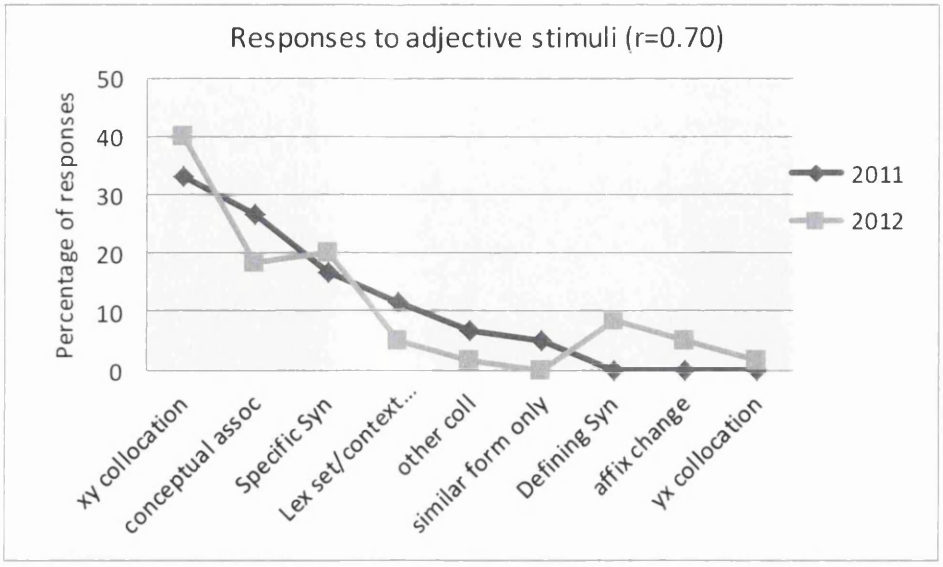


Fig 8.3 Responses to verb stimuli



Fg 8.4 Responses to adjective stimuli



Given that most (over 80%) of M's responses (see Fig 8.1) are spread between four categories (*specific synonym responses, lexical set/context related, conceptual associations* and *xy collocations*) these warrant a more detailed description. These can be viewed in Tables 8.3 – 8.5, the dominant category in each word class is in the top row.

Table 8.3 The percentage of dominant responses to noun stimuli

Response sub-category	**1st test (%Noun responses)**	**2nd test (%Noun responses)**
lexical set/context	**41.7**	**33.3**
conceptual assoc.	**21.7**	**25**
specific synonym	**10**	**10**
xy collocations	**8.3**	**13.3**
Total %	81.7	81.6

In Table 8.3 there is little variation in the proportion of nouns generated by each pair of tests, in both the initial test and the retest the *lexical set/context* category dominates, followed by *conceptual associations* then *specific synonyms* and in fourth place *xy collocations*. When given noun stimuli, it can be seen that in both tests this individual responds with a very similar pattern of responses. The findings show that the

dominant response characteristics to noun stimuli for this individual did not change in the three years between the initial test and the retest.

Table 8.4 The percentage of dominant responses to verb stimuli

Response sub-category	**1st test (%Verb responses)**	**2nd test (%Verb responses)**
xy collocations	**31.7**	**33.3**
specific synonym	**23.3**	**16.7**
conceptual assoc.	**15**	**26.6**
lexical set/context	**13.3**	**11.7**
Total %	83.3	88.3

In Table 8.4 (responses to verb stimuli) there is a little more variation between the percentage of responses in each category than in response to noun stimuli (Table 8.3). Again there is a clearly dominant category (*xy collocations*) although the second and third ranked categories (*specific synonym* and *conceptual associations*) in the initial test, switch position in the retest. The *lexical set/context related association* was fourth in both verb tests. With verb stimuli M characteristically responds with *xy collocations*. As with noun stimuli, the dominant response characteristic to verb stimuli did not change after a considerable time (two years) between testing.

Table 8.5 The percentage of dominant responses to adjective stimuli

Response sub-category	**1st test (% Adj responses)**	**2nd test (% Adj responses)**
xy collocations	**33.3**	**40**
conceptual assoc.	**26.7**	**18.3**
specific synonym	**16.7**	**20**
lexical set/context	**11.7**	**5**
Total %	88.4	83.3

In Table 8.5, adjectives can also be seen to generate a similar percentage of responses per category between the initial test and the retest. As with the verb stimuli the top category is again *xy collocations* and the second and third position (*conceptual associations* and *specific synonyms*) switch in the retest. With the fourth category there is a larger difference in the proportion of responses, although as these groups consist of less than ten responses not a lot can be read into this. With adjectives, M generally gives *xy collocations* and then a mixture of *conceptual associations* and *specific synonyms*.

The main point to take from these findings is that even though there is variation in the percentage of responses in each category at the two testing times the dominant four categories in each of the tests remain constant. The word class of the stimuli also has an effect on the type of response. With the noun stimuli M characteristically responds with associations that are mostly Meaning-based. With the verb and adjective stimuli M characteristically responds with *xy collocations*, Form-based associations. In answer to the initial research question: *No, M's general responses characteristics did not change over time.*

8.5.2 Comparing profiles in the initial and follow up word association tests

In this section we will compare the responses that the two tests in each word class generated. For each of the three word classes examined there are two graphs which show two profiles. The first graph in each set consists of responses to the high frequency stimuli at the two test times. The second graph in each set consists of responses to the less frequent words at two test times. While frequency is not being specifically examined in this experiment it was a variable in the earlier experiments; as the same word lists were used in the follow up tests the profiles in the graphs represent different frequency ranges. In Chapters 4 and 5 it was established that, for individual profiles, frequency does not seem to have an effect on individual response characteristics. Evidence from a study by de Groot (1989) also found frequency to have a negligible effect in word association tests. Despite this, it should be noted that frequency has traditionally been viewed as one of the key variables determining how associations are made between words (Deese, 1965; Cramer, 1968). More recently, Schmitt (2010:13) states that "frequency is one of the most important characteristics of vocabulary, affecting most aspects of lexical processing and acquisition", it would therefore seem premature at this stage to discount it entirely. With regards to the

evidence provided in this thesis so far though, it seems that the role of frequency in word association testing may have been overstated.

The first set of graphs (Figs 8.5 & 8.6) show M's profiles created from responses to the same noun stimuli roughly three years apart. The next set (Figs 8.7 & 8.8) show her responses to the verb stimuli (two years apart) with the most recent set (Figs 8.9 & 8.10) showing responses to adjectives (one year apart). In these graphs the response categories have been ordered so that the dominant categories are to the left.

Fig 8.5 Responses to high frequency nouns

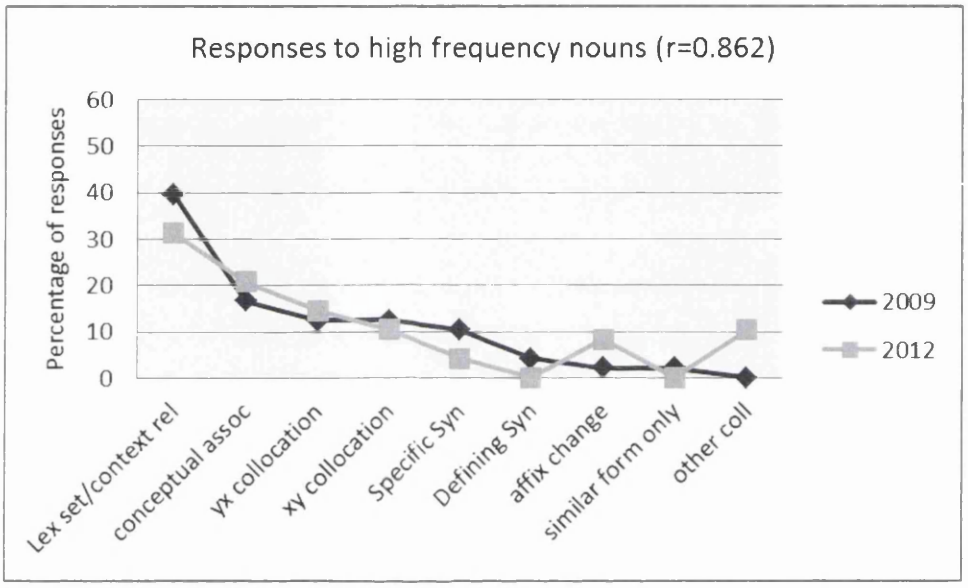
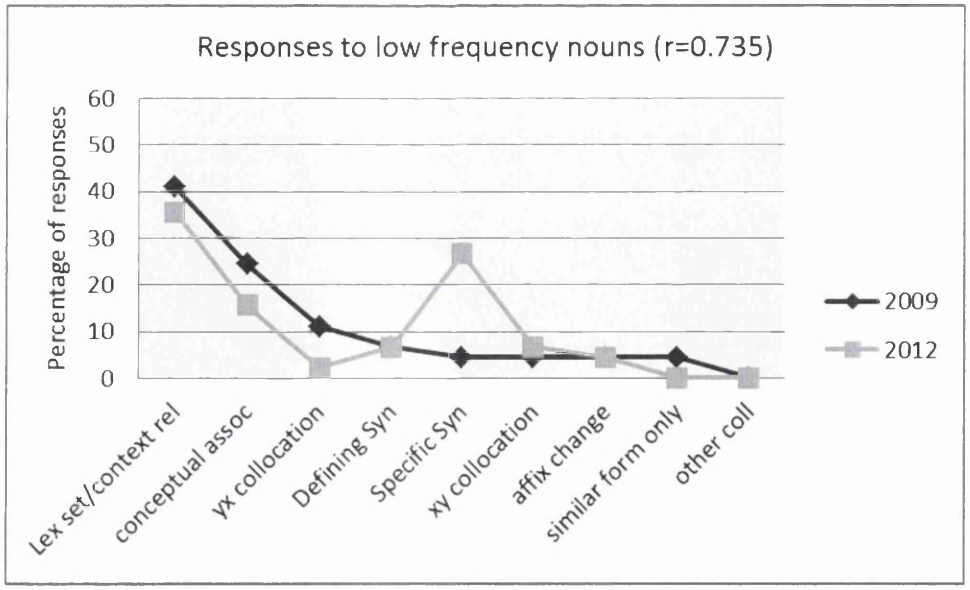


Fig 8.6 Responses to low frequency nouns



With the responses to noun stimuli the profiles from the higher frequency stimuli (Fig 8.5) correlate highly (r=0.862), the profiles from the lower frequency stimuli (Fig 8.6)

show a more moderate correlation (0.735). Despite a difference of three years between these tests the responses in each category correlate strongly. While both sets of data have high correlations there is more variation with the lower frequency stimulus. Or put another way, responses to rarer nouns are less stable.

Fig 8.7 Responses to high frequency verbs

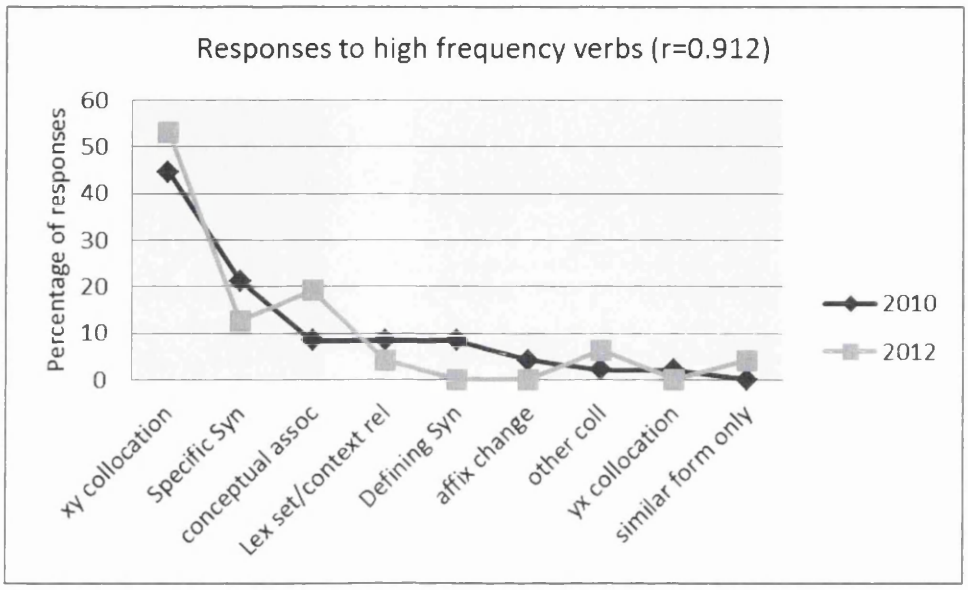
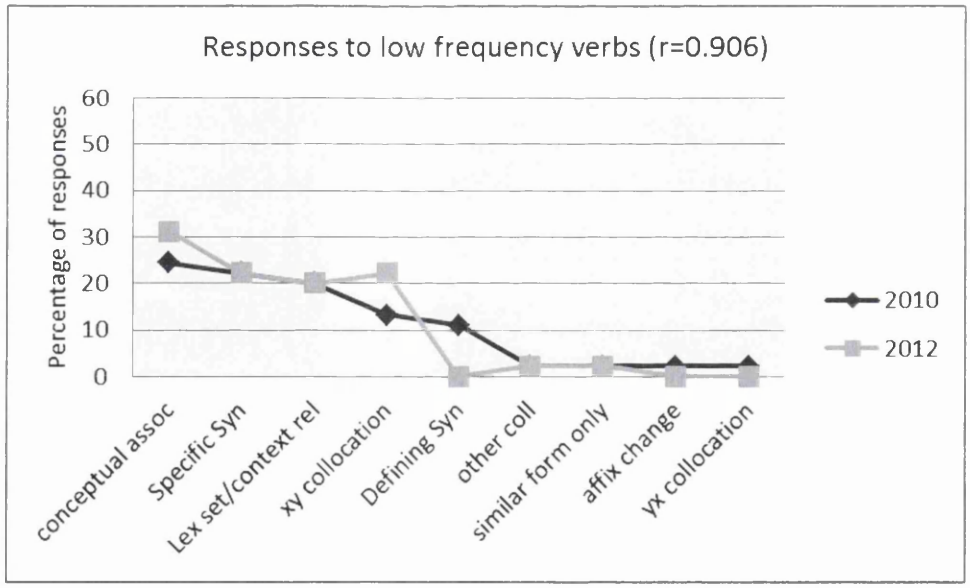


Fig 8.8 Responses to low frequency verbs



As with the nouns the responses generated from verb stimuli to both high (Fig 8.7) and low (Fig 8.8) frequency items correlate highly, with a slightly higher correlation observed for the higher frequency items. With a two year gap between these tests the number of responses in each category is quite similar in both sets of data. Unlike the noun stimuli (Figs 8.5 & 8.6) that had two dominant categories with both frequency

groups, the dominant category for the high frequency items (xy collocations) was quite different for the lower frequency items (conceptual associations). There is a change in the type of responses that M makes when given high frequency or low frequency verb stimuli.

