Second language vocabulary acquisition through dictionary use.

Ronald, James Martin

Second Language Vocabulary Acquisition through Dictionary Use

James Martin Ronald

A thesis submitted for the degree of
Philosophiae Doctor

University of Wales
Swansea

2006

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

DECLARATION

This work has not previously been accepted in substance for any degree and is not 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, if accepted, 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 12th December 2006

Acknowledgements

My thanks, first of all, go to Professor Paul Meara for his insightful guidance, his patience, and his wisdom throughout the preparation of this work. In addition, I am deeply grateful to my wife, Midori, for her unfailing love and support. I would also like to thank my friends, colleagues, and brothers and sisters in Christ, for all their encouragement and help over the years. I am thankful, too, to the many students at Hiroshima Shudo University who gave their time and effort in the collection of the data. Above all, I thank God for making this thesis a reality by putting each of these people into my life.

This thesis is dedicated with thanks to the memory of my father.

Table of Contents

Appendices

Abstract

This thesis investigates the acquisition of L2 lexical knowledge as a result of dictionary use. It main focus is on the value of monolingual learner dictionaries, both in terms of meaning comprehension of looked-up words and in the retention of lexical information. The thesis contains reports and discussions of two sets of investigations into dictionary use by Japanese learners of English. The first set of studies includes comparisons between dictionary entries and written contexts for target L2 words as resources for the comprehension and retention of lexical information, and comparisons between different word classes for these two types of resource. These studies lead us to a recognition of the limitations of traditional investigation methods using a pretest – learning condition – posttest format with small numbers of target words to address many issues central to L2 vocabulary acquisition and dictionary use. Four case studies are reported in which participants encounter very large numbers of targeted words through extensive reading and dictionary use. Participants were also repeatedly tested for knowledge of the items using the self-rating instrument *V_States* (Meara, 2000). This method has produced very rich data showing the effect of dictionary use both on looked up words and for targeted words that were not looked up. Further, the studies have shown the value of matrices derived from these data to produce projections enabling the comparison of the two conditions of reading with and without dictionary use on vocabulary acquisition.

List of Figures

Chapter One: Introduction, Outline, and Overview

1.1 Introduction

Dictionaries are central to the language learning experience of vast numbers of learners of a foreign or second language around the world. They are not only essential for much work in the language classroom, for homework and other language practice, but also for much independent learning outside formal learning environments. While learner dictionaries have long been a widely-used resource for language learners, as these dictionaries become increasingly available via various electronic media, there are signs that they are now being used more than ever, and that their importance is increasing.

Although the different electronic forms of dictionaries are perhaps the most visually striking changes to dictionaries in recent years, more far-reaching and fundamental in many ways has been the use of computer corpora in the production of learner dictionaries. Since the 1980s, beginning with COBUILD, enormous amounts of time, money, and expertise have been devoted to the creation and use of corpora with the aim of producing better learner dictionaries. While this started with English monolingual dictionaries, its effect has spread to bilingual dictionaries, whether indirectly by the widespread practice of adopting information from monolingual dictionaries or as corpus-informed dictionaries in their own right.

In part as a result of the same corpus revolution, there has been growing recognition of the central importance of vocabulary in the learning and teaching of foreign languages.

A long-neglected area of language study, the last two decades have seen renewed interest in many aspects of vocabulary description and acquisition. The insights gained through corpus linguistics – into the importance of collocation, phraseology, and word grammar, among others – are all indicators of the now pivotal position of vocabulary in language description. From the perspective of second language acquisition, the interdependence of words in the mental lexicon, the close relationship between vocabulary size and text comprehension, and the growing appreciation of the importance of word chunking and lexical networks (Wray, 2002; Hoey, 2005) (Wilks, and Meara, 2002) in both retention and production of language all testify to growing recognition of the centrality of vocabulary in second language learning.

The higher profiles both of learner dictionaries and of second language vocabulary studies have also led to greater interest in what dictionary users do with their dictionaries and in issues relating to language learner dictionary use. Especially in the past twenty or so years, important research has been conducted, usually by independent researchers rather than dictionary makers themselves, with the express purposes of investigating how language learners use dictionaries and, consequently, how learner dictionaries can be made to better serve the needs of their users. These researchers have conducted large-scale surveys of dictionary users' habits and preferences (Béjoint, 1981; Atkins, 1998), detailed observations of dictionary use (e.g., Ard, 1982) , and various studies to gain a better understanding of what helps or hinders language learners' use of learner dictionaries (Meara and English, 1988; Nesi, 1998).

One area that has not been the object of the same degree of interest as either vocabulary acquisition or dictionary use is that found at the intersection of these two disciplines: L2 vocabulary acquisition in the context of dictionary use. This is the topic of this thesis, and the primary focus of all the research reported in the thesis. Despite the growth and activity in the two fields, this intersection has remained relatively quiet, with comparatively little vocabulary acquisition research directly concerned with dictionary use, and little research into dictionary use focusing specifically on vocabulary acquisition.

There appear to be three related causes for the relative paucity of research in this area. One cause generally applicable to studies of dictionary use is that research into language learner dictionary use is not straightforward. As Nesi (2000: 1) notes, dictionary use is essentially a private activity and, as a result, observation methods will tend to affect the behaviour of dictionary users. This issue has been partly sidestepped in some cases by using computer-based dictionaries that record users' look-up behaviour but, especially until this medium becomes the norm for the dictionary users being investigated, the validity of such data cannot be accepted unquestioningly. A second cause is that vocabulary acquisition resulting from dictionary use is even more private than dictionary use itself, happening as it does within the dictionary user's head. Such data are only accessible, and then only to a limited extent, by methods that may be even more intrusive than those used simply for observation of dictionary use.

A third reason for the apparent lack of interest in L2 vocabulary acquisition through dictionary use may be that the role of dictionaries in second language learning environments has not generally been seen as one primarily related to vocabulary acquisition. Rather, dictionaries are more often perceived as tools to assist in the achievement of other language learning related tasks or goals, such as the comprehension of L2 texts or the writing of essays in the L2. On a more local level, definitions or translation equivalents of specific words in a text are consulted when reading with the primary aims of understanding what those words mean and of gaining greater global understanding of the text rather than with the goal of learning these words. In the case of L2 writing, where the goal is to produce a meaningful text, additional reasons for dictionary use may be to find guidance in how to use a word, to identify collocates, or to check the register or variety of the word. Here too, however, stated learner goals are generally to find, understand and use specific information in dictionaries, not to learn that information.

Surveys of learner dictionary use appear to confirm the importance of the above purposes for the majority of language learning dictionary users, and the apparent marginality of considerations of vocabulary acquisition. This issue may not, however, be as decisive or straightforward as survey results suggest. As Tono (2001: 67) notes, survey data are, in a sense, indirect in that respondents only report how or why they think they use dictionaries. In addition, survey data may also be distorted by the choice and style of questionnaire questions and even by the act by those surveyed of declaring particular knowledge, beliefs, or behaviour to people who they perceive as having expertise superior to theirs in the field under investigation.

One reason that survey data do not show vocabulary acquisition as an important reason for language learner dictionary use may be simply that this is not offered in questionnaires as a reason for dictionary use. In a survey of Japanese students' bilingual dictionary use, for example, Tono (2001: 235) offers respondents a choice of reasons for using dictionaries. The only reasons the list comprises are checking or looking up various types of word information, such as pronunciation, meanings, spelling, or grammar. Vocabulary acquisition is not offered as a reason for dictionary use. This is also true of Atkins and Varantola's EURALEX/AILA survey of dictionary use among European learners of English (1998: 21-81). In this case, admittedly, an *"Other (please specify)"* box is offered for reasons other than those listed (1998: 56), but since "learning a word" is of a different order of reason than "understanding a word", "finding a translation" or "checking how to use an English word", the questionnaire environment would discourage respondents from volunteering vocabulary learning as a reason for their dictionary use.

Beyond this brief consideration of dictionary use survey data lie two questions that are central to the study of language learner dictionary use:

i) Does the L2 vocabulary of language learners benefit from dictionary use?

ii) Do language learners consciously use dictionaries as language learning tools?

As stated above, this first question is, basically, the topic of this entire thesis. It will be considered in Chapter Two, as we review previous research into vocabulary acquisition through dictionary use, in Chapters Three to Ten which report various studies conducted by the author, and in Chapter Eleven as we reflect further on the benefit of monolingual learner dictionaries to language learners.

As for the second question, there are various kinds of at least partial evidence supporting the hypothesis that vocabulary acquisition is an important factor in the L2 dictionary use of language learners. One type of evidence of this purpose is to be found in the dictionaries of language learners. The widespread practice among learner dictionary users of underlining or highlighting words that they look up is an indication of their desire to retain a mental record of words they have looked up, or of words they have looked up more than once. With electronic dictionaries, the presence of a "history" function by which users can review recently looked-up words, and the fact that this function is widely used, also suggests that language learners are eager to learn at least some of the words which they have looked up. The presence and use of this function also reflects language learner recognition that consulting a dictionary entry only once may not be sufficient to ensure retention of word information found in that entry.

Further evidence of conscious vocabulary acquisition can be observed in language learners who pay attention to information in a dictionary entry that is not required for whatever purpose they are looking up the word. This occurs in Laufer and Hill's study (2000), discussed in more detail in Chapter Two, in which some learners using electronic dictionaries for reading comprehension would also deliberately investigate the pronunciation of the targeted words. This behaviour could be attributed to curiosity in the technology or in the words themselves, but even this curiosity to know more about a word may be seen as a desire to gain more vocabulary knowledge.

One further sign of language learners' desire to learn the words they look up is to be found in their use of monolingual learner dictionaries (henceforth MLDs). There may be a combination of reasons for their use, such as the belief that the information MLDs contain is more accurate or up to date than that found in bilingual dictionaries, or that these dictionaries represent and reflect the culture of the target language in a way that bilingual dictionaries do not. However, one important factor leading more advanced learners to use an MLD, despite the greater time and effort involved, is that users believe that their use increases the probability of looked-up words being retained. This belief is expressed even by learners who choose not to use an MLD, expressed as "I know I should, but..." (Tomaszczyk, 1979). Given the high rates of comprehension failure among MLD users (see, for example, Bogaards (1998) or the studies reported in Chapters Three to Six of this thesis), language learners' faith in these dictionaries may not always be justified, but the fact of this faith is unquestionable.

Finally, the learning (i.e. retention) of language is the central goal of most learners of a second or foreign language. The learning process may be largely composed of short-term tasks or goals that are not expressed in terms of language learning, such as the comprehension of a reading passage, the writing of an essay, or the completion of some other task. However, behind or beyond these short-term goals there is usually a language learning purpose that is expressed linguistically or even in terms of vocabulary acquisition: that some new language or level of language proficiency is left with the learner after the task has been completed. Dictionary use in these contexts, too, clearly has some kind of purpose in terms of language retention, however much this may be unstated or taken for granted.

1.2 Outline

This thesis is composed of four main sections. The first section is a review of research into the field of L2 vocabulary acquisition, beginning with an overview of the wider area of successful dictionary use by language learners and followed by a critical review of fourteen studies investigating L2 vocabulary acquisition through dictionary use.

The second section, spanning Chapters Three to Six, consists of reports of four studies conducted by the author into the comprehension and retention of lexical information found in monolingual learner dictionary entries. In these studies, the effectiveness of whole MLD entries, or of definitions alone, as sources of information and media for retention is compared with that of contextual information of target words as found in written texts or in example sentences. In addition to L2 word knowledge comprehension and retention, other issues investigated are comprehension and retention rates for words in different word classes and the development of more sensitive instruments to record the retention of lexical information.

The third section, from Chapters Seven to Ten, comprises reports of four longitudinal case studies. These studies deal with L2 vocabulary development in the context of the repeated reading of long L2 texts, both without and then with the aid of MLD use. A further aim of these studies is evaluate and develop an instrument for recording, predicting and comparing participants' changing knowledge of large numbers of targeted L2 items through the two learning conditions under investigation.

The final section, in Chapters Eleven and Twelve, is a review and discussion of issues that arise through the studies reported in the other chapters of the thesis but which have not received much focused attention. Issues include low entry comprehension rates for MLD users, widely varying rates of comprehension and retention for different targeted words, and widely ranging results for apparently comparable participants. These issues lead us to consider means by which rates of successful MLD use may be improved, approaches to select and test comparable sets of target words, and methods for investigating and taking account of language learner confidence regarding word knowledge as an aspect of L2 word knowledge retention.

1.3 Overview

In the context of language learning, successful (Christianson, 1997), efficient (Bogaards, 1991: 94), or effective (McKeown, 1993; Wingate, 2002: 1) dictionary use can mean a variety of different things to different people, and it not always clear what these terms are being used to mean. The purpose of this overview is to provide a clearer perception of the different types or stages, of success in dictionary use, and a better understanding of how they are related. As we unpack these different concatenated meanings of successful dictionary use in the context of second language learning we will also gain an overview of research in this area.

At this point, it is worth pointing out that while dictionary entries contain a wide range of types of word related information, including meaning, collocation, grammar, register, or pronunciation, the majority of research has focused primarily on word senses. For this reason, and also because word meanings are typically felt to be of greatest importance to language learners (Tomaszczyk, 1979; Béjoint, 1981; MacFarquhar and Richards, 1983; Harvey and Yuill, 1997), this overview will largely be limited to the same sphere. However, much of what is reported here will be relevant to whole dictionary entries, to whole senses within words, and to different types of word information.

As Nesi (2000: 68) notes, the first important step, and skill, for the language learner prior to reaching for a dictionary is that of identifying words which he or she would benefit most by looking up. From another perspective, this skill could be described as the ability to judge whether consulting a dictionary is the best response when encountering an unknown or partially known word.