Fig 8.9 Responses to high frequency adjectives

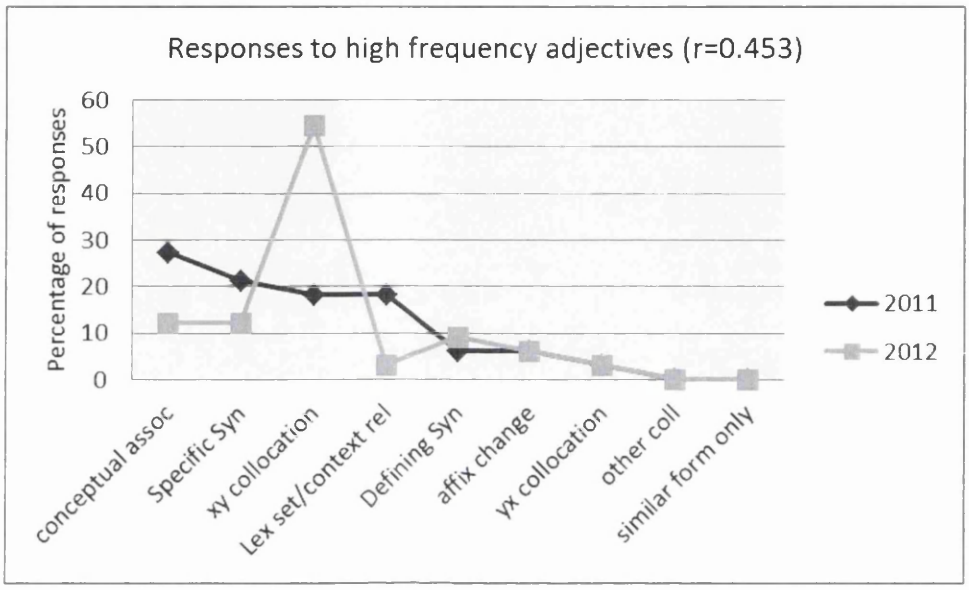
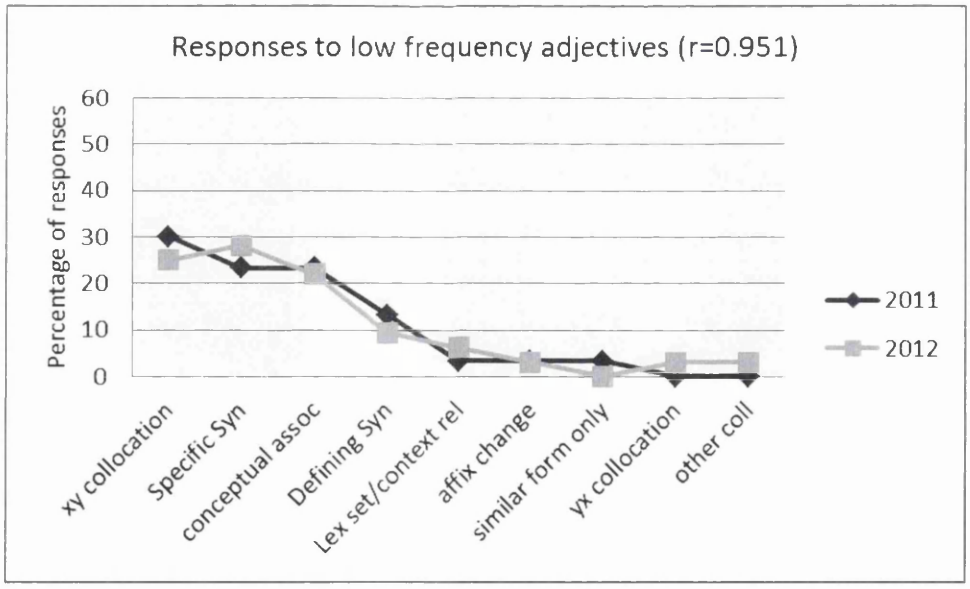


Fig 8.10 Responses to low frequency adjectives



In contrast to the noun and verb data sets, which showed high correlations with responses to both high and low frequency stimuli, the correlations between responses to adjectives are more difficult to interpret. The higher frequency items (Fig 8.9) had a weak relationship whereas the lower frequency items (Fig 8.10) correlated highly.

This goes against the pattern established for nouns and verbs (that responses to lower frequency items are less stable). With the higher frequency adjectives (Fig 8.9) it is the *xy collocation* category that varies the most, from six collocational responses in 2011 to 18 collocations in 2012. One reason for the adjective set displaying response behaviour that does not fit the patterns observed with the noun and verb responses, may be due to the shorter interval between testing. This and other possible explanations will be taken up in the discussion section.

In this section it has been demonstrated that the three word classes generally show high levels of reliability. Correlations at the two test times between profiles generated from the same stimulus words were generally very high. In answer to the second research question: *Yes, the profiles M generated were as reliable in the retest as they were in the initial test.* An interesting observation from the noun and verb response data is that the higher frequency stimuli seem to generate more stable responses than the lower frequency stimuli.

8.5.3 Changes in responses to specific words

So far we have seen that in both sets of word association tests the response patterns are generally consistent and exhibit many similar characteristics; this indicates that the basic approach is reliable. The next step in our analysis is to look at the data in more detail, comparing how this individual responds to specific words with a large gap between tests. The third question that we set out to answer in this study was: *Do responses to specific words change?* The short answer to this is *Yes*. Table 8.6 shows that on average only 15% of responses in the retests were exactly the same as in the initial tests. This will be dealt with in more detail in the discussion section.

Table 8.6 The percentage of 'same' responses

noun stimuli	verb stimuli	adjective stimuli	Mean
16 %	17%	12%	15%

8.5.4 Think aloud data

As mentioned in the introduction, one benefit of analysing single subjects is that additional insights can be gained from using techniques that are unsuitable with large group studies. One such technique, think-aloud, is time consuming although

potentially generates a large amount of qualitative data. Attempting to collect and process such data for a large group would be problematic; for an individual case study however this issue is less pressing.

Despite hopes for greater insights into how the individual was thinking while she made the associations, the findings from the think-aloud method were disappointing. During the tests the participant generally repeated the stimulus word and then said the response word, without further elaboration. On the few occasions when she did elaborate it was usually due to her being unsure of the spelling of the response word. Also, she sometimes translated the stimulus word into her L1 before responding. Although this last point indicates a further step available to learners when responding in their L2 (translating back and forth between their L1 and L2) the more interesting details of why she made particular links with words were not commented on. In short very little new information was gained from concurrent think-aloud that could not be deduced from the written responses alone. The answer to the fourth research question is therefore negative.

8.6 Discussion

An important point to come out of this study is the confirmation that the word association response characteristics of M are (in line with Fitzpatrick's 2007 findings) consistent. When replicated, the responses to three word association tests showed similarly high levels of internal reliability. While we may not wish to generalise too much from one individual, when put into the context of the previous chapters, the evidence in favour of the 'individual profiling' approach to analysing word associations seems to be growing. In this section we will focus on some questions that arise from the responses that this individual made.

Why did M's responses not become more "native-like" as her proficiency increased?

Why were there so few Form-based responses with lower frequency stimuli?

What can be inferred from the responses that were exactly the same in both the initial test and retest?

Hopefully by addressing these questions we can shed some light on the mixed findings of earlier studies indicated in Chapter 2. We will also consider the think-aloud procedure. By trying to understanding why this technique failed, perhaps we can get a better understanding of how word associations work.

8.6.1 Responding to questions raised by this study

The findings raise some interesting points. We might question why M's responses did not become more "native-like" as her proficiency increased. Studies such as Politzer (1978) and Söderman (1993b) would lead us to expect a higher proportion of Form and Position-based responses in the initial test than in the second retest. This was not found to be the case. Putting aside the problems identified in Chapter 2 with these two papers it is useful in helping us understand the mechanics of word associations to consider why 'native-like' norms were not generated. I think the main point is that learners already have at least one language in their lexicon before they start learning another, this is a crucial difference. They are not building a lexicon from scratch as children do in their native language but adding to an existing structure, often set within quite a different cultural background. Not accounting for this was a fundamental flaw in early L2 word association studies. Meara notes:

> *Teaching a language aims to produce people who are bilingual, not mere replicas of monolingual speakers. It would therefore be more appropriate to compare the associations of learners with those of successful bilingual speakers, and not native speakers.* Meara (1983:31)

More recent studies (Henriksen, 2008; Zareva, 2011) have taken this idea on board, comparing learner responses with other similar learners. Fitzpatrick (2009) for example studied how the responses of individuals in their L1 compared with responses in their L2. Such studies are I believe more valid since they are comparing like with like.

It might also be noted that the notion of the 'native speaker' as a benchmark against which to compare learners has been called into question, in a more general sense, due to the considerable variation in knowledge and skills that native groups have been found to have. In a study that looked at the affect of age, educational level and profession on the language ability of Dutch native speakers Mulder & Hulstijn (2011:491) conclude that "there are substantial differences among native speakers both in linguistic subskills and in speaking proficiency…it is impossible to define the prototypical native speaker in terms of language ability." If as they hold, the concept of "native levels of proficiency" is not as clear cut as previously assumed then this detracts even more from studies that argue learners word association responses move towards native-like ability. Despite these concerns, there are however some findings in the present study that concur with earlier studies. L1 studies such as Deese (1965)

and also L2 studies such as Söderman (1993) argue that phonological (Form-based) aspects of a word are acquired early, Form-based responses they argue are indicative of acquisition at a very basic level. If we accept this then we would expect more Form-based responses with the lower frequency words, this was partially found to be the case. When we look at the general data we can see that on the whole M rarely made Form-based responses, but when she did make Form-based responses a more detailed analysis shows they were usually with words that were either unknown or partially known. The fact that there were a low overall number of Form-based responses seems to be due to M's ability, most of the stimulus words used were fairly familiar to her. Closer inspections of the data (particularly insights from the retrospective interview) reveal she was familiar with the high frequency words and also familiar with many of the lower frequency items. Of the few items that were identified as being unknown or partially known there is evidence that the form of the word had a role. For example, with the stimulus *distinction* in the 2012 study M's response was *exterminate*, when questioned on this it turned out that she was unsure of the meaning and had made this link because "it sounded like extinction". In this case she made an association in two steps, moving from the stimulus *distinction* to *extinction* and then to *exterminate*. The first of these steps is a phonological link, the second step being more characteristically (for her) linked to the meaning of the word. Even though her comments indicated that she linked *extinction* and *exterminate* due to their meaning it ought also to be noted that they both look similar, they are long words (9 and 11 letters) and begin with the prefix *ex*. It seems likely that both form and meaning of the word contributed to the response, even when M was asked specifically about this it was not easy to untangle. How she thought she responded and what might have been occurring subconsciously was not so clear. The rater needs to be aware that retrospective interviews do not always give an accurate account of cognitive processes, people do not always know why they respond as they do. When participants try to reconstruct their own thoughts they are not always correct and may even give misleading information (telling the rater what they think is the 'correct answer' or what they think the rater wants to hear). Curiously, in the 2009 test M responded to *distinction* with *remarkable* and, when asked, demonstrated that she (at least partially) knew what it meant. This item therefore appears to have been partially acquired in 2009 however by 2012 it had largely been forgotten. Instability is of course to be expected with partially known words. Sometimes they generate Meaning-

based associations, sometimes Form-based associations and sometimes a meaningful connection cannot be made. In another example in the 2009 study, she responded with *phrase* to the stimulus *phase* as they sound and look similar. In the 2012 study she responded to this item in a similar way, she gave the phonological response *fade*. Additional questioning revealed the reason for this continued Form-based response is that this item was only partially known in 2009 and then after three years she still didn't really know what it meant. As with the previous example, this illustrates the "incremental" nature of word acquisition that Schmitt repeatedly stresses (Schmitt, 1998a, 2010). Even though *phase* is not a particularly low frequency word she made little progress with it over three years. It would seem that the reason for the lack of Form-based responses in general is that most of the stimulus words in this study were (for this learner) fairly well-known lexical items. If a set of stimulus words had been selected from a lower frequency range then probably more of these Form-based responses would have appeared in the initial responses. Based on the VLT information (Table 8.1) we might speculate that for M words within the 5000 - 6000 most frequent range would contain many more peripheral items.

As noted in the results section (Table 8.6), 'same' responses account for around 15% of the responses, for example in response to the stimulus *circle* the individual responded with *triangle* on both the initial and follow up tests. Another example is *mechanism* in response to *machine,* these responses are not predicted by native norms lists. EAT (The Edinburgh Associative Thesaurus, Kiss et al 1973) gives *round* as the primary response to *circle* and *tool* as the primary response to *machine.* The EAT norms list also rank the stimulus/response of *circle→triangle* as joint tenth, along with 21 other responses which only had one hit. The stimulus/response of *machine→mechanism* was not listed in the norms list. The fact that this student appears to know these words well yet repeatedly responds in an un-native way adds weight to the argument that we should not be comparing L2 responses against native norms but evaluating responses in terms of the individual. Given that M actually made considerable progress in her language studies over the three year period (Table 8.1) it is perhaps surprising that any of the responses were the same. There are a number of ways to interpret this. One interpretation would be that in the intervening three years these words were not met enough times for additional aspects of word knowledge to be added to her lexicon. As none of these words are of a very low frequency, lack of input does not however seem a likely explanation. My

interpretation of this would be that these words were already well integrated into the lexicon prior to the initial test, with most of the aspects of word knowledge (Nation, 2001) for these items being acquired. Therefore despite meeting the word a number of times before the retest no new aspects of this word were acquired and thus the response could not change. Even if M did begin to develop new levels of understanding for these words within the three years, a prior association might still dominate. Until there is overwhelming support for a new type of association there would seem no real reason for a prior association to be replaced: unless the initial association was erroneous. It ought to noted that such speculation is largely based on notes taken in the retrospective interviews, it would however seem preferable to triangulate this with more objective criteria. Future studies of this kind would therefore benefit from a formal measure of how well each word was known at each stage, such as Wesche & Paribakht's Vocabulary Knowledge Scale (1996).

As has been noted already, responses that were exactly the same in both the initial tests and retests were not common. What usually happened was that the responses in the two tests were not the same but were in the same response type category. For example, in response to the adjective *terrible* in the initial test the response was *miserable* (a synonym), in the retest the response was *bad* (another synonym). In another example the responses *church* (initial) and *christian* (retest) were given to *religion*, in both cases they were judged as being in the same lexical set. Again with these examples, it might be noted that the responses are not the most stereotypical responses given by native speakers: we would not be able to predict these responses from native norms lists such as EAT.

8.6.2 Rethinking the *think-aloud* procedure

A disappointing outcome of this experiment was that the concurrent think-aloud methodology generated so little useful data. With hindsight it is possible to think of a number of reasons why this might be so. This could have been due to insufficient training in the technique or that think-aloud does not suit this particular individual. However, in practice tasks before the retests (thinking aloud while writing a shopping list and making a cup of tea) the participant seemed suitably verbose. A more likely explanation is that the word association task itself is not suited to this kind of activity. Given that the mind is often thought of as having limited resources (Barcroft, 2002) one explanation could be that requiring a learner to verbalise at the same time as

making a word association is too demanding. While possible I don't think association tasks require high levels of conscious processing, particularly when most of the stimulus words are fairly well-known, so it is unsatisfactory to attribute the lack of concurrent verbalisations to task difficulty. On the contrary I think it is more likely that many word associations are automatic (subconscious) reactions, so when asked to verbally describe these thought processes many people cannot. This idea, and theoretical explanations supporting it, will be developed further in the discussion chapter.

8.6.3 Intervals between testing

In the results section it was noted that M's responses to high frequency adjectives were the most unreliable in this study. As noted in Chapter 7 this may well be due to the nature of adjectives in general or that, due to the rigorous selection procedure, the adjectives chosen for these studies were not representative of 'typical' adjectives. There is however another possibility. The gap between the initial test and second test for each of word class tests varied from one to three years. The main implication of this is that nouns had a much bigger gap between test times than adjectives, M would have therefore had far more opportunities to meet each noun through incidental exposure. We might therefore expect M's knowledge of the nouns and verbs to have been better acquired (and therefore more stable) in the time available than the adjectives. Another point to bear in mind is that it is conceivable (though unlikely) that the more recent tests interfered in some way with responses given in the retests. It would therefore have been better to have kept the time interval between each pair of tests constant: around two years would seem suitable. A weakness of this study is the lack of planning to ensure a more regular interval between tests. To remove this possible variable, a more consistent testing schedule is therefore recommended for future research projects of this kind.

8.7 Summary

The main finding is that M responds in a consistent way to word associations. Her responses can be viewed as idiosyncratic in that they are not predicted by native norms. This supports the findings of Fitzpatrick (2007) and previous chapters that learner responses are neither homogenous nor native-like, although internally reliable. Based on the general response data, M can be said to have responded in a similar way

to both sets of word association tests. When we break this data down into the separate word classes we also see that, despite some variation between the word classes, the type of responses within each word class are similar in number and correlate highly. Another important finding is that M's basic response characteristics do not change over time. While there was variation in responses and clearly some development of word meaning between the initial test and retest the general pattern of response type did not really change. This finding does not sit well with the idea of a syntagmatic – paradigmatic shift. Despite an increase in M's general language (and also vocabulary) ability between the first and second set of tests, there was no evidence for a 'shift' in how words are organised in the lexicon. The retest did not for example result in a higher proportion of Meaning-based responses. However, as the responses were given to words that M usually knew quite well (even in the initial tests) the evidence against a shift in response with increased proficiency cannot be strongly stated.

Finally, it ought to be noted that the think-aloud procedure did not enhance the methodology and is therefore not recommended with further word association studies. The partial-retrospective interview seems a better way to confirm classifications.

8.8 Conclusions

While this single subject study has helped to establish the individual profiling approach as a reliable way of analysing word association responses, due to the post-hoc nature of the research design there were a number of areas that were inadequate. This stems from the reuse of data for a purpose slightly different to the one originally intended. In the initial studies for example there were no specific aims to track proficiency changes over time. Some proficiency data were collected, although these were primarily intended to help with the selection of stimulus items. Consequently the data on general language (and specific vocabulary) ability were not as thorough as they might have been. This study would have benefited from an objective measure of depth of word knowledge. A measure such as Wesche & Paribakht's Vocabulary Knowledge Scale (1996) might have enabled more precise judgments to be made on how well each word was known at each stage. The lack of such a measure limits the confidence that can be put in statements regarding the effect of proficiency on this individual's response profile. Another criticism is that this study only really explores responses to words that are well-known. This limits the generalisations that can be

made about peripheral items in this student's lexicon, words that have been newly acquired. To measure the behaviour of partially known words less frequent items would need to be selected, for M the 5000-6000 range would seem suitable. I would speculate that were this to be done then Form-based responses would feature more heavily. It is also likely that response profiles would become less reliable with even lower frequency words, some evidence for increased instability with the lower frequency nouns and verbs used in this study having been observed.

Having established the reliability of the basic method and approach to analysis, we can now turn to the question of how we might apply the main findings. It would be going too far to suggest that word association tests are precise enough to uniquely identify learners in the same way as the "lexical signatures" derived by Meara et al. (2002) from learner's written work. The findings from this study do however point in the same direction, that "L2 learners are far from uniform in their lexical choices". These response characteristics presumably relate to the unique set of experiences and background that everyone has. If we take this a step further we might hypothesize that every learner is predisposed to a particular way of acquiring vocabulary. The data also seems to suggest that each learner's characteristic predisposition is influenced by the word class of the stimuli. If this is the case then the current one-size-fits-all approach to studying vocabulary that is adopted in many classrooms may not be the most effective. Along with other points of interest, that go beyond this particular study, the potential pedagogical application of these findings will be taken up in the subsequent discussion chapter.

Chapter Nine: General discussion

9.1 Introduction

In the conclusion to the literature review we could argue that due to a series of methodological problems, the potential of word associations to answer questions about the mental lexicon had not been realised. Following yet more inconsistent results from a replication of Wolter's 2001 study (Chapter 3) an alternative approach, developed by Fitzpatrick (2006 & 2007) was adopted that promised more reliable word association data. Using her 'individual profiling' approach to data analysis and applying careful control over the stimuli it was anticipated that more reliable data could be obtained. If this could be achieved then it was argued that word association tests could be used as a measure of the organisation of the mental lexicon. Specific problems identified in Chapter 2, that researchers had not satisfactorily agreed on were: how to select stimulus words (and how many), how to classify responses and how best to analyse them. In this chapter I will revisit these issues based on what I have learned from the series of word association studies that are reported in this thesis. Other issues that arose out of the studies were the automaticity of word associations and also the potential pedagogical applications of the findings. To help advance this promising line of research further, areas that have yet to receive attention or would benefit from a more detailed treatment will be pointed out. Before looking at these specific areas though, it seems logical to review the main findings of the experimental chapters.

9.2 General review of findings

The main finding from the series of experiments reported in Chapters 4 - 8 is that Fitzpatrick's individual profiling approach can generate reliable responses profiles. Working within this framework it was found that:

- *The frequency of the stimulus word had little effect on the reliability of responses.*
- *The word class of a stimulus word had little effect on the reliability of responses.*

These findings are important as they demonstrate that word association tests are capable of generating reliable responses, a point that has been called into question (Kruse et al., 1987). Confirming the reliability of this approach represents a step forward as we can now confidently use it as a way to measure the organisational dimension of the mental lexicon. As argued in the introductory chapter, a reliable

measure of vocabulary organisation complements the measures of vocabulary size that already exist. Measuring these two 'global characteristics' will enable researchers, and teachers, to more fully understand the vocabulary competence of second language learners.

The first of the two claims made in the previous paragraph, that stimulus frequency does not appear to have an effect on responses, was a surprising finding. It has long been assumed that frequency has some kind of effect, and is usually accounted for in word association studies. The *stimulus frequency has little effect* claim cannot be stated too strongly though as the frequency bands tested (0 -500 frequency band, 500 – 1000 frequency band, 1500 - 2000 frequency band) are all within what would usually be classed as the 'high frequency' range. As words in much lower frequency bands were not included in the experiments all that can be said with confidence is that: there was no evidence that frequency affected responses to stimuli from bands within the most frequent 2000 words in English. Also, in Chapter 8 there were slightly lower correlations between the responses to the lower frequency nouns and verbs at the two test times, hinting at increased instability with even lower frequent words. Clearly, there is still work to be done with lower frequency stimuli. It may well be the case that the reliability of individual profiling decreases with stimuli from lower frequency bands than tested in this thesis. As argued in Chapter 8, a good start might be made with testing the reliability of responses to stimuli drawn from the 5000 - 6000 frequency band.

The second of the two claims, *the word class of the stimulus has little effect on reliability*, does not mean that word class has no affect at all. A finding that came out of the word class studies (Chapters 4 – 7) and was confirmed by the detailed case study (Chapter 8) was that responses to stimuli from particular word classes are biased toward particular kinds of responses:

Nouns tend to generate *lexical set/context related* and *conceptual* responses

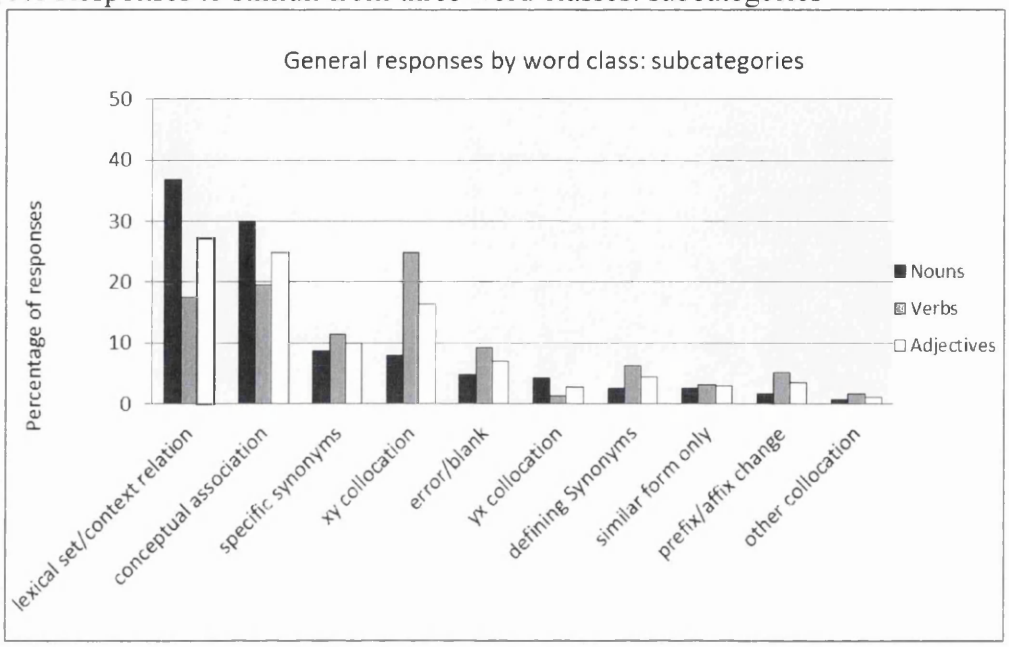
Verbs tend to generate *xy collocations* responses

Adjectives tend to generate a mix of *lexical set/context related, conceptual responses* and *xy collocation* responses.

The word class bias can be seen in Fig 9.1, which shows the percentage of responses in each subcategory for the three word class studies. The data from each of the studies has already been presented in chapters 5, 6 and 7 although it is useful to bring them together into one graph in order to see the big picture. As with the graphs in the

experimental chapters the categories have been ordered so that the dominant categories are to the left with the less used categories to the right.

Fig 9.1 Responses to stimuli from three word classes: subcategories



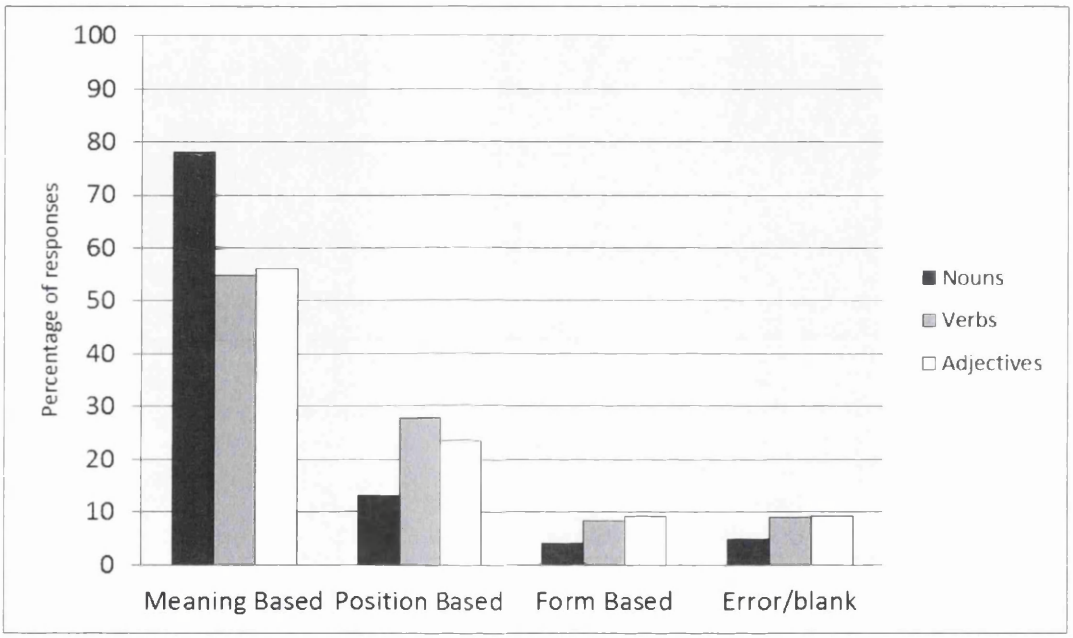
While L2 word association studies (Fitzpatrick, 2006; Zareva, 2011) routinely use stimulus lists with a variety of word classes, based on the assumption that word class has an effect, few have actually tested this assumption. One study that did attempt to measure the effect of word class on responses was Bagger-Nissen & Henriksen (2006). In that study they found that nouns generated a disproportionate number of paradigmatic responses with Danish speakers (in both their L1 and L2) and that verbs and adjectives generated disproportionate numbers of syntagmatic responses. While the data presented in Fig.9.1 doesn't contradict this, as argued earlier (2.10) due to the small number and poor choice of stimulus words their claims are not well supported. In this thesis, as each word class was treated separately in a series of experiments, far more stimuli per word class were included. Although a few words from each word list did not make it to the analysis stage, learners in the three studies were given 96 nouns, 96 verbs and 66 adjectives to respond to. There is therefore considerably more support for a word class effect than in Bagger-Nissen & Henriksen (2006) that only included 15 words per word class. In Fig 9.1, we can see that these 87 learners (Noun study, 30 students; Verb study, 27; Adjective study, 30) tended to give a lot of responses in three particular categories: *lexical set/context relation, conceptual associations, specific synonyms*. These are all Meaning-based categories, which is not what we

might have predicted from the results of some studies (Politzer 1978; Piper & Leicester, 1980; Söderman, 1993). These results do however broadly agree with the findings of Fitzpatrick (2006): in that study the predominant NNS responses were Meaning-based with few responses categorised as Form or Position-based. Of the Position-based responses it was also the *xy collocation* subcategory that dominated in both Fitzpatrick's and the experiments in this thesis. Interestingly, of the two kinds of synonym that are distinguished, the learners in this thesis made a lot of *specific synonyms* but did not make many *definite synonyms*. This is the opposite of Fitzpatrick's 2006 findings. It ought to be noted that the learners in Fitzpatrick's 2006 study were of similar ability (averaging an IELTS score of 6.6) to the learners in this study (students had TOEIC scores between 650 – 800, equivalent to IELTS 6.5 - 7). The contrary findings therefore cannot be put down to student ability, even though we might well expect higher ability students to give these more precise definitions. The main difference between the 2006 study and the studies in this thesis is the type of stimulus words used. Fitzpatrick's study used stimulus words derived from the academic word list (Coxhead, 2000) whereas the stimuli in this thesis were all within the most frequent 2000: fairly common words. A likely explanation is that academic words have a more precise meaning and are more carefully defined in written texts: it is therefore easier to give a defining synonym to these words. Another possibility, covered in the section on 'classification problems', is that these categories overlap. From the learners' perspective, it could be argued that they measure a similar concept.

A point that was commented on in previous chapters was that when these sub groups are rolled up into the three main categories (Fig 9.2) a lot of information becomes obscured. As Fitzpatrick's three main categories (*Meaning, Position and Form based*) broadly agree with the traditional *paradigmatic, syntagmatic, clang* categories that L1 and also many L2 studies used, it does however allow for some comparison. When the results of the word class studies (Fig 9.2) are compared Meaning-based responses dominate, this opposes the idea that learners generally give syntagmatic responses (Piper & Leicester, 1980; Söderman, 1993). Within the Meaning-based category there is clearly a bias for noun stimuli. To confirm this observation, three one-way ANOVAs were conducted on each of the main response categories. In the Meaning-based category the between group variance was highly significant $F(2,166) = 125.95$ $p<0.001$. Tukey's post-hoc test indicated that for all pairwise comparisons there was a significant difference ($p<0.01$). In the Position-

based category the between group variance was also significant F (2,166) = 15.47 $p<0.001$. The post-hoc test indicated that there were significant differences ($p<0.01$) between the responses in the Noun and Verb studies and also between the Noun and Adjective studies. In the Form-based category there was less between group variance, F (2,166) = 5.59 p0.0045. The post-hoc test indicated a significant ($p<0.01$) difference between the responses in the Noun and Verb studies and also between the Noun and Adjectives studies. In all the main categories there was a statistically significant word class effect.

Fig 9.2 Responses to stimuli from three word classes: main categories



Piper & Leicester's 1980 study is similar to the experiments in this thesis in that it used Japanese learners and stimuli from the same three word classes. They found that in general, both beginner and advanced Japanese learners responded syntagmatically, particularly with verb and adjective stimuli. From their findings (Table 9.1) they argue that with increased proficiency the responses begin to resemble the native group, which has the highest proportion of paradigmatic responses.

Table 9.1 Mean Proportion of paradigmatic responses (Piper & Leicester, 1980)

	Nouns	Verbs	Adjectives
Native	.62	.41	.28
Japanese ESL Advanced	.66	.34	.25
Japanese ESL Beginners	.64	.25	.16

As P&L's study exhibits many of the methodological problems highlighted in Chapter

2 though, not a lot can be read into these findings. There is a serious problem with the stimuli, not only in terms of quantity (only eight per word class) but also quality (high frequency words derived from the Kent-Rosanoff list, 1910). Another problem is the crude classification method. In P&L's study, responses were classified as paradigmatic if they were in the same word class; all other responses were assumed to be syntagmatic. The phonetic or erroneous responses were presumably lumped together with what I would term as the genuine syntagmatic responses: such as collocations. As noted in the discussion of the replication experiment (3.5.1) phonological and erroneous responses are common with learners. This means that P&L's study does not accurately show the proportion of syntagmatic responses. With flawed studies of this sort it is not so hard to explain away contradictory findings. However, against studies that use larger and more valid stimulus lists and a more carefully considered classification system (Söderman, 1993), the inconsistent findings are more difficult to understand. In Söderman's study, the main finding was that lower level learners gave more syntagmatic responses and as their proficiency increased so did the proportion of paradigmatic responses. When viewed as a group, what can be seen from the Japanese learner data in this thesis is that Meaning-based/paradigmatic responses dominate. As I have already commented on, the lack of consistent findings between studies was the main reason for abandoning the traditional 'group' approach in favour of 'individual profiling'. In this thesis, it is argued that consistency between group studies cannot be expected as the individuals within the groups do not respond homogenously.