Once the decision has been made to look up information in a dictionary, an increasingly important and relevant step is to select the dictionary best suited to the specific requirements of the language learner, to the linguistic ability of the language learner, and to the word in question. Until recently, such questions were irrelevant to the great majority of language learners, since they would inevitably turn to the one dictionary in their possession. In the past few years, however, with the advent of multiple volume full content electronic dictionaries, increasing numbers of language learners may start the look-up process by making a selection from a number of bilingual and monolingual dictionaries. Selecting wisely is one type or aspect of successful dictionary use.

Beyond the decisions of whether or not to use a dictionary and, increasingly, which dictionary to use, finding the entry for the word in question could be said to be the most basic level of dictionary use success. Researchers do not usually record success rates for finding the right entry, with some (e.g., Luppescu and Day, 1993; Fischer, 1994; Hill and Laufer, 2002) dispensing of this stage altogether by creating mini-dictionaries, whether paper or electronic, which only contain the words targeted for their research. Despite this apparent assumption that success in finding a given word may be taken for granted, success rates may for various reasons fall below 100%. Reasons include misreading or mistyping the word, mistakenly looking at the dictionary entry for a different part of speech for the word, or finding the right page but reading the wrong entry by mistake. These problems may occur for both paper and electronic dictionary users. Confirmation that the ability to find a dictionary entry cannot be taken for granted may also be found in the sections in dictionary workbooks that offer extensive explanations and practice for finding words and phrases in the dictionary (e.g., Goodale, 1995; Komuro, 2004).

In the case of multi-word items such as idioms, the task of finding the right entry is understandably more complex and has been the specific focus of some dictionary use research, such as that reported by Béjoint (1981), by Atkins and Knowles (1990: 388) and by Tono (2001: 116-142). Simple success rates in these studies are not given. Rather, we are shown that dictionary users often start looking for idioms in places other than where they are found in the dictionary. In terms of efficient dictionary use, as these authors suggest, this has implications both for dictionary design and for dictionary use training.

The next stage, finding the correct sense within the entry, is one for which we might expect lower success rates. Despite being an issue of direct relevance to dictionary makers (see, for example, the discussion by Kipfer (1984: 101-108)), few studies have been conducted into how successful dictionary users may be in finding the senses they are seeking within dictionary entries for polysemous words. This may be because research concerned with definition comprehension or lexical retention has tended to focus on previously unknown single-sense words; from a research perspective, these are easier to deal with than specific secondary senses within entries for polysemous words.

Bogaards (1998) does address the issue of secondary sense location and considers the nature of the language learner's task of finding a particular sense within a long dictionary entry. In his study, success rates in finding the correct sense ranged between 67% and 82% for different monolingual learner dictionaries. In a study comparing electronic and paper dictionary use, Ronald and Tajino (2005) also noted that although single sense entries may be located more quickly by electronic dictionary users, this advantage appears to be largely offset by the typically longer time taken by electronic dictionary users to locate specific senses within polysemous MLD entries. Tono also considers this question of sense location as he investigates the value within dictionary entries of sense "menus" (2001: 167-173) and "signposts" (2001: 174-183) as guides to the sense sought by the user. Neubach and Cohen (1988), too, in their L2 dictionary user research, consider the challenge faced by dictionary users of finding the relevant sense within entries for polysemous words.

Once a particular dictionary entry or sense has been found, the next step for the user is comprehension of the information sought. In the case of bilingual dictionary use this is, again, often taken for granted, although perhaps with less justification than for finding the dictionary entry itself. For bilingual dictionary entries, one difficulty may be understanding the information provided in one's own language, as research into monolingual dictionary use with young children (Miller and Gildea, 1985) has shown. Another difficulty is choosing correctly from a list of translation equivalents (Iannucci, 2003: 217-229). For monolingual learner dictionary users, comprehension is rarely taken for granted, and various studies have revealed remarkably low comprehension rates for definitions. In Mullich's research (1990), as many as 50% of task responses were not completely successful, much of which may be attributable to a lack of comprehension. Bogaards (1998), in his above-cited study of polysemous words, records incorrect or only partially correct comprehension rates as ranging between 30% and 40%.

There are two further types of success in dictionary use other than those listed above. The first of these is the accurate use of particular word information found in dictionaries, such as word grammar, collocation, or register. This is often described as productive dictionary use, although it should be noted that in many of the studies investigating this (e.g. Miller and Gildea, 1985; Fischer, 1994; Nesi, 2000: 84-93, and those described in Chapters Three and Four of this thesis), the dictionary users' only real purpose for production in the studies is to provide researchers with evidence of comprehension and correct use of the targeted information from the dictionary. In studies with foreign language learners using monolingual dictionaries (Fischer, 1994; Nesi, 2000: 84-93), average rates of acceptable productive use of the target items were only around 40%. From the perspective of dictionary skills training, Carduner (2003) also investigates productive dictionary use, with students proofreading their own written work with the help of a dictionary. No data are provided regarding the effect of dictionary training other than student impressions, but this approach does offer a promising alternative to the widely used LUCAS (look up, compose a sentence) method.

For many language learners, the retention of looked-up information is, arguably, the ultimate goal in terms of successful dictionary use. Retention of looked-up information is, however, dependent upon prior success in the stages of locating and understanding the information sought. As such, in this thesis, while the primary focus of our attention will be the retention of lexical information via MLD use, our overall concern throughout this thesis will be successful dictionary use by language learners.

Chapter Two: Literature Review

2.1 Introduction

In Chapter One we looked at different interpretations of successful L2 dictionary use. These included deciding wisely when to use the dictionary, finding the entry or sense, comprehension of the entry, production of language informed by dictionary use, and retention of lexical information gained through dictionary use. The final type of successful dictionary use, retention of lexical information, is the focus of this chapter. Specifically, we will look at research which investigates second language vocabulary acquisition through dictionary use. The main focus of studies in this area has been on the effect of dictionary entries, specifically definitions or translation equivalents, on knowledge of previously unknown L2 words. Since example sentences also feature as part of dictionary entries in most learner dictionaries, the effect of these, both in isolation and as part of a dictionary entry, has also been a focus of research. In general, research has been conducted by comparing groups of language learners with different learning conditions: with or without access to dictionaries while reading an L2 text, with different types of dictionary entry for targeted words, or with different parts of dictionary entries.

Related to dictionary use are marginal glosses and their use with reading passages. The information provided in glosses may be similar to that found in learner dictionaries, providing either translation equivalents or L2 definitions for targeted words in L2 reading passages. Because of the similarity of using glosses to dictionary use, one early investigation into the effect of glosses on L2 vocabulary acquisition has been included in this review, as has one study comparing the effects of glosses and dictionary use. In general, though, our focus has remained on dictionary use, in which information about targeted words in a text is not part of that text, as glosses arguably are, but is found in a separate book, booklet, sheet of paper or, for electronic dictionaries, screen or window.

As we saw in Chapter One, one type of learner dictionary use closely linked to vocabulary retention is the comprehension of previously unknown words with the help of learner dictionaries. Obviously, comprehension is a pre-condition for retention of meaning of such words, and many studies of lexical retention reviewed below do also include measures of comprehension. Of itself, as we have discussed, comprehension of unknown words cannot be equated with learning these words. For this reason we will focus mainly on studies which investigate retention of knowledge of words after their encounter in dictionaries and whatever other learning environments are under investigation.

A major problem faced by almost all research into dictionary use has been that dictionary use is usually a private, personal activity, and attempts to observe, measure or dictate its use will inevitably affect its frequency, its nature, and the attention paid to noticing or learning looked up words. Coupled to this have been the issues faced by vocabulary acquisition research generally of how to measure and compare language learners' knowledge of targeted words before and after the application of a learning condition. A third challenge, especially where incidental vocabulary acquisition is concerned, has been how to predict which words will be looked up by language learners in order to test the effect of dictionary use on knowledge of these words. In our review of the literature, we will look at the different ways in which researchers address the above challenges.

2.2 Focus

The issues touched upon above will be the main focus of our literature review. In brief, they are:

- Under what conditions does dictionary use take place? How is dictionary use recorded?
- How is prior knowledge of targeted words identified without affecting the acquisition of these words?
- What methods are employed to measure word knowledge? How sensitive are these measures?
- How does the number of targeted words, looked up words, and learned words affect the value of the studies?

In addition to the above concerns, where relevant, we will also consider the types of words targeted or looked up and the effect on learning of encountering them in a written context as compared with looking them up in a dictionary. Finally, it should be noted that in less than half of the studies reviewed are normal paper dictionaries used. In some cases, only short definitions or translation equivalents are provided, in specially prepared booklets. In others, various forms of electronic dictionaries are used, in which requested dictionary entries appear on a computer screen. While this is not a major focus of this review, where relevant we will consider how dictionary format may affect dictionary-using behaviour and resultant vocabulary acquisition.

2.3 Review of selected studies

In this section we will review, in chronological order, a series of studies that have investigated second language vocabulary acquisition through dictionary use. There are two reasons for this chronological approach. Although these studies span a period of over eighty years, there are distinct echoes, even in recent papers, of methods and approaches employed in much earlier studies. Despite differences between researchers in terms of generation, nationality, native language and target language, this review has assembled what could be called a research community: a community which, through the papers reviewed here, talks to each other and learns from each other. Especially, it is a community in which newcomers have learnt from their elders. This is one important reason justifying the adoption of a chronological approach here.

Another reason for the chronological order of papers reviewed is that there are problems with the one other obvious way of sorting and presenting this research. This is the division and consideration of papers according to whether they claim to be investigating intentional or incidental vocabulary acquisition. In some research, the distinction between these two types of acquisition are presented as cut-and-dried, and in the case of extensive reading, for example, there is some justification for this. However, with dictionaries it could be argued that all attention paid to individual words is partly form-focused, and that all such focus on language in formal learning or research contexts will be, in part at least, motivated by the dictionary user's desire to learn the looked up words.

For each study reviewed in this chapter, the review will consist of a short summary of the study followed by comments which address the questions outlined above. Within this group of studies, the approaches vary widely, as do the circumstances in which the research is undertaken. For some of the studies, the focus is quite deliberately from a teaching perspective, on the role that dictionaries can play in vocabulary-building. For others, there is a strict hands-off approach in investigating, as far as is possible, incidental vocabulary acquisition through extensive reading and natural dictionary use.

With the exception of one early study (Seibert, 1930), the scope of this review is limited to studies in which dictionaries, or information from dictionaries in a self-contained format, are used. It is also limited to studies, with the partial exception of Fischer (1994), in which learning from dictionaries is investigated, as opposed to dictionary use or word comprehension alone.

2.3.1 Grinstead (1915)

(Grinstead, W.J. (1915). An experiment in the learning of foreign words. *Journal of Educational Psychology 6*, 242-245.)

Summary

In this very early single subject case study focusing on foreign vocabulary acquisition through dictionary use, the author compares two approaches: looking up unknown German words encountered while reading a German text and looking up isolated unknown words from a list. The author was the English-speaking subject of his own study. Two sets of words were composed of 17 words each encountered during a one-hour reading session of a German text. Words included in the sets were not known to the subject and could not be guessed from context. These words were looked up in a dictionary while reading. The 17 words represented about half of the unknown words encountered in the text of around 300 words per session. Two further sets of words were composed of words encountered only in word lists, although there is little precise information about these. Lists of German words were compiled by an assistant who was proficient in German. The subject then looked over the list and deleted all the words which he already knew. He would then look up the first 17 words remaining on a list. He also did this for two sets of words.

The sets of words from the two learning conditions were tested in the same way. A short time after the completion of the reading or looking up session, those words unknown to the subject prior to the session were read out to him. He would, as far as he was able, give a meaning for each of the words. If he was not able to, he would be told the meaning. For this reason each test is referred to as a presentation of the target words: as much a learning session as a test of word knowledge. The same set of words would be tested again in the same way 24 hours later.

The results are shown in Table 2.1. There were two sets of 17 words for each learning condition. For the word list set, 15 of the words were correctly identified in the first test and all 17 in the second test. For the second set of words in context, 13 of the words were correctly identified in the first test and 15 in the second test. For the first word list set, 14 words were known at the first test, rising to 15 of the 17 at the second test. The second word list scores were 13 at the first test and 14 at the second test.

Table 2.1

Results of Grinstead's (1915) test (total of 17 items per set)

	Test 1	Test 2	Gain
Context set 1	15	17	2
Context set 2	13	15	2
Word list set 1	14	15	1
Word list set 2	13	14	1

The author's conclusion, based on the above results, is that words encountered in context then looked up were remembered better than words only encountered in word lists and then looked up. We will now consider what objections there may be to arriving at such a conclusion.

Comment

This study is valuable principally for the way it addresses various issues related to the effect of dictionary use and encountering words in context on vocabulary acquisition. Its main finding was that there will be better retention of words encountered in context then looked up than of words only encountered in word lists then looked up. This does need further investigation. From the perspective of dictionary use behaviour we can see how the two types of encounter may affect dictionary use and, as a consequence, retention. Where a word has been encountered in a written context, the language learner may first hazard a guess at the meaning the word may have in the context. The learner then looks up the word, seeks to identify the sense of the word in the dictionary entry which seems most suited to the context, and to focus on this. In this procedure, the role of the dictionary is often at least partly confirmatory, with the matching sense focused on and all other senses ignored. For a word encountered in a list, the dictionary user's attention will not be focused on a particular sense in the dictionary and neither will the dictionary user be relating the information in the dictionary entry to that in a given context. It does appear likely, then, that the focus on a specific sense within a dictionary entry would favour the learning of that one sense, as compared with a less focused reading of the whole dictionary entry, and so may help retention where this is tested by the ability to supply one sense for each of the listed words.

Although Grinstead claims that the results are clearly and definitely in favour of learning looked up words encountered in context rather than in a list, this claim is based upon at least two premises which are open to question. One is that that looking up an unknown word in a dictionary will always be successful in terms of location and comprehension of the relevant entry and sense. The other premise is that retention of this word knowledge will be affected by the following factors: the circumstances of the original encounter, the time elapsed since the encounter, and the number of encounters.

Subsequent research has shown that dictionary use is not infrequently unsuccessful, and that we cannot be sure that all 17 words that were looked up were correctly identified in the dictionary and understood. Also, in addition to the factors affecting retention identified by the author, the types of words being learned must also be taken into account. For example, at least some of the set of unknown words encountered in the text will probably share the same semantic field, while there is no reason to expect this from a supposedly random list. The typical number of senses per word in each set of words will also probably differ as may the number of initially unrecognised cognates or, for example, the number of abstract nouns. Any of these factors could affect retention scores to the extent of the one or two words' difference between learning conditions in Grinstead's research. Since there is only a very small difference between the two sets of scores, and no statistical analysis, this difference could arguably be as easily attributed to any of the above factors as to the circumstances in which the words were looked up. In addition, since all results are close to 100%, there is an obvious ceiling effect obscuring possible differences between the two learning conditions.

As Grinstead points out, the experiment described is relatively easy to perform. He suggests that teachers can use this method to learn about the processes involved in vocabulary learning. While it is important to develop practical, workable testing instruments, it is also clear that, as it stands, any results from an experiment of this kind would need to be treated with caution.

2.3.2 Seibert (1930)

(Seibert, L. C. (1930). An experiment on the relative efficiency of studying French vocabulary in associated pairs versus studying French vocabulary in context. *Journal of Educational Psychology 21* (4): 297-314.)

In this pioneering study, Seibert investigates the difference in vocabulary retention of previously unknown foreign words resulting from two different types of language data:

one is a list of L1-L2 word pairs, and the other is words presented in explanatory sentences. This research is not directly about dictionary use, since no dictionaries were used, but the two types of information under investigation mirror those found in bilingual and monolingual learner dictionaries: translation equivalents and example sentences. The methods employed and the findings of this early study into vocabulary acquisition are, therefore, clearly of relevance to the use of dictionaries in the context of language learning.

Summary

The paper reports a study in which 60 English-speaking students of French learned four sets of 12 French words through studying different types of information about the words. There were two basic learning conditions: learning from a list of French and English paired translation equivalents, and learning through sentences, one for each word, in which the meanings of the 12 words are presented in a very transparent way, and for which the L1 equivalents of the target words are also given. In addition to these two conditions, two other hybrid conditions were investigated, in which both sources of information are studied by the subjects. Here, we will look mainly at the data for single learning conditions.

Groups of students were presented with one set of materials for a set of words and asked to study the words for nine minutes. Instructions as to how the words were to be learned are very precise: they were to be studied aloud and, for the sentence condition, subjects had to read whole sentences each time they focused on a word to be learned. All 60 students experienced each of the four learning conditions for different sets of words.

The students were informed that they would be tested following the learning session, and that there would be two types of test. First, there would be a no context test in which the students would be presented with the list of 12 English words from the learning session and asked to give the French equivalents for these words. This would be followed by a context test in which the students would be presented with 12 French sentences in which the target words had been substituted for their English equivalents and would, again, be asked to provide the French words.

There were a total of four testing sessions for each learning condition: 50 minutes after the learning session, 2 days after, 10 days after, and 40 days after. As shown in Table 2.2, results for these tests show higher scores for the word pairs condition for both test types for all four testing sessions.

Table 2.2

Word pair test results of Seibert's (1930) test.

Results are combined for each condition. Figures are shown for a total of 48 items.

	t1	t2	t3	t4
Paired words	33.5	31.6	31.9	24.9
Words in sentences	28.2	28.1	27.4	20.8

For the tests 50 minutes after the learning session (t1), the word pairs condition responses showed an average of 33.5 out of 48 correct answers, while the sentences condition responses showed about 28 correct answers. Two days later (t2) the sentence condition responses at just over 28 were largely unchanged, while for the word pairs condition, correct responses fell to 31.6. There is little change in either condition at the 10-day stage (t3), although the sentence condition averages fall slightly to 27.4. At 40 days (t4), the word pairs condition responses fell to just below 25 correct answers and the sentences condition responses fell a similar amount, although proportionately more, to just under 21 correct answers. Context test results for the two conditions over the 40 days largely mirror these for the paired words tests, with the exception that attrition rates are slightly slower for both conditions with the contexts tests.

Comment

This early study bears surprising similarities to later studies into vocabulary acquisition through dictionary use, except that in this study there is no direct reference to dictionaries. Despite this, there are clear parallels between the two sources of information used in this study and those found in bilingual and monolingual learner dictionaries. The provision of word meanings through paired English and French words is basically the same as that found in bilingual dictionaries, while the explicit sentences for demonstrating word meaning within an L2 context are, to judge by the illustration provided, halfway between the sentence definitions found in some monolingual learner dictionaries and the example sentences used in both bilingual and monolingual learner dictionaries.

Seibert suggests that all the example sentences used in the learning sessions provide a context in which it is easy to identify the meaning of the target word. This may or may not be so but, in any case, as the English equivalent is also provided, we may assume that all of the target words were understood. Why, then, would subjects perform less well in this condition, when the main difference is that additional information is provided? In other words, how can we explain circumstances in which extra information seems to impede the learning of the target words? If we consider the nature of the tasks presented by the two learning conditions, we may be better able to answer this question.

For the learning session involving word pairs, the task is clear: to focus on each word pair, create some mental link between the English and French word, and memorise the form of the French word. For the sentences condition, although the same approach could be made as for the word pairs, the sentence itself may distract the students and diffuse the focus of their attention. This would mean that, to use the example in Seibert's paper – '*On met* <u>*le mors*</u> *dans la bouche du cheval*' – '*mors*' may be linked with '*met*', '*bouche*' and '*cheval*'as well as with *bit* and, possibly, *put*, *mouth* and *horse*. Further, time may also be spent in making sense of, or even translating, the example sentence, rather than memorising the target words. In any case, as the task of the student faced with the sentences is less focused than that for the student with just word pairs to learn, the nine minutes allowed for learning the 12 words, 45 seconds per word, is less likely to be spent efficiently.

One further stated aim of Seibert's aims study is to examine students' ability to actively use the words – but in fact she never tests this. The context test does not do this; it only requires the testee to provide the French words to fit the sentences provided. If she informed the students that ability to use the words correctly would be tested, this might have led sentence condition students to misguidedly spend time studying the words' contexts at the expense of focusing on learning the forms and meanings of the words.

Two failings of many later studies, small number of targeted items and of subjects, are much more adequately treated here, since there are 60 subjects and a total of 48 test words. There is no measurement of pre-knowledge of the words in this study, but the potential problem of differential levels of lexical pre-knowledge between groups is averted through rotating groups of subjects, learning conditions, and sets of test items. All sixty subjects experience all the learning conditions with different sets of words, and these sets of words are rotated among the learning conditions. This also means that there were more test items involved than might first appear. Although each word set only consisted of 12 words, since four sets of words were used for each condition this increased the total number of test items for each condition to 48 words. A disadvantage of this method is that little analysis of the data is possible beyond the sum or average number of words known out of 12 test items for any condition or averages for the 48 words in each condition, as given in Table 2.2 above.

The overall findings of this study are that learning words in L1-L2 word pairs is more effective than learning through focusing on the textual environment of target words. In this study, as indeed in many subsequent studies, at least part of the difference between the two conditions can be attributed to causes other than the learning conditions under investigation. In this case, the instructions about how to learn the words and the suggestion that in the test following the learning session students would need to produce the words in context might have led students in the sentence condition to use their learning time in a more inefficient way than the students in the word pairs condition.

Despite the above limitations of this study, it is constructed with a degree of care found in few other studies in the field, and from a perspective that would remain largely neglected for almost seventy years.

2.3.3 Black (1986)

(Black, A. (1986). The effect on comprehension and memory of providing different types of defining information for new vocabulary: a report on two experiments conducted for Longman Dictionaries and Reference Division. Cambridge: MRC Applied Psychology Unit.)

This paper reports two experiments into dictionary use: one focusing on comprehension of words through dictionary definitions and example sentences and one concerned with retention of knowledge about unknown words for which the different types of information had been provided. As our main interest is with vocabulary retention rather than with comprehension, we will only focus here on the second of the experiments, that concerned with L2 vocabulary acquisition.

Summary

For Black's experiments, low-frequency, presumably previously unknown words were encountered by readers in the context of three short L2 texts. There were four learning conditions investigated, based on the following four sources of information about the words: the L2 texts alone; the L2 texts together with sets of made up example sentences for each word; the texts and dictionary definitions for the words; and the texts, dictionary definitions and example sentences. The four learning conditions were rotated among subjects in a way similar in some respects to that used by Siebert (1930).

A total of 24 advanced level learners of English participated in this study. Their average age was 22, and they had studied English for over six years on average. They were all studying at a language school in Britain for the Cambridge Proficiency Exam. Their language backgrounds were German, Romance languages and Asian languages.

Three texts described as difficult, ranging in length between 306 and 463 words, were used in the experiment: two narrative passages and one expository passage. A total of 24 words, eight from each text, were selected as target words. Suitable words were selected by asking similar level students to underline unknown words that they felt they could not easily infer from the texts and which they would need to look up. The 24 words were randomly divided into four sets of six words, with the four learning conditions rotated among the four sets of words. This means that for each word there were six subjects for each learning condition.

The information for each word was written on a separate card: with definitions, example sentences, or definitions and example sentences for the three experimental conditions. For control condition items, only the targeted word was on the card. For polysemous words, the relevant type of information was provided for each sense.

The subjects all took part in the experiment in one room. They were given the three texts and sets of cards and told to read the text in their own time. They were also told that a comprehension test would follow. When the participants were ready, the information cards were collected and the participants were given a test booklet. During the test, the subjects kept the texts and were allowed to refer to them.

The test consisted of a forced choice multiple-choice vocabulary test for the 24 target words, testing some aspect of the meaning or pragmatic value of each of the words. For each multiple-choice question there was one correct answer and three distractors. In addition to choosing an answer for each question, subjects were asked to rate, on a scale of 1 to 5, how confident they were about the correctness of their answer. However, no details of this potentially important data are provided in the report.

The results of the test are as follows for correct responses for words in the four learning conditions: text only, 48%; text and combined definition and example sentences, 64%; text and definition, 67%; and text and example sentences, 67%. We will discuss below what these figures represent and what implications they may have for the comprehension and retention of previously unknown L2 words.

Comment

Although in some respects Black's research is carefully planned and conducted, there are serious problems with at least three aspects of the study: the mismatch between the experimental procedure employed and the information sought through the study, the small number of subjects and test items in this experiment, and the conclusions drawn from the data derived from the test.

The stated aim of this experiment is to examine the consequences, in terms of memory, of providing advanced learners of English with different types of defining information for previously unknown L2 words encountered in reading texts. Yet these items are not simply encountered while reading; they are clearly identified as target words:

underlined in the texts and with information provided only for these target items. Further, the stated purpose for the reading is to prepare for a comprehension test. In this respect, the study bears more similarities with Seibert's 1930 study reviewed above or with the second of Aizawa's 1999 experiments reported below than with other studies investigating incidental vocabulary acquisition through dictionary use while reading.

Another point is that what is being tested here is a fairly restricted interpretation of memory or retention. The test is conducted immediately after the defining information has been taken from the subjects and while they were still in possession of the texts containing the test items. In other words, it could be argued that retention is not being tested here as very little time had passed from the subjects' seeing the definitions to the test of vocabulary knowledge, and that even then the testees still had access to information regarding the test items.

The nature of the test, too, will have a considerable effect on subjects' scores. The test items and questions are not uniform except as regards their format. Some of the test questions test knowledge of the meaning of the word as used in the text, while others test awareness of pragmatic information such as speaker attitude. Further, various types of words were tested: concrete nouns and abstract nouns, adjectives, verbs, and one adverb. These two non-standardised factors might help explain two counter-intuitive aspects of the results: scores for polysemous items were not significantly lower than for single-sense items, and low-scoring items tended to be low-scoring across the four learning conditions.

Twenty-four subjects took part in this experiment but because for any of the test items the subjects were divided into four groups, only six subjects gave answers for any word in any one learning condition. Conversely, each subject only encountered six of the 24 items in any of the learning conditions. A similar rotation of learning conditions was employed here as in Seibert's 1930 study, with the exception that in Seibert's study all 60 subjects answered each of the 48 test items, producing a total of almost 3,000 individual test answers. This means that there were 750 test answers for each learning condition. In contrast, in Black's study, there were less than a quarter that number of answers: only 576 responses in total, with 144 per learning condition.

The effect of the small numbers of subjects and test items is further compounded by the treatment of the data. Perhaps most problematic is the treatment of the raw scores for the multiple-choice test as data. In this forced choice test, since there were four choices of answers presented for each question, random responses for all items would produce an average of 25% correct answers. However, where 48% of answers for the control condition were correct, the researcher interprets this as follows: "Subjects in the control condition were able to understand half of the words without any defining information". She goes on to suggest that this score is a reflection of subjects' success in making inferences about the words from the text, despite stating earlier that the target words had been chosen for the difficulty with which their meaning could be inferred from context. However, with a score as low as 48% in a four-choice multiple-choice test, we can assume that an average of around 30% of items were actually known, with the other 18% being one quarter of the 70% of items which were guessed at by control group subjects. This figure of 30% known and 70% guessed items for control group subjects compares with an average of around 54% known and 46% guessed for the other three groups' raw scores of 64% to 67%. Although this method of calculation is a simplification of the actual situation, since for some items answers will be half-guessed and for others not all distractors will be equally effective, it does at least give a more accurate representation of the subjects' knowledge than simply using the raw scores.