Idiosyncrasy within the group, rather than homogeneity, was observed in all the experiments in this thesis. Even within Japan, a society frequently stereotyped as having a strong sense of commonality, when it comes to word association responses there is a lot of variation. Although we have no firm evidence as to why the organisation of English within Japanese learner's minds is not homongenous, we can speculate. If we consider the conditions under which we might expect homogeneity perhaps we can edge closer to an understanding. For a homogenous associative network, students within a particular group would need to have a similar amount and type of L2 input in both their formal education and everyday life. Currently, neither formal study nor incidental exposure to English is particularly uniform. As a Japanese learner of English passes through the school system they have a variety of formal and informal learning experiences. This is dependent on which school they go to, if their

parents are keen for them to study English at a cram school and the amount and type of other L2 input. Some formally study English as a foreign language on a weekly basis from early childhood whereas others are not exposed to it until it becomes a mandatory school subject at 12. With the widespread availability of the internet, the typical Japanese learner now has the potential to access a wide range of L2 written and audio media. This means that learner's backgrounds, in terms of the amount and type of L2 input they receive, vary enormously. The organisation of the mental lexicon observed in this thesis appears to reflect the unique set of L2 experiences that learners have these days. Considering their backgrounds, the variation observed is not so surprising. It also supports the argument in this thesis that it is more valid to analyse the response data is in terms of the individual rather than the group.

Even though the Japanese learners in this thesis cannot be considered a homogenous group, they did exhibit a high degree of internal reliability. Table 9.2 shows that in three word class studies (Chapters 5, 6 & 7) there were high within-subject correlations. Of the 87 learners in these three studies (the same students as in Figs 9.1 and 9.2) 51 of them (59%) had proximity correlations of >0.70. There is good support for the claim that:

L2 learners generally give reliable profiles to words: irrespective of word class

Table 9.2 Proximity rankings for profiles in three word class studies

Correlation coefficient (r)	Definition of profile proximity	Noun 2 (n=30)	Verb (n=27)	Adjective (n=27)
>0.8**	very close	**12**	**6**	**11**
0.7 – 0.8**	close	**8**	**7**	**7**
0.6 – 0.7*	vaguely similar	**5**	**4**	**7**
<0.6	dissimilar	**5**	**10**	**2**

**p = <0.001, *p = <0.05

Having explored three of the major word classes it can be stated that word class does not have a particularly strong effect on the reliability of an individual learner's characteristic response profile. When an individual is given two sets of words from different word classes, or from different frequency ranges, the responses that the individual gives usually correlate highly. Although reliable findings can be obtained irrespective of the word class or frequency of the stimuli, word class does need to be accounted for due to response bias. There are options available for those interested in using WATs as a method of probing the mental lexicon. If the stimulus

list is restricted to a single word class then the response bias for that particular word class ought to be expected. Another option, a mixed class list, does however seem more useful if the aim is to obtain a profile that is representative of a person's general response characteristics in a language. In the discussion section of the Verb study (6.7.4) it was suggested that if this were the aim then a mixed list ought to reflect the percentage of words in each word class. For English that would be around 51% nouns, 20% verbs, 17% adjectives and 7% adverbs and 5% from other classes; although as noted in that section, these values vary depending on the frequency of the words.

As well as the main findings concerning the effects of word class and frequency on word association responses this series of experiments resulted in some refinements to the methodology. The use of a partial-retrospective interview in Chapters 4 -8 improved classification accuracy while adding only a little extra time to the data collection procedure. A full retrospective interview (Wolter, 2001; Fitzpatrick, 2006; Chapter 3) was deemed inefficient due to the considerable time needed and a realisation that many responses are unproblematic to classify. In the partial-retrospective interview students were only asked to give further information on responses that after a cursory inspection seemed ambiguous. The need for some kind of retrospective check was supported by a finding in the Noun 2 study (Chapter 5) that rater intuition had an error rate of around 11% when there were no interviews at all. Although 11% seems quite a lot, as demonstrated (5.6.2) when 11% of a learner's responses were systematically removed from a learner's profile and then randomly reassigned to the nine potential categories there was little change in profile shape. A margin of error of this magnitude does not appear to have much of an effect on the profiles. This gives some justification to Fitzpatrick's decision to cut interviews from the methodology in her later studies (2007; 2009). It would however be preferable to limit as many errors as possible within acceptable time restrictions. As partial-retrospective interviews offer researchers an efficient way of doing this, it is argued that they are a useful addition to Fitzpatrick's methodology. Also, knowing how good native rater intuitions are likely to be when analysing responses is in itself useful. Depending on the level of error that researchers are prepared to accept, they can now make an informed choice. They may decide to live with such a margin of error and not use an interview. The time saved could be used to extend the stimulus list, add a test of depth for the same items or test another dimension of the learners' vocabulary, such as size. As such a measure of lexical organisation would be most usefully

employed as part of battery of tests measuring lexical competence (as in Albrechtsen et al., 2008), minimising the time needed for the WAT is important.

9.3 Why are profiles internally reliable?

In the previous section I have already speculated as to why there is so much individual variation, the reason(s) behind the internal consistency of L2 response profiles is however more difficult to discern. From Fitzpatrick's 2009 findings we could argue that learners are moving towards their L1 preference. This doesn't however answer why L1 responses (Fitzpatrick, 2007; this thesis) also seem to be internally reliable.

I suspect that one factor is that the way words are learned contributes to the kind of associations made between words. This was the conclusion that Politzer (1978) arrived at. He found that some teaching methods, such as dialogue drills generate syntagmatic responses (r= 0.56) and that some methods, such as substitution drills, generate paradigmatic responses (r=0.55). Although there are a variety of ways to learn a word, an individual probably learns many new words in a similar fashion, dependent on his or her learning preferences. For some learners whenever they come across a new word they look it up in a dictionary or ask someone what it means. Others however don't explicitly try to understand every new word but allow evidence to build up from the contexts that they meet this new word in and then make an informed guess. As learners (L1 and L2) are likely to stick with the strategy (or mix of strategies) that seem to work for them, the way they learn many of their words is likely to be similar. The L2 learner who for example often looks up unknown words in a dictionary and makes a note of the definitions might be expected to consistently give meaning based responses. As Politzer's 1978 study does not convincingly support his conclusions, a more robust investigation into the role of learning strategies on associations is warranted. I think it likely that studies pursuing this line of investigation would find a relationship between an individual's word association responses and their L2 learning environment. This could be done through the use of a WAT, as explained in this thesis, and a detailed questionnaire/interview of each learner's study habits and main sources of L2 input. Although there is good reason to suspect that differences in learning strategies and a learner's educational environment are likely explanations for the lack of homogeneity in individual profiles, other non-linguistic factors in the learner's backgrounds could also have a role. A related area

that might benefit further investigation would be to try to identify other characteristics that influence profiles. It is possible that we could have predicted the profiles in this thesis based on variables such as; age, gender or general intelligence. In this thesis such background data were not collected although a recent L1 study by Fitzpatrick and colleagues (2013) suggests that age has a significant role. For L2 learners though we will have to wait for further studies focusing on age and other potential predictors.

There is another reason I would not expect individual responses to fluctuate too much. This is because I view the mental lexicon as a relatively stable cognitive system that makes gradual adjustments, as opposed to a seething mass of constant change. A lexical network that changes in a steady way would I believe allow for quicker retrieval, even if this were at the expense of retrieving the optimal word. If the primary associative links between words readjusted every time a new aspect of word knowledge were acquired then the system would be in a perpetually confused state: unable to retrieve items capable of performing a particular communicative task within a reasonable time. This tension between the need for speedy retrieval and improving accuracy of expression might mean that new aspects of word knowledge are not initially used but lie dormant until overwhelming evidence is built up to confirm their utility. The lag between acquiring a new aspect of word knowledge and it becoming an active associative link acts as a damper, limiting the fluctuations. This helps the system to operate at a reasonable speed, albeit with the occasional suboptimum word being used. Even though new aspects of knowledge for a word might be acquired, an earlier associative link is likely to retain primacy until there is strong evidence that a much better association exists.

9.4 Creating stimulus lists

As has already been noted in the review of findings, word association stimuli need to be carefully considered when constructing stimulus lists. Failure to do so may result in responses that are unintentionally affected by particularly strong associations that words have with just one another word, thereby masking the response characteristics of an individual. Of the word related variables that might unduly influence responses, early word association studies (Deese, 1965; Cramer, 1968) note "frequency, word class, emotionality, vividness and intensity". In this thesis I have only really addressed the effect of the first two on Fitzpatrick's individual profiling framework. These two seemed the most likely to have some kind of effect, and as the

experimental chapters demonstrated the word class of a stimulus does encourage some types of responses. At this stage, the potential effect of emotionality and vividness/intensity can only be speculated on. As well as the word related variables, other factors need to be considered, how to present and collect the stimuli (written, orally, via a computer), whether the words match the proficiency level of the students and the L1/nationality of the learners. An understanding of how these variables affect responses helps us to answer a crucial question, how many stimuli to use.

Of the word variables that have not yet been explored using the individual profiling approach, stimuli which have high emotional value would seem likely to have some sort of an effect on responses. With stimulus words such as *sex*, *death*, *suicide, hate,* or perhaps swear words, it is easy to imagine the testee responding uncharacteristically. In fact, due to the nature of these words they may elicit nothing and simply waste time. Which words will have such an emotional effect is probably culture and age dependent. Considering precisely who will be taking the WAT and anticipating the emotional impact of the stimuli seems a sensible step in creating a productive stimulus list. The effects of emotion words and also words that are particularly 'vivid' have received some attention in the literature (de Groot, 1989; Altarriba et al, 1999). Altarriba and colleagues, studied the effect that emotional, concrete and abstract stimuli have on L1 word association responses. A norms list was created based on the responses of 55 university undergraduates to 154 abstract words, 100 concrete words, 98 emotion words. The findings were that many of these words have strong associations to just one other word, indicating such words may be unsuitable as stimuli in a WAT aiming to uncover individual response characteristics. The concrete words had an average primary association of 35.39 %, a word such as *canoe* elicited *boat* 44% of the time. Abstract words averaged 29.80%, for example *welfare* elicited *poor* 38%. Emotion words averaged 28.62%, for example *rage* elicited *anger* 51%. As the cut-off point for acceptance into the tests in this thesis was set at 25%, many of these concrete, abstract and emotional words would therefore not be acceptable. As argued in the Adjective study (7.8.2) words that appear unsuitable based on native norms list are not necessarily unsuitable within the target learner community. In the Adjective study it was also demonstrated that while native word lists might, at best, be used as a 'rough guide' for very high frequency items their usefulness decreases as the frequency of the words gets lower. Another point to consider is that words are culturally bound, this is probably more so for the emotional

ones. As also stressed by Fitzpatrick (et al., 2013) relying on a native norms list for guidance on what to include or exclude from stimulus lists is not really justified, there is a need for local norms lists to be drawn up. Norms lists created with high level bilinguals from the same community as the learners in the study would seem to be more valid as a guide for selecting which words to use in a WAT with L2 learners. In this thesis a word's emotional burden was not specifically accounted for in the stimulus selection procedure although many of these words had already been filtered out for other reasons. Many such emotion words (e.g. *mother, hate*) are high frequency words which would have been eliminated anyway, based on their having >25% of their responses to just one word. Other words would have been eliminated on the criteria of being 'too difficult for the learners'. In retrospect when I look back over the lists used (Appendix 4.1 – 7.2) I cannot identify any words which would have been unproductive due to their emotional content. There is the odd word (such as *to die* used in the Verb study) that may have had an emotional impact on some learners, those who had just experienced a bereavement, although from the responses this wasn't noticeable. This particular item for example had a similar number of responses to other items. If it were an item that many students found offensive or embarrassing to answer I would have expected a lot of non-responses.

In the absence of a high-ability learner norms list from which to select suitable stimuli, pilot testing with learners of a similar ability and background to the learners in the main study group is recommended. This point applies not only to WA studies interested in the organisation of the lexicon, but also those in using WA as a measure of language proficiency. Piloting not only helps identify words which are L1 cognates or words that have a strong relationship to just one word, but will probably help pick up on any words which have a high emotional content or words that are in-vogue with that particular community. In the studies in this thesis the word lists were all piloted with Japanese students, and in the case of the adjectives piloted more than once. While some unsuitable words still slipped through and had to be discarded prior to the analysis the pilot tests proved to be invaluable. The importance of trialling words was demonstrated in the Noun 2 study (Chapter 5) which included nine non-Japanese learners. The reliability of the responses by two of these learners (Indonesians) was far lower than the rest of the group, which I believe was due to only piloting the words with Japanese learners. It is likely that some of the words in that study were unsuitable in some way for Indonesians. Coming back to the point about being wary

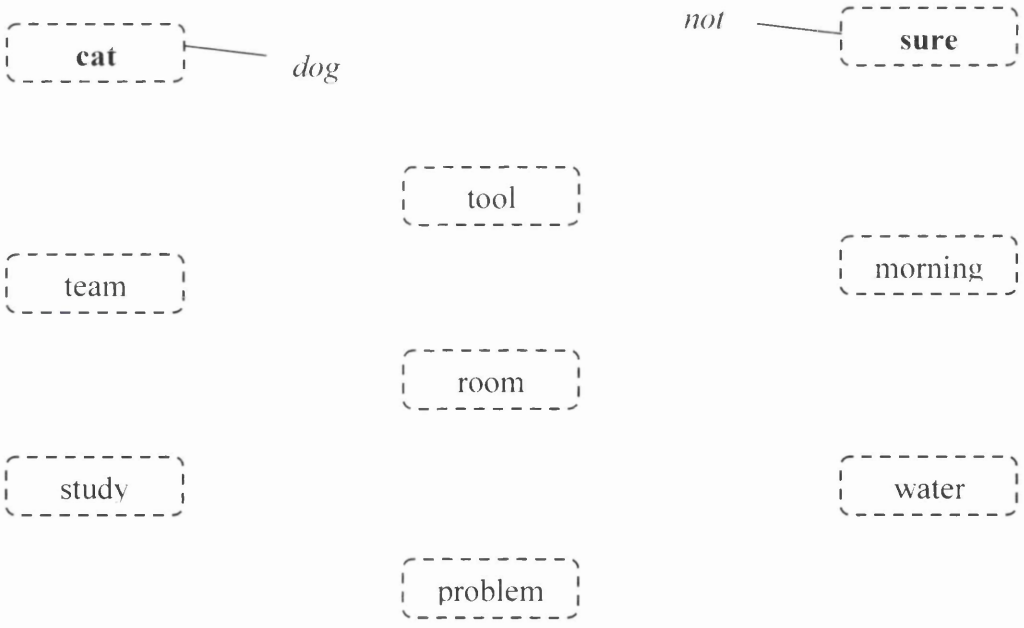
of in-vogue words, an example from the replication study illustrates this well. In that study fairly infrequent stimuli were used, one of them was the word *kindle.* When the data for the original study (late 1990s) and the replication study (early 2007) were collected it was not considered problematic. Its use was limited to the literal or metaphorical meaning of starting a fire. Nowadays of course the word *kindle* would more likely be associated with reading following the popularity of the electronic reader of that name, a product first released in late 2007. Interestingly, a common association for *kindle* these days might well be the same as one of the common associations in the replication study: *fire.* This response would still be quite likely, not because of the association with starting a fire but because the newest version of the product is called Kindle-Fire .The thinking behind the association to this new use of the word is quite different and would consequently require a different classification; the response *fire* is no longer associated based on its meaning, it is now associated based on its form. Untangling whether an association of *kindle →fire* was due to the traditional meaning or the product (or perhaps a bit of both) would currently make this in-vogue item a poor candidate for use in a word association test.