Neither the test results nor the research undertaken in this study is as impressive as first appears. Black's conclusion is that there was no difference in the effect on memory for the different sorts of dictionary information. A more accurate summary might be to state that the experimental methods employed were insufficiently sensitive to identify differences between the effects of the three types of defining information. The study did, however, show a substantial benefit of all three types of dictionary information over the text only condition. This result reflects the nature of this experiment, is in line with other research in which learners deliberately used dictionaries as a means of learning new words, and confirms the absence of a clear distinction between intentional and incidental vocabulary acquisition in studies such as this.

2.3.4 Krantz (1990)

(Krantz, G. (1990). *Learning vocabulary in a foreign language: A study of reading strategies.* Göteborg: Acta Universitatis Gothoborgensis.)

Krantz's impressive study is an investigation of how Swedish learners of English acquire vocabulary through extensive reading with the aid of monolingual or bilingual dictionaries.

Summary

For this study, Swedish English Department university undergraduates read a geography textbook written in English. While doing this, they were allowed to use a computerized dictionary: either a monolingual English dictionary or a bilingual English-Swedish dictionary. Krantz summarizes his objectives in four research questions. The first two are to what extent L2 vocabulary learning results from encounters with unknown words in context and to what extent it results from a combination of encounters in context and dictionary use. The two other questions ask which of these approaches is more effective in different conditions, and which type of dictionary, monolingual or bilingual, is more efficient for the group of learners investigated.

Fifty-two L1 Swedish learners of English volunteered to take part in this study. They were divided into two groups, a Monolingual Dictionary group and a Bilingual Dictionary group, and a diagnostic test taken earlier in their studies suggested that these were equivalent. As the students studied Economic Geography in addition to English, an English textbook on this subject was chosen as a suitable reading text. The text contained just over 50,000 tokens. The two computerised dictionaries were used to keep an accurate record of the subjects' dictionary use.

A vocabulary test was made of 148 words which occur in the text. The main criterion for inclusion of test items beyond their being in the text was that most of them would probably be unknown to most of the subjects. Perhaps for this reason, most cognates also appear to have been excluded from the list, with only a few (e.g., one equivalent for "context" in Swedish is "kontext") out of the 148 being even possible cognates. 32% of items occur only once in the text, almost 26% occur twice, and 10% three times. The remaining 32% occur four or more times. In the tests, conducted before and after the reading, subjects were asked to supply the meaning of the test items, in Swedish, English or by any other means. For 80 of the test items, only the English word was presented, and for the remaining 68 items, the words were presented in a sentence written to provide as little contextual support as possible.

In reporting the various results of the study, we will begin with the differences between the vocabulary pretest and post-test for the subjects as a whole. We will then look at the scores for the two groups. After this we will consider the direct effect of dictionary lookups on the acquisition of test items.

Overall, the subjects each learned an average of 23 items, with numbers of targeted words learned ranging between 9 and 40. When the lowest scorers for the pretest were compared with the highest scorers, the top 10 subjects learned an average of 24 targeted words while the bottom 10 learned an average of 17 targeted words. For the top 10 subjects, the 24 test items represent 29% of previously unknown items, while for the bottom 10 the 17 words only represent 11% of items unknown in the pretest. It is also worth noting that about 14% of words correctly identified in the pretest were not correctly answered in the post-test. We will consider below why this may be.

Average gains per student for the two experimental groups are almost identical: 23 items for the Monolingual Dictionary group and 22.7 for the Bilingual Dictionary group. However, in the pretest the monolingual dictionary group had correctly identified more of the test items: 39, as opposed to 30.4 for the bilingual dictionary group. This means both that the Monolingual Dictionary group started with a higher vocabulary than the Bilingual Dictionary group and that they learned a greater proportion of previously unknown words. The author suggests that as the Monolingual Dictionary group is stronger, the small differences in test scores can be attributed to this cause, and that in terms of vocabulary gains, the two groups are equivalent. As for the direct effect of dictionary lookup, as we shall see, there are differences between the two groups.

Krantz identifies two types of test item according to the different learning conditions to which the words were subject: words that were encountered only in the text and words that were encountered in the text and looked up. For previously unknown words, 9% were learned through reading only and 10% where dictionaries were also used.

Comments

Krantz's research offers many valuable insights into the role of dictionaries in L2 vocabulary acquisition. As we review the methods employed and results obtained through this study, we will consider the natural circumstances in which reading and dictionary use take place, the method employed to obtain valuable data about learner dictionary use while reading, and the large number of items from the text that are targeted for investigation in the study. We will also reflect on how data from this study may help shed light on the nature of the mental lexicon of the learner.

As the researcher points out, we should not be surprised if a set reading text in a foreign language course is read with two purposes and that these will affect reading behaviour:

to understand the content of the text and for L2 language improvement. It is only with a recognition of these combined purposes that the reading condition can be judged in terms of natural behaviour in the specific context. In Krantz's study, there was no announcement of a comprehension or vocabulary test that would follow the reading, so in this respect there was no imposed linguistic purpose for reading the text. On the other hand, the reading of the text in a monitored location with the use of computerised dictionaries does create an environment in which the importance of careful reading is evident, if not explicitly stated. The dictionary use under investigation in this study can be understood, then, as reflecting typical dictionary use of a particular type: careful reading in an academic context by highly motivated learners of English using electronic dictionaries. It is also worth noting that as the text employed is fairly specialised, specialist knowledge of the subject matter of the text would also affect knowledge of, and ability to infer, the meaning of words in the text.

As has been suggested in other studies reviewed here, readers of long L2 texts typically use dictionaries very little. In this study, however, there was a considerable amount of dictionary use; this averaged 263 lookups per student during the reading, ranging between 58 and 641 lookups, with total reading times for students taking between seven and eighteen hours. Readers using the bilingual dictionary consulted their dictionary almost twice as frequently as those using the monolingual dictionary: an average of 353 lookups for the Bilingual Dictionary group as compared to an average of 180 for the Monolingual Dictionary group.

As many as four factors may help account for the generally high levels of dictionary use by subjects in both reading conditions. One, as noted by Knight (1994), is the relative speed and ease of consulting electronic dictionaries. A second consideration is the appeal of the new technology that the electronic dictionaries may have represented for the subjects. A third factor may be the monitored reading environment, which promotes careful reading, one aspect of which is increased attention to unknown words. Finally, in this environment, advanced learners might use dictionaries more than less able learners. Although in terms of unknown words they would have less need of dictionaries than lower level proficiency learners, they may be more motivated to know the meanings of the relatively few unknown words. In addition, the dictionary use of more advanced learners should be more efficient; each lookup would be faster and interrupt the flow of the reading less than would be the case for less able learners. This final point may not seem to be borne out by the lookup data for the subjects of this study, given that the bottom ten subjects used the dictionary more than twice as often as the top ten subjects. However, we need to remember that all the subjects in Krantz's study, Swedish learners who chose to major in English at university and who volunteered to assist with this study, may be described as motivated learners of English with a high level of proficiency in the language.

A further impressive aspect of this research is the exceptional record of dictionary use it provides. Although the use of electronic dictionaries will affect dictionary use in some respects, it does provide an ideal means of keeping an accurate yet non-intrusive record of dictionary use. This makes possible a very clear and detailed indication of how actual dictionary use affects vocabulary acquisition for the words that were known to have been looked up, and even allows for calculations of the relationship between time spent on lookups and learning. The advantages of this electronic recording of dictionary use are clear, especially with studies of larger numbers of readers, when we note that in many studies not using this technology, the only comparison available is of overall test results for readers with different reading conditions who may or may not have used the dictionaries at their disposal for some of the test items.

Also of particular note is the large number of test items included in this study. This is especially valuable with investigations into vocabulary acquisition resulting from extensive reading through which tens of thousands of words are encountered. Even the 148 test words drawn from the text, a very large number when compared with the 20 or 30 items in other studies, by no means represent a comprehensive survey of unknown words in the text. With an average of 35 of the test words correctly identified in the pretest, this leaves 113 test items unknown to the average subject. When we compare this to the estimated average of 507 unknown word types in the text, the previously unknown test words represent under a quarter of the estimated average unknown words in the text. Even 148 test words, then, are not sufficient to provide a full picture of dictionary use and vocabulary acquisition through reading the text. On the other hand, they do produce sufficient data to make extrapolations for all unknown words in the text. Overall, with this number of items a much more accurate and reliable picture of vocabulary acquisition through dictionary use becomes available than is possible for those studies with much smaller numbers of test items.

While there are a number of valuable aspects to this study, one piece of the data reveals problems with the assumptions upon which this, and much other research, is based. Of the average 35 items correctly identified in the pretest, 14% (5 items) could not be correctly identified in the post-test. Compared to the average gains of around 23 items between the pretest and the post-test, this figure is not large but it does reveal a challenge to the view that vocabulary development proceeds uniquely forward, whether in numbers of items known or in extent of knowledge of individual items. There are various possible explanations for the circumstances in which 14% of correctly identified items in the pretest were not recognized in the post-test. It could be argued that in the period of a few days between the two tests, the subjects forgot some of the words they had known. This theory relies on the subjects' failing to encounter, or notice, these words in the text. Another possibility is that the contexts or definitions of these items were misleading and challenged the subjects' prior, correct, understanding of the meanings of the words. While these factors may play a part in the apparent loss of vocabulary knowledge, two further causes appear much more likely: the instability of much of a language learner's word knowledge and the instability of language learners' confidence about word knowledge. That 14% of "known" items should become "unknown" a few days later could be seen as reflecting the unstable nature of our mental lexicon, perhaps amplified by the use of a test in which there is no way of indicating partial knowledge. The nature of the test may also be involved in the second possible cause; confidence about knowledge of lexical items does waver but as the test allows for no indication of confidence about word knowledge, this can only be reflected as apparent "loss" of words for which subjects do not feel sufficiently confident to supply the meaning.

As we shall see, few other studies have succeeded in producing comparable data as regards the relationship between dictionary use and vocabulary acquisition, and this is especially admirable in the context of extensive reading.

2.3.5 Bogaards (1992)

(Bogaards, P. (1992). French dictionary users and word frequency. In: H. Tommola, K.Varantola, T. Salami-Tononen and J.Schopp (eds.). *EURALEX '92 Proceedings* (51-59), Tampere: Department of Translation Studies, University of Tampere.)

This is a relatively small study which addresses the usefulness of different types of dictionaries in translation, and incidental L2 vocabulary acquisition that may result through using these dictionaries. This focus on translation as the context of dictionary use is almost unique among the studies reviewed, but it is especially valid when we reflect on how much dictionary use, both in and outside of formal learning contexts, takes place in the context of translation work.

Summary

In this study, Dutch university students of French were asked to translate a Dutch text into French with the aid of various types of dictionary. Altogether, 42 students completed the required tasks of translating the text followed, two weeks later, by a test of 17 less common words which appear in the text. The students were divided into four groups according to the kind of dictionary provided during the translation task: a bilingual (Dutch-French) dictionary group, a target language (French) monolingual learner dictionary group, a standard French dictionary group, and a no dictionary group. Students using dictionaries were asked to keep a record of which words they looked up.

A further 14 students served as a kind of control group, taking the vocabulary test without having translated the text.

It is not indicated how the experimental groups were formed. Summaries of student grades for previous work indicate that the groups are similar but that they are not the same; this suggests that these groups may have been already pre-existing groups of students, such as classes or seminar groups. As for the 17 test items, they are French translation equivalents of Dutch words in the text, described as not normally being in the productive vocabulary of the type of learner taking part in the study. The text to be translated was a 150-word Dutch passage, judged to be relatively easy with the exception of the 17 targeted words. The vocabulary test, conducted two weeks after the text translation task, consisted of the list of 17 isolated Dutch words, which the subjects were asked to translate into French.

There are two sets of results: for the translation of the targeted words in the text and for the translations of the same words as isolated test items two weeks later. These results are shown in Table 2.3 below. In addition, within these results, scores are provided for words which had been looked up and words which had not. For the bilingual dictionary group, an average of 12.0 of the 17 targeted words were looked up, with 10.3 of these correctly translated and a further 3.2 of the 5 words which were not looked up being correctly translated. This compares with an average of 7.6 targeted words looked up by the monolingual learner dictionary group, only 3.6 of which were translated correctly, with 4.0 of the 9.3 words not looked up being correctly translated. For the standard French dictionary group, too, success rates were low: an average of under 5.9 words were looked up, with 2.4 of these correct. Of the 11.1 words not looked up, 5.6 were correctly translated. As for the no dictionary group, 5.6 of the 17 words were correctly translated. This compares with totals of correctly translated words of 13.5 for the bilingual dictionary group, 7.2 for the monolingual learner dictionary group, and 7.5 for the standard French dictionary group.

Table 2.3

Results of Bogaard's (1992) test (total of 17 items)

	Looked up	Total translated (= looked up items)	Total translated 2 weeks later
Bilingual dictionary	12	13.5 (10.3)	8.2
MLD	7.6	7.6 (3.6)	8.8
Standard French dictionary	5.9	8.0 (2.4)	7.6

In terms of targeted words correctly translated in the test two weeks later, the figures for the three dictionary groups are much closer. They are 8.2 for the bilingual dictionary group, 8.8 for the monolingual learner dictionary group, and 7.6 correct words for the standard French dictionary group. We will now go on to consider what these figures may tell us about vocabulary acquisition attributable to the use of the three types of dictionary.

Comment

This study, at first glance, appears very straightforward both in terms of the procedures employed and with regard to the results gained through the study. There may be problems with the small numbers of targeted words, with the small numbers of subjects in each group and with the composition of the groups, but the purposes and the procedures employed investigate dictionary use and vocabulary acquisition in a careful and coherent way. Bogaards sums up the findings by pointing out that the bilingual dictionary was most useful as an aid to translation and that the usefulness of monolingual dictionaries depends on the type of word being investigated, but that there was no clear advantage for either dictionary type as regards vocabulary retention.

In some respects, the answer to the question about which type of dictionary is most useful for translation into the L2 is self-evident. After all, a bilingual dictionary provides translation equivalents for looked up words while a monolingual dictionary does not. The use of a monolingual target language dictionary in translating from the native language will inevitably be indirect. All a translator with only a target language monolingual dictionary can do is guess at the meaning and look up supposed synonyms for the word. Monolingual dictionaries may, however, be useful in a confirmatory role in this context; translators can make sure that they are using the right word.

As a simple comparison of results from the translation and from the test indicate, benefits of different types of dictionaries in terms of vocabulary acquisition are less clear-cut, and harder to interpret. For the bilingual dictionary group, just under two-thirds the number of words correctly translated in the text were correctly translated in the test two weeks later. For the learner dictionary group, a slightly greater number of words were translated correctly in the test than in the text, while for the standard French dictionary group, the figures for the text translation and for the test were almost the same.

These results suggest that while the bilingual dictionary was useful as a tool for translation, much of the information looked up was not retained two weeks later. On the other hand, the lack of difference between text and test results for the two monolingual dictionaries suggests that these dictionaries were of little use in the translation task. It also suggests that little retention took place: that the students already knew most of the targeted words for which they could give translation equivalents in the text and the test.

The use of translation as a means of investigating vocabulary acquisition is valuable since it is in this context that much dictionary use takes place. It does, however, bring various problems with it. If we reflect that words other than the specified targeted French words are acceptable answers in the translated text and in the test itself, we may ask what is being tested and what the results may tell us. While a bilingual dictionary will provide translation equivalents for unknown L2 words, the absence of a bilingual dictionary may force learners to use known L2 vocabulary which approximates in meaning to the L1 word. A monolingual dictionary will serve the purpose of confirming whether these known words may be acceptable substitutes for the unknown translation equivalent.

As far as the vocabulary test is concerned, for test items with previously unknown translation equivalents, bilingual dictionary group subjects will either be recalling the looked up translation equivalent or, for the first time, trying to think of a satisfactory equivalent. For the two monolingual dictionary groups, what the test requires of them is much closer to that for the translation task, this time without the aid of context but with the aid of the memory of the experience two weeks earlier. In addition, we should not underestimate the motivation of language learners to consult their bilingual dictionaries after the translation task to check whether their guesses or impressions were correct. This behaviour may help to account for the monolingual dictionary groups' apparent ability to 'retain' words that they had been unable to translate correctly during the translation task.

As with many other studies, the small number of test items leaves us with as many questions as answers. The author suggests that success in finding or retaining unknown words depends on the type of word, but with so few targeted words of any particular type, this study offers little support for this theory. The focus on translations is attractive and useful, and does highlight the insufficiency of monolingual dictionaries for the purpose of translation, but it also brings problems of its own. On the other hand, the surprisingly low retention rate for bilingual dictionary users serves as a reminder that the comprehension of L2 words, as demonstrated through translation, cannot automatically be equated with their retention.

2.3.6 Luppescu and Day (1993)

(Luppescu, S. and Day, R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning 43* (2), 263-287.)

This study was one of the first to ask directly, in the context of L2 reading, what effect dictionary use has on vocabulary acquisition, as opposed to text comprehension. It is interesting principally for demonstrating the issues involved in considering the effect of dictionary use without actually documenting the dictionary use that took place.

Summary

In this experiment, 303 Japanese university students were asked to read a short story of 1853 words written in English. They were divided into two groups by using whole classes of students for each group; 148 of the students were not allowed access to a dictionary while reading and 145 were allowed to use to use their English-Japanese dictionaries while reading. The text was chosen for having been largely comprehensible and of interest to a similar group of learners of English. There was no set time limit for the reading. Immediately following the reading, all the subjects were given a vocabulary test in which knowledge of 17 words from the short story was tested. The test was a multiple choice type, with the choices for each item being the test answer, three distractors, and an *I don't know* option. Two points were awarded for a correct answer, one point for *I don't know* and no points for a wrong answer.

Scores for the group permitted to use dictionaries were significantly higher than for the group that was not. According to the scores, the chances of subjects allowed to use dictionaries getting a right answer were 1.86 times greater than for the no-dictionary group, although there was a wide variation among test items. It is worth noting that as the Dictionary group chose the *I don't know* response almost twice as often as the No Dictionary group, their raw scores are not directly comparable. Another important statistic is that the group permitted the use of dictionaries took, on average, almost twice as long to complete the reading: over 21 minutes as compared to under 12 minutes.

Comment

This experiment, one of the first to investigate the effect of dictionary use on learning in a fairly natural context, and one that remains much cited in the literature, raises various questions relating to the investigation of L2 vocabulary acquisition through reading and dictionary use. Three aspects of this paper are of particular interest: the challenge of investigating the effects of dictionary use in such an environment, the comparability of the two groups of subjects, and the quantity and quality of the test items used to measure vocabulary development through reading and dictionary use.

One point the authors make is that their aim is to investigate natural dictionary use. Because of this, although they divided the subjects into two groups, a group with dictionaries and a group without dictionaries, they could not dictate that subjects in the with-dictionary group had to use their dictionaries, let alone tell the subjects which words they should look up. Neither did they feel that any observation or monitoring of dictionary use would be sufficiently non-invasive as to leave the subjects' dictionary use behaviour unaffected and natural. This means that all they are able to state with authority is that one group had possession of dictionaries and one group did not, and that the group with dictionaries spent, on average, almost twice as long reading the short story as the group without. If we can assume that the groups were comparable in terms of L2 proficiency, it is reasonable to infer that the extra time was spent on dictionary use. Other than this, the main, again indirect, evidence of dictionary use is that the group with dictionaries had better test scores than the group without; since the two groups were believed to be the same except for the possession or not of dictionaries, it is assumed that this must account for the difference in test scores. As for whether the test items had been looked up in the subjects' dictionaries, the only indication of this, again circumstantial, is of a negative kind. The advantage for the group with dictionaries was smallest, or did not exist, for words with the most demanding dictionary entries: those with a large number of different senses listed in the entry and for which it would be harder to locate the sense used in the story.

The inferences made above about dictionary use accounting for the differences between the groups' results can only be justified if we are sure that the two groups are the same in all respects other than whether or not they had access to a dictionary during the reading. For this experiment there was no pre-test or other data to show that the two groups were equally proficient in English either generally or in terms of vocabulary. We are told that the groups were made up of whole classes of students and that students had been assigned to these classes solely on the basis of their surnames. The authors claim that the assigning of whole classes of students to the two groups can be seen as equivalent to randomly assigning individual students to the two groups. For a test of grammatical knowledge, this reasoning may have some validity, but in terms of vocabulary it is quite conceivable that different classes of students may have studied different material with different vocabulary, some of which may be included in the 17 test items. Further, attitudes towards study, such as how much time and effort students are willing to devote to a task, may often be largely shared among students in a class, especially for tasks conducted in the class such as the reading task in this experiment. A final question is whether the experiment was conducted with all the classes in the same circumstances. For example, if for some classes the experiment were followed by a lunch break and the students were free to leave the classroom on completion of the test, this may have had a greater effect on the time subjects spent reading the text than whether or not they had access to a dictionary.

Related to the above questions about the comparability of the two groups is the small number of test items. Initially, there were 17 words in the short story that were chosen as test items, although data from two of the items were later excluded from analysis. This left 15 items, some of which are relatively high frequency items (*happen*, *appear*, *worse*) and some which are widely known in Japanese as loanwords from English (*slide projector*, *clear*). The problem with these two types of words is that there is a reasonable likelihood of their having been encountered in some classes but not in others. This alone may cause a significant difference between the two groups' results to be recorded.

Finally, there are also some problems with the answers and distractors for the test items. For the target word *chant*, two of the choices offered might be acceptable: *speak* and *sing*. With *clear*, for which an antonym was requested, three of the four choices may be acceptable: *dirty*, *vague*, and *dull*. As for *happen*, the correct answer is *occur*, a markedly less frequent word than the target word. In any test there may be one or two items with which there are problems. Here, as many as seven items are unsatisfactory in some way, accounting for almost 50% of those used for analysis. With such a small total number of test items, there is a greater likelihood that problems with individual items would have a considerable effect on overall results.

The approach employed by Luppescu and Day in this study does initially appear to offer valuable insights into vocabulary development through dictionary use in a relatively natural environment. It does do this to some extent, although only with regard to issues that are of interest but are peripheral to this study: the longer time taken by the dictionary group and the lower scores for words with more senses in the dictionary. However, the problem regarding the comparability of experimental groups, the lack of direct evidence of dictionary use, the non-comparable nature of the test results, the small number of test items, and the fact that a large proportion of these are faulty in some way, mean that there is little reliable data to support the central claims made in this study. Although Luppescu and Day's paper remains a much-cited study in the field, these weaknesses do require us to re-evaluate its contribution to our knowledge of L2 vocabulary acquisition through dictionary use.

2.3.7 Fischer (1994)

(Fischer, U. (1994). Learning words from context and dictionaries: an experimental comparison, *Applied Psycholinguistics, 15* (4), 551-574.)

The research described in this paper investigates how the use of different sources of information about previously unknown L2 words enables advanced German learners of English to understand, use, and remember the words.

Summary

The purpose of this study is to investigate the effect on the comprehension, use, and retention of 12 unknown L2 words of three different learning conditions: learning the target words through consulting a set of monolingual dictionary entries; learning them by focusing on an extract of a novel in which the target words are imbedded; and learning them by using both the dictionary entries and the text together. A total of 87 German high school students taking part in the study were divided into four groups: one for each source of information and a control group which was used to test the comprehensibility of the text by itself. There is no explanation as to how the groups were formed. The 12 target words were selected as words that should be unknown to the subjects and that are difficult to understand. They were all low frequency items divided equally into nouns, verbs and adjectives. These were inserted into the text by replacing easier words for the target words or by adding a phrase containing a target word. Apart from the target words, only relatively high frequency words were contained in the text.

The experiment lasted two hours and was composed of four parts. It began with a pretest in which the subjects were asked to give the meaning of the target words. This was followed by a learning condition phase during which they were given their specific set of learning materials and asked to write a sentence in English for each of the target words. Once they had all done this, they were asked to translate their sentences into their mother tongue. The learning materials were then taken away and the experiment ended with a post-test which was the same as the pretest. The control group subjects were given a gapped version of the text, with the target words removed, and asked to write a summary of the story.

Two raters evaluated the English sentences according to whether they were idiomatically meaningful, while two other raters evaluated the German translated sentences according to whether they included an accurate, partially accurate or inaccurate translation of the target word. Discernible strategies for the subjects' English sentences were also rated with reference to the learning materials to which they had access.

The pretest confirmed that very few of the target words were known. The best known word, *insidious*, was known to 13, or about 20%, of the 67 subjects, while no other word was known to more than one subject in any group.

The English sentences, and their demonstration of the use of the target words, were evaluated in terms of omissions, idiomatic correctness, and questionable or correct usage. Raters found no overall difference between the groups in this respect. The mean number of adequate German translations of the 12 target words for the dictionary, text, and mixed groups were 5.13, 4.74 and 3.55 respectively, between just under 30% and a little over 40% of the total. The dictionary group's average scores were significantly better than those of the two other groups. Results are also provided for assumed strategies employed in writing the sentences. As this is not the focus of our interest, we will not consider this aspect of the study further. As for post-test results, no information is provided in Fischer's paper about this aspect of the paper.

Comment

First, we should note that although the title of the paper is "Learning words…", no report of learning or retention of the target items is made in this paper. It may appear that understanding or comprehension is equated with learning but clearly, whether for contextual information or information contained in dictionary entries, unless some of this information is retained it would be hard to argue that the words or meanings had been learned. In fact, the description of the study does include a post-test measuring retention but these data are neither presented nor discussed in Fischer's paper. We will consider below why this may be.

Before we consider the results, we should consider the approach employed by Fischer and the target words used in this study. Here, the focus is clearly not on incidental vocabulary learning but on the deliberate use of written contexts and of monolingual dictionary entries as means of understanding the meaning of the target words. It also differs from many other studies (although is similar to one reported by Nesi, 2000: 71-92) in that it focuses on the productive use of the target words, which includes an evaluation of syntactic and idiomatic success in using these words in a sentence. This focus on productive use of vocabulary does not extend to the post-test, but whether or not subjects are able to use words appropriately while referring to learning materials is surely an indication of whether they might be able to do this when these materials are taken away.

There are only 12 target words. This is understandable when we consider the demands required of the subjects in writing an English sentence for each word, then translating the sentence into their mother tongue. But when we consider that the 12 words are subdivided into 4 adjectives, 4 verbs and 4 nouns, we may wonder whether the results from this study may tell us anything of value about these parts of speech. Further, we cannot expect the sets of dictionary entries or the contexts in which the words were placed to be representative of those for unknown words that learners might encounter. This is not, apparently, the concern of the researcher, since we are told that the words were selected on the basis of their being difficult to comprehend. The learning of such words may differ from that of other words and, as a consequence, this will restrict what claims may be made for the results of the study. The small number of items is also likely to make the results of the study less reliable and the differences observed between groups less easily attributable to the different learning conditions under investigation.