The next point that will be discussed is how the stimulus words were collected. In the replication the data were collected orally, whereas in the subsequent experiments the tests were in a pencil and paper format. The main reason for the switch to a written test format was time. The oral method only allows one or two learners per hour, the written format allows around 10 students to be tested (and then interviewed on a limited number of responses) within the same time frame. There do seem to be some benefits to an oral approach. Classification is sometimes easier as the interviewer can use facial expressions, body language and tone of voice to help understand how the respondent is associating words. If there is still some ambiguity, then immediate follow up questions are also possible. These benefits however need to be weighed against the increased time needed to collect data in this way. If an experimental design involving a large number of participants is envisaged then a written format is easier to administer, if however a case study experimental design is used (Chapter 8) then a researcher might be able to more accurately classify oral data. A point concerning the layout of the written forms, identified in the Verb study (6.7.1), is that they may increase the number of xy collocations generated due to the participants being asked to write their responses to the right of the stimuli. In the Verb study a lot of xy collocations were generated but there were only a few yx

collocations. Although various possibilities for the lack of yx collocations were put forward, the idea that it is due to the format of the test paper seems easily testable. One way to do this would be to examine how stimuli that prior word association studies have shown to elicit mainly yx collocations behave when tested using first an oral and then a written format. If the left to right written format does encourage xy collocations then we would expect to find fewer xy collocations in the oral format. Alternatively, the format of the written test forms could be altered so as to allow participants more freedom with where they write. Rather than a stimulus with a white space to the right of it, the stimulus item could be presented in a format exemplified in Fig 9.3. Such a format might also make it more visually obvious, in the case of ambiguous responses such as *hot →pot* or *flower →power* whether the respondent was making a Form-based link or whether the participant was thinking of a collocation.

Fig 9.3 An alternative way to present the stimuli

Please write the first word you think of when you read the words below. For example: with **cat** you might think of **dog**. With the word **sure** you might think of **not**.



It ought to be noted that the alternative method of presenting the stimuli (Fig 9.3) is only intended to illustrate how the usual layout (see Appendices) might be adapted. There are not nearly enough stimuli (seven) to obtain a representative sample of responses. As will be argued in the subsequent paragraph, even three times this would be a minimal amount.

A point to emerge from Chapter 2 was that past studies vary considerably as to the number of stimulus words used. Some such as Kruse et al. (1987) feel that nine stimuli are enough where as others (Fitzpatrick, 2009), erring on the side of caution, have 100 words per list. As the number of words determines how much time the test will take a question that was raised fairly early on in this thesis was 'how big a sample is necessary to generate a reliable profile?'. Following past WA studies (Deese, 1965) in the experiments in this thesis it was decided to give learners stimuli lists of between 90 and 100 words each, which when split into two frequency levels (45 -50 words per list) could be used to verify the reliability. In this way it was established that word lists of between 90 -100 words could be used to generate reliable learner profiles. While tests of this length can be completed within acceptable time limits (30 – 40 minutes) it doesn't really allow for the measurement of anything else in the same session. If, as suggested in 9.2, the WAT is to be used as part of a battery of tests (Albrechtsen et al., 2008) then it would be useful to know if a shorter test would give profiles that were similarly reliable. Researchers wishing to clarify the thinking behind responses with interviews, measure depth of word knowledge, or perhaps the proficiency of the test taker, need to know the minimum number of items necessary for the word association component. Without an answer to this basic question it is probable that many experiments over burden students with unnecessarily long lists of stimulus words, and by so doing, limit the kind of questions that can be addressed. In order to give a more precise answer to the question of how many stimuli to use in a word association test, the 248 responses given by M (Chapter 8) in the word class study retests (2012) were randomly sampled a number of times. The data were randomly sampled six times, each time a progressively smaller sample was drawn. The idea being to identify the point at which a profile becomes unreliable due to not enough stimulus words being used. As the retest data correlated very strongly ($r=0.92$) with the initial test data, the full set of responses were judged to reliably reflect M's characteristic response profile.

Table 9.3 Random samplings of M's responses (2012)

Responses randomly sampled	Lex set/ context related (%)	Xy collocation (%)	Conceptual association (%)	Specific synonym (%)	Defining synonym (%)	Yx collocation (%)	Affix change (%)	Similar form only (%)	Other collocation (%)
All (248)	**22.98**	**20.16**	**20.16**	16.53	8.06	5.65	3.63	2.02	0.81
100	**23.00**	**19.00**	**20.00**	17.00	10.00	6.00	3.00	2.00	0.00
50	**22.00**	**18.00**	**16.00**	10.00	14.00	12.00	4.00	4.00	0.00
40	**35.00**	**20.00**	**22.50**	10.00	7.50	5.00	0.00	0.00	0.00
30	**33.33**	**16.67**	**26.67**	10.00	3.33	6.67	3.33	0.00	0.00
20	**35.00**	10.00	**15.00**	**20.00**	10.00	0.00	5.00	0.00	5.00
10	**30.00**	0.00	0.00	**20.00**	**30.00**	10.00	10.00	0.00	0.00

note: **bold** values indicate top three ranked categories

As can be seen in Table 9.3, when 100 responses were sampled from the original 248 responses the ranking of the categories was identical and the percentages were also very close: 100 stimuli gives reliable data. When 50 were sampled, the top three response categories were ranked in the same order as in the profile generated from the 248 responses, the percentages were also similar. With 40 and 30 responses, again the order of the rankings of the top three response categories did not change and the percentages were still quite close; the lower ranked categories however started to show some variation. When 20 responses were drawn a different pattern began to emerge, although the top and third ranked categories were the same the category ranked 2nd became joint 5th. When only ten responses were drawn, other than the top category retaining its position the profile generated was quite different to the original. This random sampling procedure indicates that for a classification system with nine discrete categories a sample of below 20 generates unreliable data whereas more than 30 responses is enough to give a profile that corresponds well with a much larger sample of responses. This seems to justify the intuitive decision (initially adopted in Chapter 4) to reject students who made less than 25 responses in any of the word association tests.

9.5 Classification problems

A common gripe amongst researchers (Meara, 1987; Sökmen, 1993; Wolter, 2001; Orita, 2002; Henriksen, 2008; Shimotori, 2013) is the difficulty in establishing a classification system that unambiguously categorises all responses into discrete categories. Without some sort of retrospective check a common failing of word association studies is that many responses inevitably end up being misclassified or discarded into an *erroneous* or *other* category. The classification system in this thesis follows Fitzpatrick (2007) which was in part motivated by a wish to develop a more objective and thus more efficient system. On the face of it, Fitzpatrick's classification is well thought out. By mapping directly onto the aspects of word knowledge (Nation, 2001) it ensures the inclusion of all potential responses. Another clever feature is that the subcategories can be rolled up into main categories which are similar to the traditional paradigmatic, syntagmatic and phonological categories: allowing some comparison with previous studies. In practice I also found it (compared to the syntagmatic/paradigmatic division used in Chapter 3) to be fairly user-friendly due to the clear definitions and examples for each subcategory. Despite these positive points

there were still some responses that proved difficult to classify, even when coupled with an interview aimed at digging deeper into the thinking behind ambiguous responses. As noted earlier, collocation categories were a cause for concern, there also seems to be a problem with the synonym categories. A recent criticism (Shimotori, 2013:87) of Fitzpatrick's classification system is that "it is probably impossible for a participant to think of a word without having that word's meaning in mind". Shimotori's study of Japanese and Swedish learners therefore rejects Fitzpatrick's classification on the grounds that Position-based and Form-based associations ought to be viewed as subcategories of Meaning-based associations. I think though that this somewhat misses the point. The classification system is not aiming to identify the only association between two words but, of the multiple associations (meaning included) that a person has, identify the strongest. Still, an indication that the categories are not quite right, comes from the observation that responses in this thesis are unequally distributed.

As suggested in section 9.4 the problem of collocations might be due to the collection sheet favouring xy collocations. A solution might therefore be to adapt the collection procedures with a better WAT format (Fig 9.3). Another possible explanation is that *yx collocations* are probably not as common in the language as *xy collocations*. One of the assumptions that Fitzpatrick's classification system rests on is that with each stimulus there is an equal chance for any of the nine categories to be chosen. If it is the case that certain responses are more (or less) likely to be generated than others, then the classification system will need to be rethought. As well as the collocations, other sub categories that have been questioned are the *defining* and *specific synonym* categories. With native speakers this distinction seems valid as both participant and rater are usually aware of whether the response means the same as the stimulus or whether it can only be used as a synonym in specific circumstances. As noted in a recent replication of Fitzpatrick's 2006 paper (Racine, 2012) the problem is that for learners this distinction becomes blurred, learners do not have such a clear understanding of the language. Sometimes they think they are defining a word but by native standards they often fail to do so, giving what a native speaker would judge as a close synonym instead. When a learner gives a response such as *big* to the stimulus *wide* she may well be giving this response as she thinks it is a definitive synonym. A native speaking rater would probably not realize this though and judge it as a specific synonym. For the native speaker a 'definitive' response to *wide* would be a response

such as *vast* or *broad*. If such words are unknown to the learner though, it could be argued that with the response *big* the learner is in fact responding with as definitive a response as possible. Even if the learner is asked to explain why she gave this response, or a panel of NS judges were asked to arbitrate, it seems unlikely that a satisfactory classification would be made. Does one classify *big* as a *specific synonym* because by native standards it is not considered definitive or does one classify it as a *defining synonym* because for the learner it is the most definitive response she can possibly make? I think that for many learners the distinction between the *defining synonym* and the *specific synonym* categories lacks a certain amount of validity. Due to a concern that these two types of response are not easy to discriminate, there is a good case for conflating the categories. It is notable that in her study of L1 WA responses Fitzpatrick (et al, 2013) does indeed do this.

The split-half analysis of learner responses in the initial Noun study (Chapter 4) showed surprisingly high corrrelations. These were then confirmed in the subsequent experiment. As high correlations continued to be seen, a nagging doubt began to emerge that perhaps these correlations were too good to be true. The suspicion was that the high correlations might be artifactual in nature. The correlations were made by comparing the response type each learner gave to two sets of around 50 stimuli (33 for Adjectives). The assumption was that every stimulus had an equal chance to generate any of the nine potential response types. The problem is that if one of the categories is hardly ever chosen by any learner then this category will nearly always be ranked lowest. When compared using a correlation analysis the zero responses for that category in one array will perfectly match the zero responses in the other array. This perfect match will mean that of the nine categories only eight are truly free to vary, this creates an inflated correlation value.

With the initial experiments there was not enough data to confirm or deny this doubt although as more data began to be collected it became clear that some response categories were more equal than others. In Table 9.4 the results of a random sampling of 60 learner profiles can be seen. These 60 profiles were selected randomly from the three word class studies (20 from each study - Chapters 5 - 7). The number of times that each category occurred at each rank was counted. As can be seen, the *definite synonym* category for example was never a top ranking category for any of the 60 learners sampled. Some of the categories were also never ranked lowest. With this number of students I would have expected a more even distribution.

Table 9.4 Number of categories at each rank from 60 randomly selected profiles (Noun study 20, Verb study 20, Adjective study 20)

	Defining Synonym	Specific Synonym	Lex set/ context related	Conceptual association	Xy collocation	Yx collocation	Other collocation	Affix change	Similar form only
rank 1	0	11	17	18	16	0	0	0	0
rank 2	2	11	12	27	8	0	0	2	3
rank 3	5	16	14	8	10	3	1	9	1
rank 4	19	11	11	3	6	4	2	2	1
rank 5	14	8	5	2	8	13	2	17	14
rank 6	4	3	1	2	9	14	15	11	17
rank 7	8	0	0	0	1	12	20	11	15
rank 8	6	0	0	0	2	12	16	8	8
rank 9	2	0	0	0	0	2	4	0	1
Number of profiles	60	60	60	60	60	60	60	60	60

Table 9.5 The number of M's responses to adjective stimuli (2012 WAT)

Subcategory	Profile 1	Profile 2
defining synonym	3	3
specific synonym	4	9
same lexical set/context related	1	2
conceptual association	4	7
xy collocation	18	9
yx collocation	1	1
other collocation	0	1
affix change	2	1
similar form only	0	0
Total	**33**	**33**

Reasons for the low number of *defining synonym* responses have already been discussed in the previous section although as Table 9.4 shows there is an uneven distribution in many of the categories. If some categories are rarely used then this will affect the probability of a category being placed at a particular ranking. The problem of the response classes not being equal therefore means that the correlation values are likely to be inflated to some extent. One solution to this, allowing us to calculate more accurate statistical values might be to combine some of the underused categories that measure similar response types. Another possibility would be to exclude any unused categories in the correlation calculations. To exemplify the kind of change in correlation values that could be expected with such an adjustment, a profile made for M in the previous chapter will be reexamined. When we look at M's profiles for the responses to adjective stimuli in 2012 (Table 9.5) the Pearson correlation between the two arrays is 0.731. We might note that in these two profiles the *similar form only* category was unused, this category matches perfectly thereby inflating the value. When the correlation value is recalculated without this category the value shows a slightly weaker relationship (r=0.711).

Although there seems to be a cloud over the 'too good to be true' correlations that were calculated from the response data it was also demonstrated in other ways that a person's word associations are reliable. In Chapter 4 for example the top two dominant categories were combined (Table 4.3) to create pair categories for each individual. As using the top two categories gave an average coverage of 66.19% of all responses it was argued that a good description of a learners' typical response behaviour could be made. When each individual's 'top pair' in the first profile was compared with their 'top pair' in the second profile, 78% matched. Rather than attempting to make profiles from all the response data, a simpler analysis that only aims to identify the top two or perhaps top three response categories is perhaps good enough. From a practical point of view, it would seem sufficient to be able to state for example that: *K's profile is dominated by xy collocations and same lexical set/context related responses*. Going into the details of the minor responses categories would probably be unnecessary for researchers or teachers and perhaps overwhelming for learners.

9.6 Automaticity of responses

The failure of the think-aloud method in Chapter 8 was disappointing although it does seem to highlight an interesting feature of word associations: many of them are generated automatically. Of the possible reasons given for the inability of M to verbalise her thinking during the word association tasks, the most likely explanation is that word associations are often processed too quickly for us to consciously analyse.

An everyday example might help to explain what I mean by 'automatic generation'. When an experienced car driver makes a right turn he does not consciously think through all the sub-tasks (taking his foot off the accelerator, applying the brakes, pushing the clutch pedal, selecting a lower gear, activating the indicator lights, checking his mirrors, turning the steering wheel, looking left etc.), he just turns. The driver probably attends to a few of these tasks consciously but many of them will have been practiced so often as to be automatic. Due to the limitations of conscious processing, were he to try to attend to all these sub-tasks consciously he would probably crash the car! In a similar way, a word association can be thought of as a sub-process of language production that a person usually does not need to consciously attend to. Were this person to attempt to consciously process each association at the same time as attending to the other sub-tasks necessary when talking, the conversation would proceed extremely slowly – perhaps even crashing. Consequently, it is difficult for respondents to verbalise why or how they are making specific associations because they are often processed automatically.

An additional perspective on automaticity can be gained by considering word associations through the Construction-Integration model proposed by Kintsch (1998). In this model, which aims to explain how written texts are understood, Kintsch argues that when a sentence is read all the possible meanings and associations are activated. Based on the readers' background knowledge the irrelevant interpretations are suppressed and thus the reader constructs their image of the text. It is only when this automatic default fails that the reader has to work out the image consciously. As Kintsch's experiments demonstrate, the conscious processing of texts takes more time and so is automated whenever possible. Similarly, I would argue that when people make associations between words, by default these are made automatically, unless the word is only vaguely known and thus requires conscious processing. In the set of WATs detailed in Chapter 8, which pressed M to work quickly through a list of well-known words, she did not have the time or need to consciously process her thoughts,

thus there was nothing to verbalise. Even if given more time and more encouragement to think-aloud it seems doubtful whether this technique can enhance word association studies of this sort. In Henriksen's 2008 study we might note that in the word association task, part of a larger linguistic project into how vocabulary and writing develops (Albrechtsen et al. 2008), a retrospective task was used rather than the think –aloud verbalisations used in the other parts of the project. The decision to use an alternative introspective method was presumably due to the realisation that word association tests often tap into an automatic rather than a conscious mental activity.