When we turn to the results, the above fears about their reliability appear to be confirmed. Especially for the subjects' German translations, through which they demonstrate whether they have understood the target word, scores for all three groups are low, suggesting that neither from context nor from a dictionary definition were most subjects able to work out the meaning of the majority of the target words. Further, subjects with access to both sets of materials do worse than both groups with access to only one type of material to assist comprehension. This is counter-intuitive, and suggests that subjects with two sources of information did not have time to make the most of them or to complete the writing and translation tasks. The low scores for comprehension, together with the small difference between groups, may also help to explain why the author presents no data for retention of word knowledge; we would expect retention rates to be quite a lot lower than these already low comprehension rates, and for there to be no significant difference between the groups' results.

Almost exclusively, other studies in the field have focused on word meaning: the comprehension and retention of the meaning of unknown L2 words. As such, they have ignored issues more relevant to production, such as syntagmatic, collocational, and pragmatic acceptability. This study does, ambitiously, address those issues but, in order to do so, must limit itself to a very limited number of target words. This small number of target words, together with problems of target word typicality and of interrater reliability, in turn affects the value and reliability of the findings, leaving the questions posed through this study largely unanswered. Further, the apparently central question of

L2 vocabulary acquisition, somewhat mysteriously, remains unaddressed. Despite these weaknesses, however, this study does highlight the very limited focus of the majority of studies in the field and the consequent neglect of issues that are central to the investigation of the effect of dictionary use on vocabulary acquisition.

2.3.8 Knight (1994)

(Knight, S. (1994). Dictionary use while reading: the effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal, 78* (3), 285-299.)

This thoughtfully designed study investigates L2 vocabulary acquisition through contextual guessing of unknown words encountered through reading and through consulting a bilingual dictionary while reading.

Summary

A total of 105 intermediate level American university students of Spanish took part in this study of text comprehension and vocabulary acquisition through reading Spanish texts and bilingual dictionary use. Based on their American College Test verbal scores, the subjects were divided into two equivalent groups: one in the Dictionary condition and the other in the No-Dictionary condition. Two sets of two Spanish texts of up to 250 words were prepared, and from each set 24 words were selected, making 48 words in total. Each of the groups was further divided so that half the subjects in each group read one set of texts and the other half read the other set.

Two weeks prior to the experiment itself, the subjects were given three tests: one a vocabulary checklist for eighty words for which subjects were asked to mark which words they knew, and then two tests for the 24 words drawn from the text set they would not read: a test in which they had to supply English equivalents, followed by a multiple-choice test. For the experiment itself, conducted using computers, each subject was given a computer disc containing one set of texts and two vocabulary tests relating to vocabulary in these texts. The discs for the subjects in the Dictionary condition also enabled access to a computerised Spanish-English dictionary. Immediately following each text reading, there was a recall protocol in which the subjects had to write, in English, as much as they could remember from the text. Following the recall protocol for the second of the readings, the subjects took the two vocabulary tests. In the first of these, subjects were required to give the meaning of 24 words from their text set. This was followed by a multiple-choice type test, with for each item the correct answer, three syntactically similar distractors, and a *don't know* option. Two weeks after the reading, the subjects were given the same two tests again.

For the supply-meaning tests, and for the multiple-choice tests, there are three sets of results from the tests immediately after the reading, from the tests two weeks after the reading and, as a kind of control, from the tests for the 24 words in the text-set not seen by the subjects. These are shown below in Table 2.4. The average scores for subjects in the Dictionary and No Dictionary conditions for the immediate post-tests are as follows (all scores are for a possible total of 24): for the supply-meaning tests, mean scores are 4.95 with dictionaries and 1.72 without dictionaries. This compares with a mean score of only 0.15 for the test of the items from the other text-set: i.e., for words not seen either in reading or when using dictionaries. As for the 24 items in the multiple-choice tests, subjects in the Dictionary condition scored an average of 14.56 as opposed to 8.75

for the No Dictionary condition. For this test, the mean for words from the unseen text set was 1.80. For the set of tests two weeks later, not all scores were as might have been predicted. For the supply-meaning tests, the Dictionary group's mean scores had fallen from an average of 4.95 to 3.37 while the No Dictionary group's scores had risen from 1.72 to 2.30. The multiple-choice test scores for both groups fell slightly over the two weeks: from 14.56 to 12.24 for the Dictionary group and from 8.75 to 8.06 for the No Dictionary group.

Table 2.4

Results of Knight's (1994) test (total of 24 items)

	Immediate post-tests		Tests 2 weeks later	
	Meaning	M/C	Meaning	M/C
Dictionary group	4.95	14.56	3.37	12.24
No Dictionary group	1.72	8.75	2.30	8.06

Comment

This study impressively addresses many of the challenges involved in identifying and measuring dictionary use during reading and its effect on vocabulary acquisition. These include identifying previously known vocabulary and its effect on vocabulary acquisition data, and using sufficient numbers of test items to produce meaningful results. Test results show substantial vocabulary gains for both reading with a dictionary and reading without a dictionary, with that for dictionary users significantly higher than for those without access to dictionaries. Two related aspects of the results of this study appear worthy of further investigation and comment: the high levels of vocabulary growth for both learning conditions in the tests following the reading and the high scores in the retention tests two weeks later.

Compared to other studies aiming to investigate incidental L2 vocabulary acquisition from reading in formal educational or research settings, there is a high level of vocabulary growth recorded for the 24 items tested with each set text, both for subjects reading the texts alone and for those reading with dictionary use. This is true for the supply-meaning test results and even more for the multiple-choice test results. No explanation is suggested for these high scores, and comparisons with the only study reporting similar levels of word learning (White, 1988) are rejected. That is because in White's study the subjects were given the meaning of targeted words and had to write definitions for them: resultant vocabulary acquisition could clearly not be described as incidental.

One possible reason for the high retention rates recorded in Knight's study may also be found in the task the subjects were set: studying the text for as long as necessary until they were ready to write all they could recall from the text. Although the researcher justifies the use of a recall protocol from the perspective of evaluating text-reader interaction, from the perspective of incidental vocabulary acquisition through reading its use is more questionable. Despite this, the rate of learning is still surprisingly high, especially for the No Dictionary condition since, typically, for only a small proportion of unknown words in a text is the context usually sufficiently explicit to allow confident and accurate inferring of the meaning (see, for example, Schatz and Baldwin, 1986).

Regarding explanations for the high levels of learning, we may exclude the presence of easily recognisable English-Spanish cognates among the targeted words, since there were few if any such items. More significant is the extent to which targeted words from each text belong to fairly restricted semantic fields. For the first text of Set A, for example, eight of the twelve targeted words are related to either the sea or to giving birth: *la ballena* (whale), *la bahia* (bay), *yubarta* (humpback or finback), *el parto* (giving birth), *dar a luz* (give birth), *la cria* (an animal's young), *la partera* (midwife), and *la hembra* (the female). In the multiple-choice tests, none of the distractors used for these items are from either of these two semantic fields. A consequence of this is that their power to distract is very limited, while the likelihood of testees' guessing the correct answer is increased.

The above issue is related to the unexpected rise in the average number of correct items for the No Dictionary group between the supply-meaning test directly following the reading and the test two weeks later. Unless an additional source of information about the items were encountered between the two tests, we might expect this knowledge of targeted words to fall over the two weeks between tests. In fact, between the two supply-meaning tests there is one source of knowledge made available to the subjects: that contained within the multiple-choice test immediately following the first supply-meaning test. The difference between the scores for the supply-meaning tests and the multiple-choice tests here illustrates the difference in degree of confidence required to answer the two tests. In terms of task, a multiple-choice question is requesting recognition of the targeted word. As a source of knowledge, its effect is confirmatory. In other words, its effect is similar to looking up a polysemous word in a bilingual dictionary and identifying the correct sense. Since the multiple-choice test followed directly after the supply-meaning test, and especially because of the shared semantic field of many test items, it may have served precisely this purpose of confirming word meaning for the subjects. This may help explain why this rise in delayed post-test scores is not also evident in the Dictionary Group's results, since for these subjects the dictionary entries consulted would already have provided more information than may have been gleaned from the multiple-choice test.

This study does show high rates of vocabulary acquisition, especially for recognition of targeted words. It confirms results from other studies focusing on the inference of word meaning from context, which demonstrate that in many cases contextual information will either be misleading or will be insufficient to enable correct guessing from meaning.

Despite the positive results of this study, we can see how the results of a study may be adversely affected by small numbers of targeted words, shared semantic fields for many of the items, and poor distracting qualities of distractors in multiple choice questions.

2.3.9 Hulstijn, Hollander and Greidanus (1996)

(Hulstijn, J.H., Hollander, M. and Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign-language students: The influence of marginal glosses, dictionary use, and reoccurrence of unfamiliar words, *The Modern Language Journal, 80* (3), 327-339.)

This carefully constructed study investigates the influence of two different variables on incidental L2 vocabulary acquisition: different conditions in which texts are read, with or without the aid of dictionaries or marginal glosses, and varying times of occurrence of targeted words in the texts used.

Summary

A total of 78 advanced learners of French at three Dutch universities took part in this research. They were asked to read a short story in French, 1306 words in length, following which they were told they would be given a comprehension test. Sixteen words were selected as targeted items from this text, which was edited so that eight of the words occur once in the text and eight occur three times. The subjects were randomly divided into three groups and each group was assigned a different text reading condition: reading the text with L1 marginal glosses, reading the text while being able to consult a bilingual French-Dutch dictionary, or reading the text without access to either glosses or a dictionary.

Twenty-five minutes were allowed for reading the text, following which three tests or surveys were administered. Firstly, the subjects were asked to indicate which of 32 words in the test they recalled having seen in the text, and to write the meaning of these words. The group with dictionaries were also asked to indicate which words they had looked up. Secondly, the subjects were asked which of the 16 targeted words they had been familiar with prior to reading the text. Thirdly, half of the targeted words were presented in a few lines of the context in which they had appeared in the text and subjects were asked to provide the meanings of the words. The other eight targeted words were presented in isolation and subjects were asked to give their meanings.

There are four main types of data presented from this study: the subjects' self-reported preknowledge of the test items, the Dictionary group subjects' reporting of their dictionary use, the test results of the subjects in the three reading conditions, and the test results for the words occurring once or three times in the text. For the latter two sets of results, there are results for the tests of the words presented in isolation and presented in context.

Of the items reported as previously known, only those for which a correct answer was given were counted: out of 16 targeted words, for the Marginal Glosses group an average of 0.7 words per subject were already known, for the Dictionary group 0.1 words (i.e., an average of only one word per 10 subjects) and for the Control group 0.4 words.

Reported dictionary use by the Dictionary group subjects was very low: 4 subjects did not use their dictionary at all, while the remaining 20 subjects looked up targeted words a total of 38 times; this represents an average of under two targeted words looked up by each of these subjects.

Post-test retention scores were also low, especially in the test with words presented in isolation, where scores ranged from 0.2 to 1.4 out of a possible score of 8, equivalent to between 2.5% and 17.5% of the total. For the words presented in context, scores ranged between 1.1 and 3.4 (14.25% and 42.5%). The Marginal Glosses group's scores were significantly higher than those for the other two groups, for test words both in isolation and in context, for words occurring both once or three times in the text, and for both whole points only or whole and partial points. There was no significant difference in the results of the Dictionary group and the Control group, although, as we will discuss below, analysis of scores for individual items do reveal interesting data about the effect of dictionary use.

Comment

This study illustrates how even a carefully planned and executed study can be marred by the unpredictability of L2 readers' dictionary use. The research does provide valuable data about the effect of frequency of occurrence of words on their retention and of the effect of providing L1 marginal glosses, but little data relating directly to the effect of dictionary use. This is because very little dictionary use took place: overall, on average, fewer than one in ten targeted words were looked up. This tells us little more than that little or no dictionary use will have little or no effect on vocabulary retention. Although judicious dictionary use providing knowledge of just one or two key words in a text may affect comprehension of the whole text, in this case the very selective dictionary use does not appear to have aided vocabulary learning in any global way. We will review what effect of dictionary use there was before going on to consider why dictionary use was so limited.

Only when there is analysis of scores for items that were looked up by individual subjects do we see what effect there may be for dictionary use. For the 38 lookups of individual targeted words by individual subjects, there was a retention score of 59% for words presented in isolation, as compared with an average of 8.75% for the Dictionary group responses overall and 26.25% for the Marginal Glosses group. For individual items the scores are still higher; the most looked up word, *pépinière* ("tree nursery"), was looked up by 17 subjects, 15 of whom were later able to give a correct or partially correct response. This represents an 88% rate of retention for these subjects, as opposed to 70% for the Marginal Glosses group. These figures should, however, be treated with caution. While we may accept that the three original groups of subjects are equivalent, a subgroup distinguished by its dictionary use will not be. This may be especially pertinent for this study since the high levels of standard deviation for the means of test scores alert us to the wide variation among subjects within each group.

A further point is that, as the researchers note, for four subjects in the Marginal Glosses group, retention levels for words in context was in the order of 13 wholly or partly correct items out of 16. We are, perhaps tellingly, not told how many lookups were made by individuals in the Dictionary group, but it may well be that the most able learners were also the most avid dictionary users, and that it is this factor that is reflected in the analysis of scores for items that were looked up. Two further factors complicate this issue. If we assume that all the glosses were actually referred to, the vocabulary learning burden for the Marginal Glosses group can be said to be all 16 words, while the learning burden for the Dictionary group members would only be the couple of targeted words that they looked up. We should not be surprised that the small numbers of looked up words are learned better than the larger number of items with glosses – provided, that is, that the glosses were actually consulted. It is assumed that glosses were consulted, and the test scores seem to confirm that they were, but there is no external evidence of the extent to which the marginal glosses were consulted.