9.7 Pedagogical applications

Given that a well organised network of words is considered to be one of the requirements of full language competence we might ask what teachers ought to do to help their learners develop such networks. The idea that *the way in which language is taught is reflected in the mental lexicon* was suggested by Politzer (1978), and later Sökmen (1993). The question Sökmen asks is "Which associations are useful to teach?" At around this time a number of commentators (White, 1988; Holland,1990) argued in favour of directly teaching word association networks. White (1988) offers activities that EFL teachers might incorporate into their classes as a way to "review and refine" word knowledge. The teacher could for example give a cue word and instruct students to "write down all the associated words that come to mind…within a minute" and then compare with others in the group. In Holland's paper there is an explanation of a computer based system that uses hypertext to allow learners to work through a pre-prepared L2 network. For the US soldiers that learned words in this way she argued that it was more motivating than the rote memorisation method that it replaced. It should be noted however that neither White (1988) nor Holland (1990), support their claims with empirical evidence.

There is probably some benefit in the occasional class activity aimed at raising learner's awareness of what a well-developed associative network might look like. It does not however seem appropriate as the main method of vocabulary instruction within a language course, as in Holland (1990). The passive nature of Holland's computer program for example gives students little encouragement to produce the words, find out what they sound like or identify typical situations in which they are used: they are learning words in a decontextualized way. A further problem would be identifying what to include/exclude from the associative networks. With low

frequency technical words (Holland's soldiers were learning L2 equivalents to: *howitzer*, *firearm* and *garand*) it was possible to work out simple networks but with higher frequency words, and far more polysemy, the complexity of the network might become overwhelming. Given the amount of learner idiosyncrasy observed in this thesis, developing a set of 'common associative networks' for learners to study could even be counterproductive. Teaching a group norm might conflict with a learner's predispositions and impede the development of a more natural network. In fact, the idea of 'teaching associations' strikes me as a case of putting the cart before the horse. Rather than directly teaching word associations I view a well-developed word network as the outcome of good language teaching. I think associative knowledge is better achieved by implicit methods and is not something that needs to be explicitly stressed within a language class. Exposure to large samples of the language (reading and listening extensively) and being given corrective feedback on attempts to produce the language will in my opinion lead to a well-developed network. It does not really follow that it would work better the other way around. It might be noted that this approach has not been adopted by current proponents of explicit vocabulary instruction. In recent vocabulary teaching guide books (Nation, 2008; Zimmerman, 2009) there are no activities aimed at developing the organisational dimension of the lexicon of the kind described by White (1988) and Holland (1990). Although one activity "semantic mapping" described in Nation (2008:95) is similar to the activity detailed by White, the purpose is quite different: to prepare students for a writing task. So, rather than asking how we can 'teach' word associations, I think it is better to ask how word association research findings can be applied to help learners develop their own network of words in their own way.

The findings of this thesis, echoing the findings of Fitzpatrick (2007; 2009), suggest that a learner's characteristic response profile is idiosyncratic yet internally reliable. A useful application of this might therefore be to use WATs as a way to tailor current vocabulary learning strategies to suit each student. This could be done through giving a student a word association test and then picking learning activities that match the main response category(ies) that are identified in the student's response profile. One might argue that a student who generally responds to prompt words using words in the *same lexical set* would respond to learning activities that help build on words thematically. When for example a new word is met in class, the students identified by the WAT as being *same lexically set* orientated could be encouraged to make 'word

families' in their vocabulary notebooks. Of course these students would need to add to the other aspects of word knowledge later on in their studies, but such a task might give them a handle on the word in the initial stages of acquisition. In Chapter 8, M for example showed a preference for *collocations* with verb stimuli. When M comes across a new verb perhaps it would be beneficial for her to build on this preference and as part of her personal learning strategy for verbs to use online concordancing software. Looking up the kinds of potential phrases and pairings that a particular verb often has might suit her. While this link between word association response characteristics and learning strategies for vocabulary has yet to be established, it does seem to be a promising area for further study. It would be interesting to give a student a word association test in order to identify his/her response characteristics, and then give that student two sets of words to learn. The first set would be learned in harmony with the main response characteristics identified while the second set would be learned in the way that the student usually learns words. If my hypothesis is correct then the words learned in harmony with the response characteristic would show better retention in a subsequent L1/L2 vocabulary matching test.

It is possible that a person's unique word association response characteristics reflect a 'best study path', a path that would also seem to vary with the individual and also with the word class of the item being studied. Whether strategies based on WA tests would result in an improved uptake of words is beyond this study but I think it would be an interesting avenue to pursue. Various vocabulary learning strategies have been proposed to help learners acquire words: using word cards, connecting words with places or situations, guessing from context, the keyword method and saying words out-loud (see Pavičić, 2008 and Nation, 2008 for a wider discussion). The general advice to teachers is to raise students' awareness of the available strategies, in the hope that they will adopt the ones that suit them best. Given that there are so many potential strategies available though, it would only ever be practical to introduce students to a few of these within a regular language course. It is therefore necessary to have a method of objectively narrowing down the strategies and learning styles most suitable for particular students. Pavičić (2008: 83) suggests interviews and questionnaires, although responses to word associations might also help to predict which vocabulary learning strategies and activities would work best for certain students. As noted in the previous section, many associations between words seem to be made automatically, without the learner being particularly aware of how or why

they occur. Therefore asking students to make a conscious choice (such as a questionnaire) on how best to develop their mental lexicon might not be so effective, as they may lack the metacognitive awareness to do this. Also, as many students have an incomplete knowledge of the potential strategies available, a word association test could be more efficient as it does not require them to make a metacognitive decision. A word association test might turn out to be particularly useful in identifying 'best study paths' for younger students. Those under the ages of 12 or 13 would probably not have much experience in the potential learning strategies available to them, and even if they did, they may lack the maturity to think through what would suit them.

9.8 Summary of General Discussion

In the previous sections the main findings were reviewed and various proposals made as to how research into the organization of the mental lexicon through the use of word association tests might proceed. One proposal, based on the finding that the word class of a stimulus word has an effect on the type of response generated is that stimulus lists ought to consist of a sample of words representative of the language. The next proposal is that as this thesis was limited to stimuli drawn from the most frequent 2000 words, more research needs to be conducted with lower frequency stimuli. Although a frequency effect was not evident in the ranges tested there was a faint suggestion within the individual case study data that with lower frequency stimuli the data might become less reliable.

The series of experiments also highlights areas that the methodology can be improved. It is argued that a partial retrospective interview, as opposed to a full interview, is a useful addition to the methodology outlined by Fitzpatrick (2007). This is supported by the finding that native intuitions are reasonably accurate and so a time consuming verbal confirmation of the thinking behind every response given is unnecessary. It is also argued that the usefulness of native speaker norms lists in helping to identify productive stimuli is limited. In order to weed out the words that are unlikely to generate responses that show a person's characteristic response preferences, norms lists based on high level bilinguals from the learners own country would be more useful. In the absence of such a norm list though, pilot testing of all stimuli is advised. Another problem identified was with the format of the tests that might have favoured xy collocation responses. An alternative kind of WAT is suggested that addresses this issue, a further study is needed to assess the

effectiveness of the proposed format.

Concerning the classification system used (Fitzpatrick, 2007) a number of problems were identified. Some of the categories appear redundant for the learners in this study with an unequal distribution suggesting that some revision to the classification system is needed. Some concern was also raised over the correlation values, which may not be reflecting the relationship between the learner profiles as accurately as I would like. However, evidence from other parts of the thesis support the main claim that the approach is reliable.

From the finding that the think-aloud procedure did not generate much useful data a little more was learned about the way in which word associations work. It appears they are often processed at a subconscious level and so when asked to verbally explain these thought processes often a person cannot, as there is little conscious activity to verbalize. The implications of this being that we should not put too much reliance on introspective data as people are not always aware of why they make particular responses. While retrospective interviews can often enlighten they might also mislead.

The potential pedagogical applications of the main findings suggest another interesting area for further research. The learners within this thesis were characterized by their lack of homogeneity, it is therefore argued that a common path to acquiring vocabulary is unlikely. Rather, each learner probably has a 'best learning path'. It is hypothesized that WATs of the type detailed in this thesis might help to identify such an optimal learning path for each learner. Of the many kinds of word learning strategies available to learners it seems possible that a WAT could predict what strategies would suit that learner. It might even prove to be sensitive enough to predict the kind of learning that would suit a learner for each word class.

As well as the potential pedagogic applications of considering learners in terms of their dominant WA response categories (e.g. a *synonym orientated* or a *collocational + lexical set orientated* learner) it may also prove to be a useful way to group learners in further research on the structure of L2 lexicons. While the approach taken in this thesis was to analyse 'individuals' rather than 'groups', for research into L2 lexicons that does require some kind of grouping of learners a WAT seems to offer a more precise alternative. Currently, it is typical for learners to be categorized using indirect measures of cognitive development, such as age, gender or educational background. Despite there being good reasons for such groupings, the considerable

variation that has already been commented on has often led to unclear findings. It would therefore seem more logical for such research to attempt to categorize L2 learners, perhaps as an initial step, through measuring their cognitive structure more directly. The WATs and 'individual profiling' style analysis discussed in this thesis offer a reliable method for categorizing L2 learners in this way.

Chapter Ten: Conclusions

A consistent thread running through this thesis is that using word associations as a probe into the mental lexicon is not easy: great care is needed at every step. Inadequate preparatory work in selecting the stimuli will lead to response data that fails to satisfactorily answer the questions posed. The kind of analysis that is used to interpret the data is also an important consideration; in this thesis an individual approach was adopted, which appears to have benefits over group data in terms of reliability. Through repeatedly attacking Fitzpatrick's *individual profiling* approach from a number of different angles various issues have been addressed and some progress has been made, although as explained in the previous chapter there is still plenty of work left to do. Fitzpatrick's classifications for example, while an improvement on traditional systems, still needs some fine tuning. The lack of data on how L2 learners respond to lower frequency stimuli is another gray area. Despite this, the main finding that an individual profiling approach can deliver reliable response data, is encouraging. Further research in this field, which had stalled due to an inadequate methodology and approach to data analysis, can now be expected.

We are still a fair way from a comprehensive model that could satisfactorily explain how learners integrate new words into their mental lexicons and organise, retrieve and deepen knowledge of acquired words. The findings from this thesis do however underline three elements that any such model would need to incorporate. The first is that it would need to recognise individuality. A model based on group norms seems unworkable due to the large amount of within-group variation found in this and other recent studies. Secondly, any fully inclusive model would need to recognise that although some words (and the various aspects of word knowledge) are processed at a conscious level, many are processed at a subconscious (automatic) level. The third is that different kinds of words are probably stored and processed in different ways. In this thesis there was a bias observed towards particular types of response from the three kinds of stimuli (nouns, verbs and adjectives). This suggests that words with fundamentally different functions are not organised (or perhaps retrieved) in the same way. While a comprehensive model represents a long-term goal, there are immediate benefits to the findings of this thesis. At the very least, a reliable method of measuring the production of word associations gives us the opportunity to better understand how a learner's lexicon is organized. Coupled with a test of vocabulary size, this also ought to enable teachers and researchers to better assess L2 vocabulary competence.

Bibliography

Aichison, J. (1987) *Words in the Mind. An Introduction to the Mental Lexicon*. Oxford: Blackwell.

Aitchison, J. (1992) Good birds, better birds and amazing birds: the development of prototypes. In Arnaud and Béjoint *Vocabulary and Applied linguistics.* 71-84. London: Macmillan.

Albrechtsen, D., Haastrup, K and Henriksen, B. (2008) *Vocabulary and Writing in a First and Second Language: Processes and Development.* Basingstoke: Palgrave Macmillan.

Altarriba, J., Bauer, L. and Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, and Computers* 31/4, 578-602.

Bagger-Nissen, H. and Henriksen, B. (2006) Word class influence on word association test results. *International Journal of Applied Linguistics, 16/3,* 389-408.

Bauer, L. and Nation, I.S.P (1993) Word Families, *International Journal of Lexicography 6/4*, 253-279.

Beck, J. (1981) New vocabulary and the associations it provokes. *Polyglot, 3/3*, C7–F14.

Bleuler, E. (1924) *Textbook of Psychiatry*: (trans A.A.Brill) New York: Macmillan.

Brown, R. and Berko, J. (1960) Word association and the acquisition of grammar. *Child Development 31,* 1-14.

Burgess, E. W. (1924) The growth of the city. *The American sociological society. XVIII*, 85–99.

Channell, J. (1990) Vocabulary acquisition and the mental lexicon. In J. Tomasczyk and B. Lewandowska-Tomasczyk (Eds.), *Meaning and lexicography,* 21-31, Amsterdam: Benjamins.

Churchill, E. (2007) A dynamic systems account of learning a word: from ecology to form relations. *Applied Linguistics 29/3*, 339-358.

Cohen, L., Manion, L. and Morrison, K. (2006) *Research Methods in Education (5th Ed).* New York: Routledge Falmer.

Coxhead, A. (2000) A new academic wordlist. *TESOL Quarterly, 34/2*, 213-38.

Crossley, S., Salsbury, T., and McNamara, D. (2009) Measuring L2 lexical growth using hypernymic relationships. *Language Learning 59/2,* 307-334.

Cramer, P. (1968) *Word Association.* New York: Academic Press.

Cremer, M., Dingshoff, D., De Beer, M. and Schoonen, R. (2011) Do word associations access word knowledge? A comparison of L1 & L2 child and adult word associations. *International Journal of Bilingualism, 15*, 187-204.

Davies, M. (2008) *The Corpus of Contemporary American English: 425 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

De Groot, A.M. (1989) Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15/5,* 824-845.

Deese, J. (1965) *The Structure of Associations in Language and Thought*, Baltimore: John Hopkins Press.

Ebbinghaus, H. (1885) *Über das Gedächtnis.* Leipzig: Dunker and Humbolt.

Emmerson. H.F and Gekoski,W.L (1976) Interactive and categorical grouping strategies and the syntagmatic-paradigmatic shift. *Child Development, 47/4*, 1116 – 1121.

Entwistle, D.R. (1966) *Word associations of young children*. Baltimore: John Hopkins Press.

Ervin, S. (1961) Changes with age in the verbal determinants of word association. *American Journal of Psychology, 74*, 361-372.
Fitzpatrick, T. (2006) Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook, 6/1*, 121-145.
Fitzpatrick, T. (2007) Word association patterns: unpacking the assumptions *International Journal of Applied Linguistics, 17/3*, 319-331.
Fitzpatrick, T. (2009) Word association profiles in a 1st and 2nd language. In T. Fitzpatrick and A. Barfield (Eds.), *Lexical processing in second language learners,* 38-52. Bristol. Multilingual Matters.
Fitzpatrick, T. and Izura, C. (2011) Word association in L1 and L2: an exploratory study of response types, response times, and interlingual mediation *Studies in Second Language Acquisition, 33*, 373-389.
Fitzpatrick, T., Playfoot, D., Wray, A., and Wright, M. J. (2013) Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics*, Advanced Access published September 24, 2013. doi:10.1093/applin/amt020.
Freud, S. (1900) *The interpretation of dreams.* New York: Avon, 1980.
Galton, F. (1883) *Inquiries into human faculty and its development.* London: J. M. Dent and sons.
Gardner, D. (2007) Validating the construct of *word* in applied corpus-based vocabulary research. A critical survey. *Applied Linguistics 28/2*, 241-265.
Gentner, D. (1982) Why nouns are learned before verbs: linguistic relativity verses natural partioning. In S. Kuczaj (Ed.) *Language development: Language, thought and culture*, 301-334. Hillsdale, NJ: Lawrence Erlbaum.
Greenbaum, S. and Quirk, R. (1998) *A Student's Grammar of the English language.* Harlow: Longman.
Greidanus, T. and Nienhuis, L. (2001) Testing the Quality of Word Knowledge in a Second Language by Means of Word Associations: Distractors and Types of Associations. *The Modern Language Journal, 85/4,* 567-577.
Henriksen, B. (1999) Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*, 303-317.
Henriksen, B. (2008) Declarative lexical knowledge. In Albrechtsen, D., Haastrup, K and Henriksen, B. (Eds), *Vocabulary and Writing in a First and Second Language: Processes and Development.* Basingstoke: Palgrave Macmillan.
Higginbotham, G. (2010) Individual learner profiles from word association tests: the effect of word frequency. *System, 38,* 379-390.
Hofland, K. and Johansson, S. (1982) *Word Frequencies in British and American English.* Bergen: The Norwegian Computing Centre for the Humanities.
Holland, M.V. (1990) *Teaching a foreign language lexicon: A rationale for hypertext.* Virginia: US Army Research Institute.
Hughes, J. (1981) Stability in the word associations of non-native speakers. *Unpublished MA project.* Birbeck College, London.
Hulstijn, J.H. (2000). Mnemonic methods in foreign language vocabulary learning: Theoretical considerations and pedagogical implications. In J. Coady and T. Huckin (Eds.), *Second Language Vocabulary Acquisition,* 203-224, Cambridge: Cambridge University Press.
Hulstijn, J.H. (2007). Psycholinguistic perspectives on second language acquisition. In J. Cummins and C. Davison (Eds.), *The international handbook on English language teaching,* 701-713, Norwell, MA: Springer.

Japan Association of College English Teachers.(2003) *JACET list of 8000 basic words.* Tokyo: JACET.

Jarema, G. and Libben, G. (2007) *The Mental Lexicon: Core Perspectives.* Bingley: Emerald Group Pub Limited.

Jung, C.G. (1902) The associations of normal subjects. In: *Collected Works of C. G. Jung, 2*, 3-99, Princeton, NJ: Princeton University Press.

Jung, C.G. (1918). *Studies in Word Association.* Zurich: Taylor and Francis.

Kent, G.H. and Rosanoff, A.J. (1910) A study of association in insanity. *American Journal of Insanity 67*, 37-96 and 317-390.

Kintsch, W. (1998) *Comprehension: A Paradigm for Cognition*, Cambridge: Cambridge University Press.

Kiss, G.R., Armstrong, C., Milroy, R. and Piper, J. (1973) An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley and Hamilton-Smith, N. (eds.): *The Computer and Literary Studies.* Edinburgh: University Press.

Kruse, H., Pankhurst, J. and Sharwood-Smith, M. (1987) A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9*, 141-154.

Lambert, W.E. (1956). Developmental aspects of second language acquisition. *Journal of Social Psychology 43*, 83-104.

Lambert, W.E. and Moore (1966) Word –association responses: Comparisons of American and French monolinguals with Canadian monolinguals and bilinguals. *Journal of Personality and Social Psychology 3*, 313-320.

Leow, R.P. and Morgan-Short, K. (2004) To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26*, 35-57.

Meara, P.M. (1978) Learners' word associations in French. *Interlanguage Studies Bulletin 3/2*, 192-211.

Meara, P.M. (1980) Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics: Abstracts 13/4*, 221-246.

Meara, P.M. (1983) Word Associations in a Foreign language. *Nottingham Linguistic Circular, 11*, 28-38.

Meara, P.M. (1995) Single-subject studies of lexical acquisition. *Second Language Research, 11/2*, i-iii.

Meara, P.M. (1996) The dimensions of lexical competence. In Brown G., Malmkjaer, K. and Williams, J. (Eds.), *Competence and Performance in Language Learning,* Cambridge: Cambridge University Press. 35-53.

Meara, P.M. (2007) Simulating word associations in an L2: the effects of structural complexity. *Language Forum. 33/2*, 13-31.

Meara, P.M. (2009) *Connected Words: Word Associations and Second Language Vocabulary Acquisition.* Amsterdam: John Benjamins.

Meara, P.M. (2011) Gossamer or bindweed? Association links between common words *EUROSLA Yearbook, 11*, 94-114.

Meara, P.M. and Jones, G. (1990) *Eurocentres Vocabulary Size Tests 10KA*. Zurich: Eurocentres Learning Service.

Meara, P.M. and Fitzpatrick, T. (2000) Lex 30: An improved method of assessing productive vocabulary in an L2. *System, 28*, 19-30.

Meara, P.M., Jacobs, G. and Rodgers, C. (2002) Lexical signatures in foreign language free-form texts. *ITL Review of Applied Linguistics 135-136*, 1-12.

Meara, P.M. and Wolter, B. (2004) V Links: Beyond vocabulary depth. *Angles on the English speaking world, 4*, 85-97.

Mestres-Missé. A, Rodriguez-Fornells. A and Müntea. T.F. (2010) Neural differences in the mapping of verb and noun concepts onto novel words. *Neuroimage 1, 49/3*, 2826-2835.

Miller, G.A. and Fellbaum, C. (1991) Semantic networks of English. *Cognition, 41*, 197-229.

Milton, J. (2009) *Measuring Second Language Vocabulary Acquisition.* Clevedon: Multilingual Matters.

Morrison, R. (1981) Word association patterns in a group of bilingual children. *Unpublished MA project.* Birbeck College, London.

Moss, H. and Older, L. (1996) *Birbeck Word Association Norms.* Hove: Psychology Press.

Mulder, K. and Hulstijn, J. H. (2011) Linguistic skills of adult native speakers, as a function of age and level of education. *Applied linguistics*, *32*(5), 475-494.

Namei, S. (2004) The Bilingual Lexicon from a Developmental Perspective: A Word Association Study of Persian-Swedish Bilinguals. *International Journal of Linguistics, 14/3*, 363-87.

Nation, I.S.P. (1983) Testing and teaching vocabulary. *Guidelines 5/1,* 12-25.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I.S.P. (2001) *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation, I.S.P. (2007) Fundamental issues in modelling and assessing vocabulary knowledge. In Daller. H, Milton.J and Treffers Daller. J. (Eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Nation, I.S.P. (2008) *Teaching Vocabulary: Strategies and techniques.* Boston: Heinle Cengage Learning.

Nation, I.S.P. and Beglar, D. (2007) A vocabulary size test. *The Language Teacher 31/7*, 9-13.

Nelson, D. L., McEvoy, C. L. and Schreiber, T. A. (1998) *The University of South Florida word association, rhyme, and word fragment norms.* Available online at http://www.usf.edu/FreeAssociation/.

Nishiyama, M. (1996) A Study of Non-Fluent Bilinguals' Mental Lexicon by Means of Word Association Test. *Research Memoirs of the Kobe Technical College, 34*, 117-124.

Orita, M. (2002) Word associations of Japanese EFL learners and native speakers: Shifts in response type distribution and the associative development of individual words. *Annual Review of English language in Japan, 13,* 111-120.

Palermo, D. S. (1971) Characteristics of word association responses obtained from children in grades one through four. *Developmental Psychology*, *5*, 118-123.

Pavičić Takač, V. (2008) *Vocabulary learning strategies and foreign language acquisition.* Second Language Acquisition 27. Clevedon: Multilingual Matters.

Pawley, A. and Syder, F.H. (1983) Two puzzles for linguistic theory: nativelike selection and native like fluency. In Richards, J.C. and Schmidt, R.W. (eds), *Language and communication*: London, Longman. 191-226.

Piper, T. H. and Leicester, P. F. (1980) Word association behavior as an indicator of English language proficiency. *ERIC Document Reproduction Service, ED227651.* The University of British Columbia.

Politzer, R. (1978) Paradigmatic and syntagmatic associations of first year French students. In Honsa,V. and Hardman-de-Bautista, M. J. (eds), *Papers on linguistics and child language: Ruth Hirsch Weir memorial volume*, 203-210 The Hague: Mouton.

Pollio, H.R. (1966) *The structural basis of word association behavior.* The Hague: Mouton.

Postman, L. and Keppel, G. (Eds). (1970) *Norms of word associations.* New York: Academic Press.

Racine, J. (2008) Cognitive processes in second language word association. *JALT Journal 30/1*, 5-26.

Racine, J. (2012) Replicating rabbits: Toward a Comprehensive Analytical Framework for Word Association. *Vocabulary Education and Research Bulletin 1/1*, 7-9.

Randall, M. (1980) Word association behavior in learners of English as a foreign language. *Polyglot 2/2,* B4-D1.

Read, J. (1993) The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10/3*, 355-371.

Richards, J.C. (1976). The role of vocabulary teaching. *TESOL Quarterly 10*, 77-89.

Richards, J.C. and Schmidt, R. (2002). *Longman Dictionary of Language Teaching and Applied Linguistics*, 3rd edition, London: Longman.

Riegel, K.F, and Zivian, I.W.M. (1972). A study of inter and intralingual associations in English and German. *Language Learning, 22/1*, 51-63.

Rosenzweig, M.R. (1961) Comparisons among Word-Association Responses in English, French, German, and Italian. *The American Journal of Psychology, 74/3,* 347-360.

Ruke-Dravina, V. (1971) Word associations in monolingual and multilingual individuals. *Linguistics, 74*, 66–85.

Russell, W.A. and Jenkins, J. J. (1954) *The Complete Minnesota Norms for Responses to 100 Words from the Kent-Rosanoff Word Association Test*, Technical Report No.11.

Schmitt, N. (1996) An examination of the behavior of four vocabulary tests. In D. Allan (Ed.) *Entry Points: IATEFL*: Whitstable, 34-39.

Schmitt, N. (1998a) Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*/2, 281-317.

Schmitt, N. (1998b) Quantifying word association responses: what is native-like? *System, 26*, 389-401.

Schmitt, N. (1999) The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16/2*, 189–216.

Schmitt, N. (2000) *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Schmitt, N. 2010. *Researching vocabulary: A vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Schmitt, N. and Meara, P.M. (1997) Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition 20*, 17-36.

Schmitt, N., Schmitt, D., and Clapham, C. (2001) Developing and exploring the behavior of two versions of the Vocabulary Levels Test. *Language Testing,18,* 55-88.

Schoonen, R. and Verhallen, M. (2008) The assessment of deep word knowledge in young first and second language learners. *Language Testing 25*, 211-236.

Shimotori, M. (2013) Conceptual contrasts: A comparative semantic study of dimensional adjectives in Japanese and Swedish. *Unpublished PhD thesis*, Umeå University.

Shin, D. and Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal, 62/4*, 339-348.

Söderman, T. (1993a) Word associations of foreign language learners and native speakers – Different response types and their relevance to lexical development. In B. Hammerberg (Ed.) *Problems, process and product in language learning*. Abo: Abo Akademi.

Söderman, T. (1993b) Word associations of foreign language learners and native speakers – A shift in response type and its relevance for a theory of lexical development. In H. Ringbom (Ed.) *Near-native proficiency in English.* Abo:Abo Akademi.

Sökmen, A. J. (1993) Word association results: a window to the lexicons of ESL students. *JALT Journal, 15/2*, 135-50.

Sowell, J. (2006) Word association testing and vocabulary learning. An investigation of cultural and linguistic influences on the lexicon of non-native speakers. *Unpublished MA dissertation,* Colorado State University, presented at JALT 2007,Tokyo.

Stolz, W. S. & Tiffany, J. (1972) The production of "childlike" word associations by adults to unfamiliar adjectives. *Journal of Verbal Learning and Verbal Behavior*, *11*, 38–46.

Van den Haak, J.M., Dejong, M.D.T and Schellens,P.J. 2003. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour and Information Technology. 22/5*, 339-351.

Van Ginkel, C.I. and Van der Linden, E.H. (1996) Word associations in foreign language learning and foreign language loss. In K. Sajavaara and C. Fairweather (Eds.). *Approaches to Second Language Acquisition.* Jyvaskya.

Wang, H. and Zhang, J. (2012) A study on professional word association patterns in L2 mental lexicon. *British Journal of Social Sciences. 1/1*, 47–75.

Wesche, M., & Paribakht, T. M. (1996). Assessing vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*, 13–40.

Wharton, C. (2011) Changing associations: The effect of direct vocabulary instruction on the word associations of Japanese college students. *Unpublished MA project.* Birmingham University.

White, C.J. (1988) The role of associational patterns and semantic networks in vocabulary development. *English Teaching Forum 26/4*, 9-11.

Wilks, C. and Meara, P.M. (2002) Untangling word webs: graph theory and the notion of density in second language word association networks. *Second Language Research,18/4*, 303-324.

Wilks, C. and Meara, P.M. (2007) Implementing graph theory approaches to the exploration of density and structure in L1 and L2 word association networks. In Daller. H, Milton.J and Treffers Daller. J. (Eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press.

Wolter, B. (2001) Comparing the L1 and L2 Mental Lexicon: A Depth of Individual word Knowledge Model. *Studies in Second Language Acquisition 23*, 41-69.

Wolter, B. (2002) Assessing proficiency through word associations: is there still hope? *System 30*: 315 - 329.

Woodworth, R.S. (1938) *Experimental Psychology.* New York: Holt.

Wray, A. (2002) *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Zareva, A. (2005) Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System 33/4*, 547-562.
Zareva, A. (2007) Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second Language Research 23/2*, 123-153.
Zareva, A. (2010) Multicompetence and L2 users' associative links: being unlike nativelike. *International Journal of Applied Linguistics, 20/1,* 2-22.
Zareva, A. (2011) Effects of lexical class and word frequency on L1 and L2 English-based lexical connections. *The Journal of Language Teaching and Learning, 1/2*, 1-17.
Zareva, A. and Wolter, B. (2012) The 'promise' of three methods of word association analysis to L2 lexical research, *Second Language Research.*
Zeno, S. M., S. Ivens, H., Millard R. T. and Duvvuri R. (1995). *The Educators Word Frequency Guide.* Touchstone Applied Science Association (TASA).
Zimmerman, C.B. (2009) Word knowledge: A vocabulary teacher's handbook. New York: Oxford University Press.

Appendices

Appendix 3.1 Prompt Word List used in replication study

Prompt Word List 1			
stimulus	*response*	*stimulus*	*response*
attach		**rely**	
doorway		**container**	
enjoy		**tolerate**	
cherish		**vast**	
anticipation		**vacant**	
temporary		**brave**	
startle		**genuine**	
human		**express***	
venue		**discovery**	
pathetic		**suspicious**	
cartoon		**recreation***	
concentrate		**tourist**	
truth		**useful**	
serious		**disciple**	
regulate		**exert**	
reactor		**volatile**	
assist		**powerful**	
undertake		**confine**	
limitation		**loyal***	
fragile		**divert**	
multiple		**foolish**	
conductor		**prohibit**	
trend		**thrive**	
beneficial		**hill**	

*omitted from final analysis

Appendix 3.2 Prompt Word List used in replication study

Prompt Word List 2			
stimulus	*response*	*stimulus*	*response*
permeate		**extrapolate**	
rejoice		**wrath**	
audacious		**irascible**	
cringe		**narcissism**	
supplant		**pander***	
pith*		**putative**	
imbibe		**surmount**	
rapport		**interject**	
unfurl		**enigmatic**	
scour		**inept**	
jaunt		**kindle**	
incipient		**noxious**	
painstaking		**miraculous**	
propensity		**tome**	
horst		**profane**	
cloister*		**judicious**	
utensil		**innovate**	
apprehension		**salivate**	
blatant		**dowry**	
enrage		**facile**	
opulence		**ulterior**	
purveyor		**amplitude**	
rostrum		**ensnare**	
gleeful		**boisterous**	

*omitted from final analysis

Appendix 3.3 Statistical analysis of data in the replication study.

A)
The table below reports the mean ranks, the number of responses according to each VKS category and chi-square values (H) as determined by the Kruskall-Wallis test. The H values for both the NNS and NS data were not significant ($p<0.05$). These findings, in direct contradiction to the original study (Wolter, 2001), do not allow us to accept the hypothesis that depth of word knowledge is a key indicator of response type.

Kruskall-Wallis GH07	df	H value	Vocabulary Knowledge Scale				
Non-native Speakers	4	2.13 n.s*	1	2	3	4	5
Mean Rank			7.5	9.5	10.8	11.5	13.3
n			114	69	71	106	349
Native Speakers	4	3.47 n.s*					
Mean Rank			11.3	10.1	9.4	7.1	14.6
n			46	37	28	17	304
*p<0.05							

It ought to be noted that the Kruskall-Wallis test assumes that each cell has >5 items. For the NNS data this requirement was met, although for the NS data some of the cells had <5. As the NS data does not strictly conform to the assumptions made by the test, the H value calculated for the NS data ought to be considered as merely an approximation to the chi-square test.

B)
The table below reports a comparison of the means between NNS and NS groups according to the VKS categories, as determined by the Mann-Whitney test. The U scores at each VKS level indicate there were no significant differences ($p<0.05$) in the ranked data. While on the surface this seems to support the hypothesis that L1 and L2 lexicons are structured in a similar way it ought to be noted that the findings in Wolter's original study were different. In the original study the scores for VKS levels 3 and 5 were significant, indicating that at higher levels of word knowledge the lexicons of native and non-native speakers are structured differently.