The second question we need to ask is why dictionary use was so limited. The main reason proposed by the researchers is that the students in the Dictionary group did not feel that looking up words would help them with the comprehension questions they were expecting. Other contributory factors mentioned include the advanced level of the learners, their natural aversion to using dictionaries, the length of the text, and the trouble of using paper dictionaries as compared to electronic dictionaries. Two factors not mentioned but perhaps also relevant are the reading environment and the limited time available for dictionary use. Two-thirds of the students in the same room as the Dictionary group were not using dictionaries at all; this may have created an atmosphere in which students with dictionaries felt uncomfortable about using them even if they did have time to use them.

In conclusion, while this study does provide us with various insights into dictionary use, and non-use, it provides little reliable data about the effect of dictionary use on incidental L2 vocabulary acquisition. On the other hand, the much more impressive results for the Margin Glosses group may tell us as much about dictionary use as about the use of glosses themselves. The test results for this group suggest that where meanings of unknown words in an L2 text are provided in an easily accessible form, they will be consulted and substantial levels of vocabulary learning may result. Presumably, how accessible the meaning is – whether as marginal glosses, in an electronic or paper dictionary, in a bilingual or monolingual dictionary – will affect the extent to which the meanings of unknown words are consulted. So, too, will the purpose of reading a text. If, for example, subjects had been told that no test would follow the reading, or that a vocabulary test would follow, we might expect different dictionary use behaviour to result. Although Hulstijn et al.'s purpose through this study was to investigate incidental vocabulary use in a natural L2 reading environment, we can see that under test conditions there is no such thing as a natural reading environment. Reading followed by a vocabulary test is arguably as *natural* as reading followed by comprehension questions and it is not unreasonable to suggest that reading in class that is not followed by a test of any kind might, in some students' or teachers' eyes, seem quite unnatural.

2.3.10 Aizawa (1999)

(Aizawa, K. (1999). *A study of incidental vocabulary learning through reading by Japanese EFL learners.* Unpublished PhD thesis, Tokyo Gakugei University, Tokyo.)

Two studies described in Aizawa's thesis investigate the effect of bilingual dictionary use as an aid to reading an L2 text on the learners' comprehension and vocabulary development. The two papers differ mainly in the degree to which subjects are given guidance in dictionary use, and the effect that this guidance has on dictionary use behaviour.

Study 1

Summary

In the first of these studies, Aizawa investigated vocabulary development through four different reading conditions: with vocabulary priming exercises before reading, with marginal L1 glosses, with bilingual dictionary use permitted, and with the reading text alone. A total of 230 university students took part in this study and were divided as whole classes into four groups. The 2,000-word and 3,000-word levels of Nation's

Vocabulary Levels Test, administered one week prior to the reading, suggested that the four groups were equivalent, although we shall consider this question further below. A reading text was chosen; it is described as relatively easy but for 29 words that would probably be unknown to most of the subjects. Sixteen of these words were selected as targeted items, with a further six items retained as dummy items. The text was then edited so as to replace the seven remaining unknown words and to adjust the frequency of occurrence in the text of the test and dummy items. After editing, the text length was just under 1,000 words.

For the Vocabulary Priming group, 15 minutes was allotted for vocabulary guessing exercises; this was included in the 30 minutes allowed for reading the text. The Marginal Glosses group were instructed to use the glosses as they read the text, and to read the text twice in the 30 minutes allowed. The Dictionaries group were instructed to use their dictionaries, and to spend the 30 minutes reading slowly through the text just one time. Finally, the instructions for the Text Only group were to read the text twice, and to try to guess the meanings of unknown words from the contexts.

After the reading session, the texts were collected. The subjects were then given two short multiple-choice tests: a comprehension test and a vocabulary test of the 16 targeted items. Two weeks later, a re-ordered and slightly different vocabulary test was administered. In all three tests there was a ‘don’t know’ option, and a penalty for wrong guessing. The test results suggest that there was no significant difference between any of the groups in terms of text comprehension. The results are summarised in Table 2.5.

Table 2.5

Results of Aizawa's (1999, study 1) test (total of 16 items per set)

	Immediate post-test	Test 2 weeks later
Dictionary	5.29	4.76
Text only	4.32	4.32
Vocabulary priming	8.68	6.92
Marginal glosses	7.99	6.23

As for vocabulary learning, average scores for both immediate and delayed tests for the Dictionary group (5.29 and 4.76 respectively) and the Text Only group (4.32 and 4.32) were significantly lower for than the Vocabulary Priming group (8.68 and 6.92) and the Marginal Glosses group (7.99 and 6.23). Differences between Dictionary group and Text Only group scores were not significant.

Comment

Before considering the results of this study, it should be noted that the experimental groups are composed of whole classes of students. As with Luppescu and Day's (1993) research, prior to the learning condition there may be significant differences between the groups in their knowledge of the small sets of targeted words. This effect may be all the more evident in studies where the vocabulary tested is grouped around a topic area, as in this test *stable*, *rein*, *stall* and *hitch* are related to horses. If one class had recently spent time on the topic, this may have a considerable effect on vocabulary test scores. As there was no pretest of the items in this study, it is not possible to evaluate how important this factor of varying prior knowledge of targeted words for different groups may have been in this particular study.

As we try to account for the low Dictionary group scores, only slightly better than those for the Text Only group, it is worth considering whether much dictionary use actually took place. As the experiment provides no direct information about this, we need to find another means of speculating regarding this question. Hulstijn et al's (1996) research has suggested that levels of vocabulary retention through successful use of bilingual dictionaries can be roughly equated with those obtained through using L1 marginal glosses. If we follow this reasoning, the difference in scores in this study between the Text Only group and the Marginal Glosses and Dictionary groups may provide an indication of the extent to which dictionary use for the targeted words took place. If we assume that the mean 3.37 word advantage of the Marginal Glosses group over the Text Only group is a reflection of reading all 16 glosses, the 0.97 advantage for the Dictionary group over the Text Only group represents an average of only about 4.5 successful lookups per subject for the 16 targeted words.

One further point worthy of comment is the time spent by each group. Although each group was allotted 30 minutes for the reading session, their respective tasks during this time were very different. For some tasks, 30 minutes may have been the optimum time, but for others it may have been too little or more than necessary. It may be misleading, then, to infer that having an equal amount of time for each learning condition provides a direct indication of their relative efficiency. We will now go on the second of Aizawa's studies to be considered here.

Aizawa 2

Summary

This study focuses on the learning of vocabulary from English L2 texts by Japanese high school students. This time, although the stated aim is still to investigate incidental vocabulary acquisition, target words are underlined and numbered to ensure that they stand out from the text so that learners will be guided to words worth spending time on, in the hope of increasing the apparently low level of successful dictionary use reported in the previous study. The specific issues addressed here are whether bilingual dictionary use helps learners read an L2 text more accurately, whether this dictionary use increases the retention of unknown words, and whether vocabulary acquisition levels are higher for more proficient learners both with and without dictionaries.

A total of 308 high school students took part in this study. The learners were divided into two nearly equivalent groups according to their scores in a vocabulary test, with one group assigned a reading with dictionary condition and the other group a reading without dictionary condition.

Two texts were selected as reading passages, and both edited to include 12 target words for which the meanings should be inferable from context. The texts were each accompanied by eight comprehension questions. Forty minutes was allowed for reading the two texts and answering the 16 questions. Following this, after materials had been collected, the participants were given a surprise 10-minute multiple-choice definition test of the 24 target words. The same test, with test items reordered, was conducted two weeks later.

The results for the reading comprehension questions showed that the Text Only group scored slightly higher than the Dictionary group. For the immediate vocabulary test, the Dictionary group's scores were almost 50% higher than the Text Only group's scores (a mean score of 15.60 out of 24 as against 10.88), while in the delayed vocabulary test, the Dictionary group still scored significantly higher than the Text Only group, but with a much reduced difference between the two groups (13.01 as against 11.42). While the Dictionary group's score fell over the two week interval, the Text Only group's score rose. When the groups were divided into higher proficiency and lower proficiency groups, of most note is that for higher proficiency learners there was almost no difference in the delayed vocabulary test scores between the Dictionary group and the Text Only group. We will consider why this may be below.

Comment

Various aspects of the above test results are counter-intuitive and require further examination: the higher reading comprehension scores for the group without access to dictionaries; the markedly higher vocabulary test scores for the Dictionary group; and the rise in the Text Only group's vocabulary test scores in the delayed test.

Aizawa suggests two possible reasons for the comprehension test results: i) that the Dictionary group participants spent too much time using the dictionary to allow enough time for the comprehension questions or ii) that the participants were not good or efficient dictionary users. While the first point does provide a likely explanation of these scores, confirmation of this would be available if scores for each question were provided, showing that Dictionary group participants either answered later questions less well or failed to answer them. These data are not provided. As for the second point, the author expands on this by suggesting that learners will often look up a word without trying first to guess its meaning from the context or trying to relate the meaning in the dictionary with that in the text.

While participants may well have looked up targeted words in the text without prior guessing of their meaning and settled for the first sense listed for polysemous words, this is not likely to have adversely affected their scores. This is because for over 20 of the 24 targeted words, the sense in the text and tested in the vocabulary tests is the first or only sense generally given in the most widely used high school or college level English-Japanese dictionaries. This means that even without checking that the first sense they encounter matches that in the text, dictionary users' focusing on the first sense will usually be to their advantage for these vocabulary tests.

The markedly higher scores for the Dictionary group in terms of vocabulary retention are not in themselves surprising but do contrast unexpectedly with their low comprehension test scores and with those in Aizawa's other study reported above. This might be explained by the underlining and numbering of the targeted items in this study and not in the other. In this study, the combination of low text comprehension scores and higher vocabulary retention scores seems to confirm that for the Dictionary group the reading task has become a vocabulary learning task, whatever the intention of the researcher. Although the distinction between incidental and intentional vocabulary learning is fuzzy and open to discussion, as we will see in Laufer and Hill's (2000) paper, the dictionary use and high vocabulary test scores of the Dictionary group participants in this study suggests that for them their vocabulary acquisition is intentional.

Finally, how do we account for the rise in the delayed vocabulary test scores for the Text Only group participants? The relatively small number of test items, the use of highly motivated participants, and the particular rise among more proficient Text Only participants all suggest that at least some of these participants would have looked up some of the targeted words between the two tests. The effect of looking up an average of only one or two targeted words would be sufficient to account for the rise in delayed vocabulary scores for the Text Only group.

This study demonstrates how small changes to the text, such as the highlighting of targeted items, can have a large effect on participants' dictionary use. Unfortunately, though, it also confirms the various problems inherent in using a small, unfocused, collection of targeted words.

2.3.11 Fraser (1999)

(Fraser, C.A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21 *(2),* 225-241.)

The main focus of this longitudinal study is the strategies L2 readers employ when they encounter unknown words in a text, including ignoring the words, inferring meaning from context or word form, and looking up the words in a dictionary. It also records levels of retention of word meanings acquired during the contact with the text.

Summary

This summary will begin with an outline of Fraser's study but will focus specifically on the issue of word meaning retention through inference from context or from the morphological form of words and from dictionary use.

A total of eight Francophone university students with intermediate proficiency in English took part in this study. These were drawn from a pool of 19 potential participants to represent lower and higher proficiency learners, based on scores for Nation's (1990) Vocabulary Levels Test and scores for a section of the Institutional TOEFL test.

The study was conducted before, during, and after an Academic English course in which there was a focus on L2 reading development, both in strategies and in language input. Over a period of five months, there were eight data-collection sessions. Each of these consisted of four stages. First, participants would study a set of comprehension questions for a text. They would then read the text and answer the comprehension questions, using bilingual and monolingual dictionaries as they pleased. Following this, the participants would skim through the text again and identify the words that had been previously unknown. Finally, each session would end with a structured interview about the strategies employed to arrive at the meaning of up to 15 of the previously unknown words in the text. One week later there was a vocabulary test, using the Vocabulary Knowledge Scale (Wesche and Paribakht, 1996), of 10 of the words focused on in the interview.

The results for retention of word knowledge need to be understood in the light of two considerations. One is that answers were rated according to whether they accorded with the meanings given in the interviews, even where these were mistaken. The other is that although the five-point Vocabulary Knowledge Scale was used, the scores were collapsed into two categories: recalled and not recalled. There are three sets of results relating to retention of word meanings. The first compares retention for words which were looked up, words whose meanings were inferred, and words whose meanings were inferred then looked up. We are told that for inference alone there was a retention rate of 31%, for lookup alone 30%, and for inference followed by lookup 50%. The second set of results compares types of inference: when the L1 was used as the basis for inference there was a retention rate of 50%, when the L2 was used 39%, and there was a 28% retention rate for inference from context. Finally, the researcher compares retention rates for words the participants recalled having seen before, 42%, and words they felt they had never seen before, 25%. We are only provided with percentages for any of the results. The numbers of responses these represent will be among the issues discussed below.

Comment

We will focus on two aspects of the results of this study: the high levels of retention in the different conditions investigated, and the numbers of word encounters represented by the percentages used to report results.

According to the figures reported above, retention rates for different learning strategies and word types range from 25% to 50%. Three factors seem likely to account for these high scores: the nature of subjects' encounters with the words; the training received in comprehension strategies as part of their studies in L2 reading development; and the counting of recalled wrong answers as retention.