Mann-Whitney U test GH07					
	Number of Responses		Mean Rank		U score
VKS	NNS	NS	NNS	NS	
1	114	46	4.3	4.8	9 ns*
2	69	37	5	4	6 ns*
3	71	28	5.5	3.5	4 ns*
4	98	17	6	3	2 ns*
5	349	304	4.8	4.3	7 ns*
*p<0.05.					

The conflict in the findings of the two studies, when analysed using the same statistical tools, leads us to conclude that the method of data collection (or perhaps the data itself) is in some way unreliable.

Appendix 4.1 Prompt Word List 1: used in Noun 1 study

これから出てくる各英単語から連想する英単語を記述してください。例えば、はじめに出てきた英単語が CAT ならば、それから連想する単語は DOG、次に出てきた英単語が PEN ならば、それから連想する単語は PENCIL というように、英語で記述してください。上記のようにこのテストは正誤を問う問題ではありません。自分が連想した英単語を正直に全て記述してください。

student*		member	
body		bank*	
month		moment	
book		money	
car		business	
case		morning	
paper		number	
church		hand	
class		other	
game		child	
mind		person	
eye		police	
staff		price	
family		problem	
food		road	
foot		room	
door		school	
office		face	
head		event	
hour		word	
house		team	
world		time	
letter		study	
line		water	
year		idea	

* items rejected after post-test analysis

Appendix 4.2 Prompt Word List 2: used in Noun 1 study

これから出てくる各英単語から連想する英単語を記述してください。例えば、はじめに出てきた英単語が CAT ならば、それから連想する単語は DOG、次に出てきた英単語が PEN ならば、それから連想する単語は PENCIL というように、英語で記述してください。上記のようにこのテストは正誤を問う問題ではありません。自分が連想した英単語を正直に全て記述してください。

air		heart	
feeling		animal	
baby		science	
character		care	
shop		choice	
picture		relationship	
effort		page*	
design		blood*	
chapter		goal	
officer		model	
chance		environment	
evening		competition	
music		help	
culture		data	
doctor		good	
energy		meeting	
garden		hospital*	
history		difficulty*	
manager		hair	
love		teacher	
style		skill	
horse		space	
size		computer	
town		worker	
sound		window	

* items rejected after post-test analysis

Appendix 4.3 Fitzpatrick's 2007 Classification System

Category	Sub category	Definition	Examples
Meaning based Association	Defining synonym	x means the same as y	collapse – fall
	Specific synonym	x can mean y in some specific contexts	reluctant – unhappy
	Lexical set/context related	x y same lexical set/coordinates/meronyms/ super ordinates/provide context	odd – even
	Conceptual association	x and y have some other conceptual link	voluntary – kind immigration – politics
Position – based association	Consecutive xy collocation	y follows x directly (includes compounds)	classical - music
	Consecutive yx collocation	y precedes x directly (includes compounds)	file – nail
	Other collocation association	y follows/precedes x in a phrase but with other content word(s) in between	specific – disability (specific learning disability) cream – cat (the cat got the cream)
Form based association	Change of affix	y is x plus and/or minus a prefix or suffix	construction- constructive conceived – conceive
	Similar form only	y looks or sounds similar to x but has no decipherable link	label- lapel quote quite
Others	Erratic association	y has no decipherable link to x	involved – brow
	blank	No response given	

Appendix 4.4 Chi-square matrix for 9 randomly selected profiles: Noun 1 study

	1	2	3	4	5	6	7	8	9
1	-	**17.22**	12.62	**27.44**	**33.44**	**19.25**	**23.04**	**39.97**	**15.63**
2	-	-	**22.03**	**20.09**	**98.38**	10.91	**21.47**	12.58	9.68
3	-	-	-	**31.56**	**42.84**	**33.74**	**30.04**	**73.45**	**26.50**
4	-	-	-	-	**41.13**	8.97	**20.22**	**16.31**	**48.02**
5	-	-	-	-	-	**25.24**	**49.88**	**55.07**	**37.47**
6	-	-	-	-	-	-	**17.63**	13.16	**21.00**
7	-	-	-	-	-	-	-	**16.62**	**102.27**
8	-	-	-	-	-	-	-	-	**29.46**
9	-	-	-	-	-	-	-	-	-
Significant values, at 0.05 confidence level, marked in **bold**									

Nine individuals from the 50 in the database were randomly selected. Those profiles that did not fit the requirements of the chi-square test (i.e. the profile had a category with zero responses) were not included and an alternative profile selected. The profiles were generated from responses to 94 stimuli. In order to keep the number of responses in each profile equal the blank/erroneous category was also included. In the analysis each of the nine profiles were compared to the other eight, as shown above. In the matrix the chi-square values marked in bold are greater than the critical value, indicating that the pair of profiles are significantly different. There are only six profile pairs out of the 36 that show no statistical difference.

Appendix 4.5 A note on calculating profile similarity

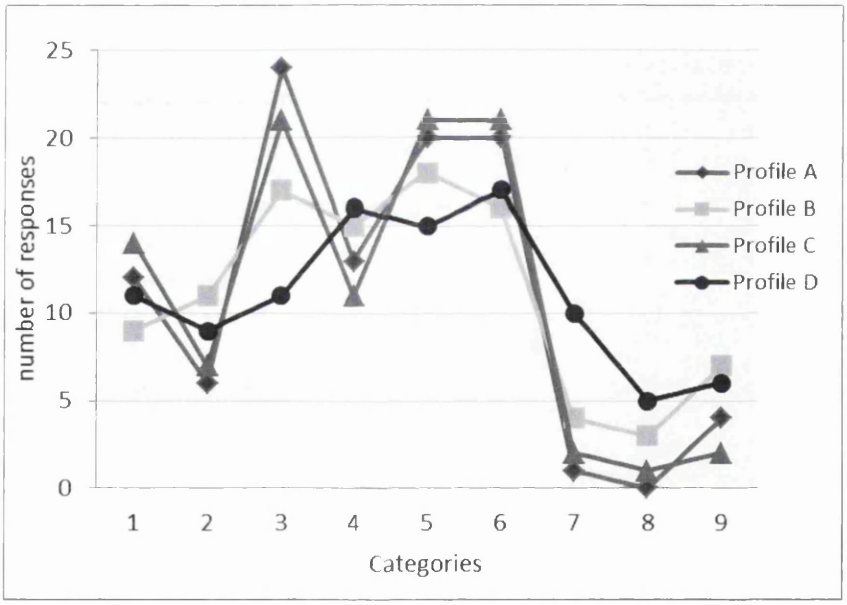
The *Euclidean Distance* and *Pearson Correlation* calculations are both metrics, which can be used to measure the similarity between two arrays. In Fitzpatrick (2007) the Euclidean distance measure was used to calculate the similarity between profiles, in this thesis the Pearson Correlation is preferred.

The calculation for Euclidean Distance is: $\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$.

The calculation for the Pearson Correlation is: $r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$

In Fig 11.1 four 100 item profiles were created to exemplify these two measures. As with the experiments in this thesis, the responses in the example profiles are dispersed over nine categories.

Fig 11.1 Four hypothetical profiles.



With the hypothetical profiles we could argue that profiles A, B and C all have a similar shape; they all peak in varying degrees in categories 3, 5 and 6 and then fall sharply in categories 7 and 8. While Profile D is not completely different to the other profiles, the shape of this profile is in some ways dissimilar; this profile peaks at categories 4 and 6, with category 7 being a fairly important category. The similarity values generated between these profiles, when measured using the two calculation methods, are shown in Table 11.1. In this table the values are ranked, the profile pair with the highest Pearson Correlation value is at the top of the table.

Table 11.1 A comparison of two metrics: Pearson Correlations and Euclidean Distance

Profiles	Pearson Correlation	Euclidean Distance
A & C	0.979	5.099
A & B	0.935	11.576
B & C	0.905	11.832
B & D	0.795	**9.798**
C & D	0.766	16.911
A & D	0.736	18.221

When we compare the values that the two metrics generate, they can be seen to relate quite well. The closest profiles using the Pearson measure are A and C (r=0.979), as we would expect these two profiles also have the smallest Euclidean distance (5.099). As we move down the table, the Pearson's calculation shows weaker correlations and in most cases the Euclidean distance increases, reflecting the increasingly dissimilar profile pairs. In general it might therefore be argued that irrespective of the calculation used a similar conclusion will be drawn about the similarity/distance of the profiles.

The only profile pair in Table 11.1 that have Euclidean and Pearson values that do not correspond are profiles B and D. Using the Pearson Correlation the similarity between profiles B and D are ranked fourth, whereas using the Euclidean distance measure it would rank second. Looking back at Fig 11.1, I would argue that the Pearson calculation better reflects the similarity in the shapes of the profiles. Profiles B and D do not appear to be more similar than profiles A and B (or B and C) as the Euclidean value suggests. The reason for this anomaly is that the Euclidean measure calculates the distance between each category whereas the Pearson calculation looks more at the overall trend. Profiles that have a small distance between points yet a different shape will have a small Euclidean distance but a weak Pearson value. As in this thesis it is the shape of the profile that is of most interest, the Pearson Correlation value is preferred.

Appendix 5.1 Prompt Word List 2: used in Noun 2 study

これから出てくる各英単語から連想する英単語を記述してください。例えば、はじめに出てきた英単語がCATならば、それから連想する単語はDOG、次に出てきた英単語がPENならば、それから連想する単語はPENCILというように、英語で記述してください。上記のようにこのテストは正誤を問う問題ではありません。自分が連想した英単語を正直に全て記述してください。

alternative		mechanism	
plate		metal	
rain		negotiation	
bridge		origin	
circle		output	
soldier		phase	
comparison		justice	
desk		pleasure	
assumption		priority	
construction		expense	
distinction		religion	
youth		revenue	
examination		selection	
magazine		significance	
factory		enemy	
lawyer		tool	
flat		surprise	
fruit		trend	
guest		violence	
human		welfare	
index		wing	
instrument		tooth	
coal		observation	
faith		border	

* items rejected after post-test analysis

Appendix 5.2 Chi-square matrix for 9 randomly selected profiles: Noun 2 study

	1	2	3	4	5	6	7	8	9
1	-	**37.21**	**125.90**	**32.05**	**41.98**	**35.97**	**109.84**	**33.76**	**45.31**
2	-	-	**53.96**	**93.24**	**53.80**	**119.62**	**47.92**	**36.99**	**32.02**
3	-	-	-	**41.45**	**87.77**	**36.58**	9.60	14.98	**51.20**
4	-	-	-	-	**67.41**	**147.05**	**87.26**	**63.84**	**37.81**
5	-	-	-	-	-	**99.29**	**104.87**	**77.29**	**63.50**
6	-	-	-	-	-	-	16.30	**20.61**	**85.38**
7	-	-	-	-	-	-	-	**21.00**	**75.36**
8	-	-	-	-	-	-	-	-	**155.36**
9	-	-	-	-	-	-	-	-	-

Significant values, at the 0.05 confidence level, are marked in **bold**

Nine individuals from the 30 in the database were randomly selected. Those profiles that did not fit the requirements of the chi-square test (i.e. the profile had a category with zero responses) were not included and an alternative profile selected. The profiles were generated from responses to 96 stimuli. In order to keep the number of responses in each profile equal the blank/erroneous category was also included. In the analysis each of the nine profiles were compared to the other eight, as shown above. In the matrix the chi-square values marked in bold are greater than the critical value, indicating that the pair of profiles are significantly different. There are only three profile pairs out of the 36 that show no statistical difference.

Appendix 6.1 Prompt Word List 1: used in Verb study

to call*			**to hold**		
to believe			**to allow**		
to appear			**to know**		
to hear			**to return**		
to become			**to get**		
to keep			**to die**		
to bring			**to leave**		
to meet			**to ask**		
to say			**to find**		
to continue			**to act**		
to decide			**to carry***		
to put			**to talk**		
to develop			**to play**		
to suggest			**to describe**		
to do			**to receive**		
to understand			**to tell**		
to fall			**to expect**		
to offer			**to move**		
to produce			**to show**		
to follow			**to try**		
to see			**to force**		
to turn			**to send**		
to happen			**to speak**		
to feel*			**to take**		
to help			**to think**		

*words omitted from the analysis

Appendix 6.2 Prompt Word List 2: used in Verb study

to advise		**to escape**	
to afford		**to recover**	
to explore		**to assess**	
to blow*		**to disappear**	
to fix		**to fear**	
to recommend		**to generate**	
to attach		**to realize**	
to destroy		**to vote***	
to gather		**to burn***	
to investigate		**to estimate**	
to reject		**to contact**	
to climb*		**to declare**	
to damage		**to rely**	
to promote		**to impose**	
to remind		**to satisfy**	
to secure		**to conclude**	
to connect		**to rest**	
to illustrate		**to hurt**	
to separate		**to shout**	
to influence		**to succeed**	
to shut		**to invite**	
to contribute		**to persuade**	
to organise		**to consist**	
to propose		**to deliver**	
to divide		**to surround**	

*words omitted from the analysis

Appendix 7.1 Prompt Word List 1: used in Adjective study

これから出てくる各英単語から連想する英単語を記述してください。例えば、はじめに出てきた英単語が CAT ならば、それから連想する単語は DOG、次に出てきた英単語が PEN ならば、それから連想する単語は PENCIL というように、英語で記述してください。上記のようにこのテストは正誤を問う問題ではありません。自分が連想した英単語を正直に全て記述してください。

social			**basic**		
national			**wide**		
sure			**appropriate**		
general			**significant**		
particular			**foreign**		
political			**private**		
likely			**recent**		
important			**free**		
public			**individual**		
real			**popular**		
special			**necessary**		
international			**previous**		
different			**natural**		
clear			**various**		
certain			**current**		
available			**concerned**		
useful			**similar**		
modern			**common**		
normal			**professional**		
serious			**original**		

Appendix 7.2 Prompt Word List 2: used in Adjective study

immediate			**average**		
entire			**wonderful**		
familiar			**vast**		
married			**upper**		
bright			**vital**		
reasonable			**external**		
alternative			**official**		
limited			**constant**		
permanent			**corporate**		
perfect			**ancient**		
rare			**bloody**		
apparent			**used**		
criminal			**urban**		
terrible			**mental**		
detailed			**capable**		
attractive			**quiet**		
careful			**odd**		
educational			**elderly**		
severe			**overall**		
sufficient			**chief**		

Appendix 7.3 Adjectives rejected after 1st pilot study

The following adjectives were piloted with 28 Japanese university students on July 14th 2011. Items were rejected if they had >25% primary association to one word.

Prompt Word List 1	Primary association	Prompt Word List 2	Primary association
full	empty 26.9%	academic	school 26.1%
bad	good 57.7%	democratic	democracy 50%
heavy	light 40%	afraid	scary 30.4%
far	near 42.3%	scientific	science 45.5%
great	good 26.9%	equal	same 33.3%
easy	difficult 57.7%	historical	history 37.5%
dark	black 28%	narrow	wide 40%
little	small 38.5	critical	hit 44.4%
new	old 73.1	typical	type 38.9%
possible	impossible 53.8	secondary	second 26.1%
right	left 50%	suitable	suit 38.1%
local	city 26.1%	busy	free 36%
good	bad 52%	northern	south 36.4%
low	high 65.4%	tiny	small 32%
main	sub 42.3%	twice	two 30.8%
small	big 56%	wild	animal 32%
long	short 50%	expensive	cheap 53.8%
big	small 57.7%	lovely	cute 38.5%
early	morning 26.9%	single	double 26%
last	first 30.8%	open	close 64%
high	low 48%	short	long 76%
large	small 38.5%	black	white 80.8%
true	false 34.6%	simple	complex 30.8%
poor	rich 56%	central	city 26.9%

Appendix 7.4 Adjectives rejected after 2nd pilot study

The following list was piloted with a group of 30 Japanese students aged 17 – 18 on September 20th 2011. These words were cut from the list due to a strong association (over 25%) to just one other word or other problems, such as being easily misread.

Prompt Word List 1	Primary association	Prompt Word List 2	Primary association
specific	special 31.58%	financial	money 38.1%
able	can 44%	increased	decrease 57.69
major	minor 40%	thin	thick 46.15%
total	sum 48.15%	leading	leader 33% many responses to *reading*
legal	illegal 50%	initial	32% wrote their initials
personal	computer 36.67%	correct	many responses to *collect*