One consequence of the method the researcher employs is that the participants have, effectively, three encounters with each targeted word prior to the retention test. The initial encounter is when subjects read the text in order to answer the comprehension questions. This may include dictionary use or some type of inferring. The next encounter is when subjects are asked to read through the text again and identify words that were previously unknown. The third encounter takes place during the interview, in which they are asked to give the meanings of these words and describe the process by which they gained an understanding of their meanings. Although the first of these encounters may legitimately be described as incidental, the identification and discussion of previously unknown words do seem to be separate, word-focused, activities. Together, the three encounters do help explain the high retention rates.

A second point is that this study takes place in the context of intensive training in inferring meaning and dictionary use. Not only would we expect participants' proficiency in these areas to increase with this training, as the results seem to bear out, we would also expect participants to deal with texts presented to them during this time from this perspective of lexical processing strategies. Both directly and indirectly, then, this focus would seem to be a factor in the high retention rates. Thirdly, retention rates are calculated according to the proportion of meanings given in the retention test that accord with those given in the interviews, rather than the number of meanings given in the retention test which are correct. If this latter standard were applied, we might expect retention rates to be about 80% of the figure given for words which were looked up and nearer to 70% of that for words whose meanings were inferred.

These last points lead to us consider what the results mean, and the numbers of word encounters on which they are based. For example, the percentages given for retention from inference and from dictionary use alone are almost identical but, as we have noted, the rate of successful comprehension for previously unknown items is considerably higher for dictionary use than for inference. Put simply, there was a higher rate of successful comprehension and retention through dictionary use than for inferring meaning, despite the impression of parity given by the test results. We will now look more closely at what other results may mean.

For none of the sets of results of the retention tests are we given any indication of the numbers of words for each category, but from inference we can gain some idea of how many words are involved. A total of 622 individual retention test answers were analysed. If the numbers of answers proportionately match the numbers of words which were reported as looked up, inferred, or ignored, we might expect this total of 622 words to be comprised of answers for about 20 items not noticed during the reading, 150 ignored items, 180 looked up items, and 275 inferred items, with about 135 of these last two sets of items inferred then looked up. This means that the percentage given for retention of items which were only looked up is based on under 50 encounters with unknown words, and that the percentages given for inference from L1 or L2 knowledge alone rely on similarly small numbers of encounters. The small numbers of items and of participants involved also help explain why no significant difference in retention between higher and lower proficiency groups was recorded.

In conclusion, although this study seems to demonstrate high rates of incidental vocabulary acquisition through reading, many factors conspire to make acquisition of unknown words through these encounters far from merely incidental. Rather, we should look at these figures in the context of focused vocabulary learning, and from this perspective the levels of vocabulary retention are not surprisingly high. Finally, the small numbers both of participants and of total encounters with words should alert us to the danger of placing too much importance on any but the two clearest of results: the benefit of inferring meaning followed by use of a dictionary (similar to that found in Iwai, 2000, below), and the increased likelihood for retention of words with which the participants have some initial familiarity.

2.3.12 Iwai (2000)

(Iwai, Y. (2000). *Dictionary use in context and vocabulary acquisition*, Unpublished MA dissertation. University of Birmingham.)

In his research, Iwai asks what difference there may be in the vocabulary acquisition of language learners reading an L2 text depending on whether or not they guess the meaning of unknown words from context prior to looking them up in a dictionary. Especially with the growing use of electronic dictionaries, coupled with the absence of training in dictionary use for most language learners, this question is an increasingly important issue but about which little other focused research has been conducted.

Summary

Twenty-four Japanese learners of English, in their final year of high school, took part in this study. They were all from the same class and were randomly divided into two groups. Class grades and a vocabulary test confirmed that the two groups were equivalent in terms of language proficiency. The experiment began with a brief questionnaire about dictionary use habits, followed by a reading passage of over 1,300 words from an English-language Japanese newspaper. Fifteen words in the passage were underlined and numbered. They were selected on the basis of being probably unknown to the students but among words that may be encountered in university entrance examinations.

The two groups of learners were given different instructions about how to deal with the target words when encountered in the text. The Guessing group learners were asked to identify the part of speech of each item, to look at the surrounding context in terms of syntactic and semantic relationships, to guess the meaning from the context, and then to look up the word in their English-Japanese dictionary and write it down. The No Guessing group learners were simply asked to give the meaning of the word if known, and if not to look it up in their English-Japanese dictionary and write it down.

Following the reading, the text and answer sheets for the targeted words were collected and the learners were given a simple test in which they were asked to give translation equivalents for the 15 target words. This test was administered three more times: one day, five days, and twenty days after the reading.

Reported prior knowledge of the target words averaged just over three words for the No Guessing group, 20% of the total, and just over two words (13.3%) for the Guessing group. Look-up figures per subject averaged over 9.5 (63%) of target words for the No Guessing group and under 6 (38.9%) target words for the Guessing group.

For the test results, figures were adjusted so that results for previously known words could be discounted. The test results presented by the author, and shown in Table 2.6, are only for target words that were looked up.

Table 2.6

Iwai's (2000) retention test results

	tl	t2	t3	t4
Guessing group	83%	75%	79%	68%
No guessing group	47%	40%	37%	33%

At t1, Guessing group participants gave an average of 83% correct Japanese equivalents for looked-up words, compared to 47% for such words by the No Guessing group participants. At t2, the respective figures were 75% and 40%, at t3 79% and 37%, and at t4 68% and 33%. In other words, Guessing group participants were, on average, able to give Japanese equivalents for near to twice as many looked-up target words as No Guessing group participants. Also, as might be expected, average retention rates for both groups fell steadily over the four sets of tests. We will consider these figures in more detail below.

Comment

The issue of how careful guessing of L2 word meanings from their contexts may affect acquisition has been addressed in other studies, but either not with regard to dictionary use (e.g., Schouten van Parreren, 1985) or only in terms of incidental acquisition (Aizawa, 1999). For Iwai in this study, however, conscious language learning is the purpose of the task, with the focus on text comprehension and vocabulary acquisition. This purpose is evident in the choice of target words – words that participants may need in the approaching university entrance exams – and in the dictionary use training included in the instructions for the Guessing group. This study is also unusual in that it shows clear differences in learning from the two learning conditions investigated, despite the relatively small numbers of both participants and target words. We will begin by considering how these results were arrived at.

Although the two groups are very similar in terms of language proficiency as far as can be judged by school grades, there are surprisingly large differences in the average number of words known prior to the experiment and in the number of words looked up. For the No Guessing group, an average of over three words, 20%, are correctly identified as previously known while the figure for the Guessing group is just over two words. The No Guessing group looked up an average of over 9.5 words (63%) while the Guessing group, on average, looked up under six words (38.9%). The author addresses this issue by pointing to the limited time available for the task and the longer time required for the guessing procedure prior to dictionary use as compared to using the dictionary without guessing. His response to this is to calculate scores for the two groups using only the data for words that were looked up. The problem with this is that for most participants in the Guessing group no more than the first eight or so items were dealt with, while all 12 items were attended to by the No Guessing group participants. With so small a number of test items, and with the words covered by the two groups differing so considerably, it is possible that any differences in retention may be attributable to the differences in words looked up, since words vary in the ease with which they may be learned, dictionary definitions in their clarity, and lexical contexts in the extent to which they are informative. One way of restoring the balance would be to focus on the results for the first eight items for each group. While such a move would further reduce the already small number of test items in the experiment, it would at least mean that the results for the same items for the two groups could be compared.

A further related issue is the differing learning load for the two groups resulting from the number of test items that were completed, with one group having the meanings of almost ten new words to remember and not confuse with each other and the other group having only six. Given this situation, it is also not inconceivable that highly motivated students would review or look up target words between tests. With such a small number of test items, looking up just a couple of words could easily have a substantial effect on subsequent test results.

The above problems could be rectified in further studies. If participants were given more time they would all be able to finish the tests and the two groups' results would be comparable. If there were not a single stretch of time available for a longer test, the study could, for example, be conducted with sets of words over a number of weeks as in Seibert's (1930) study. This study does, however, again demonstrate the dangers inherent in using a small number of vocabulary test items and, in this case, insufficient time for task completion.

2.3.13 Laufer and Hill (2000)

This study investigates the research applications of a kind of hybrid source of lexical information, half dictionary entry and half marginal gloss, increasingly common both in computer assisted language learning and in on-screen reading of electronic texts. The focus in this study is on vocabulary comprehension and retention in the context of reading comprehension.

Summary

In this study, a short reading text of about 120 words was used, and 12 single-sense words from the text were identified as words that the participants would be unlikely to know. These were highlighted in the text. Prior to the reading, the participants were presented, on-screen, with a list of the 12 target words and asked to give the meaning of any of the words they knew. Following this, the participants were allowed 10 minutes to read the text on-screen. They were told that a comprehension test would follow the reading, and were encouraged to look up the dictionary entries for the target words. This was done by clicking on the particular highlighted word. On doing this, they would be presented with three options: to see a monolingual dictionary entry for the highlighted word, to see the entry in the learner's L1 as in a bilingual dictionary, or to listen to the word's pronunciation. There was a test of the 12 target words immediately after the reading; as in the pre-test, the participants were asked to give the meanings of all words they knew. This post-test, however, was written on paper rather than conducted on-screen. Following this, there was a short comprehension test.

Initially, 97 students of English took part in this study. Of these, data for 25 was rejected as a pre-test indicated that they knew two or more of the target words prior to the study. The focus in this study, then is on 72 advanced level students of English, 32 Israeli and 40 Hong Kong Chinese, who knew none or only one of the target words before the study.

The vocabulary post-test results showed that the Israel group gave correct meanings for an overall average of 4.0 words, while the Hong Kong group gave correct meanings for an average of just under 7.5 words; this represents 33% and 62%, respectively, of the target words. These figures were adjusted to exclude any words correctly identified in the pre-test. The use of dictionary information accessed also varied widely between groups. The Israel group participants averaged over 1.5 lookups of some kind per word, while for the Hong Kong group this figure was just over 2.5. For the Israel group, the majority of lookups were for definitions in their L1, while the Hong Kong group participants typically accessed both L1 and L2 dictionary entries, with a large number also requesting pronunciation of the individual words. The most successful sources of information, at least in terms of raw numbers of previously unknown words known in the post-test, were the L2 definitions for the Israel group and L1 and L2 entries together for the Hong Kong group.

Comment

An interesting feature of this study is the similarities it bears with some earlier studies, despite its use of modern technology. These include the small number of target items, the intensive focus on the target words by the participants, and the high levels of retention of target words. We will look here at two main aspects of the paper: the claims regarding incidental vocabulary acquisition and the effect of accessing different types of dictionary information.

The authors go to some lengths to claim that they are investigating incidental vocabulary acquisition in this study, and although the validity of this claim depends to some degree on how incidental vocabulary acquisition is defined, in many respects it does seem hard to justify. From the design perspective, the following factors combine to give a focus on the target items that may for the participants have seemed far from incidental: the pre-test being limited to the 12 target items, the highlighting of the items in the text, the use of "dictionaries" containing only these words, the encouragement to look up these words, and the large amount of time made available for reading a very short text. As Wingate has observed (2002: 14), perhaps most important in determining whether acquisition is incidental are the perspectives and purposes of the participants; if for them the main purpose of the task is to increase their vocabulary, it would be hard to argue that vocabulary acquisition is incidental, regardless of the intentions of the researchers. In this study, the behaviour and apparent acquisition levels of the participants strongly suggest that for them the learning of the target words is a central, not incidental, concern. This is especially true for the Hong Kong group, who averaged over 2.5 lookups per word, often including the pronunciation of the words, and who in the post-test were able to give correct meanings of 7.45 of the 12 target words. This dictionary use seems excessive if the goal were only to prepare for a short comprehension test, even given the ease of accessing dictionary information on-screen as compared with using printed dictionaries.

Whether or not we accept the vocabulary acquisition described here to be incidental, we do need to ask why it is apparently so high as compared with other studies. The participants' intensive focus on the target words is surely an important factor in accounting for the exceptionally high post-test scores for both groups. Clearly, acquisition of the target words was for many of the participants in this experiment the purpose of the task, and they were given ample time to focus on learning the target words. The use of a text much shorter than this paragraph (about 120 words, as compared to over 1,300 words in Hulstijn et al's 1996 study or about 500 words in Knight's 1994 study) will also increase the focus on the target words. A further factor might be the use of an on-screen pre-test and a paper post-test; especially where different scripts are used, as with Hebrew and Chinese, recording the meaning equivalent using a computer is a more complex and time-consuming task than writing it on paper. The medium for the test itself may account in part for the higher post-test scores is alone may produce higher scores for the post-test than the pre-test.

Although the authors suggest that there is little or no correlation between the number of lookups and post-test scores, at least for in-group comparisons, there would clearly be a correlation if lookups and test scores for all 72 participants were investigated together. Put simply, both the Hong Kong group's lookup rate per word and their average post-test scores are almost twice those of the Israel group. This relationship is confirmed when we consider other research in terms of lookup rates and vocabulary retention. In other studies, with usually no more than one lookup per word, the highest retention rates are between 20% and 25% for looked up words. The only exceptions are studies where there is a deliberate focus on word learning, such as Seibert's (1930), Iwai's (2000) and, arguably, Aizawa's (1999). In this study, the Israel group's score of 33% for an average of 1.5 lookups per word and the Hong Kong group's score of 62% for an average 2.5 lookups suggest that there is a definite link between lookup rates and acquisition rates.

As for which type of accessed dictionary information is most likely to lead to acquisition, the situation is more complex than that reflected by raw figures showing information types accessed and retention success rate. For example, we are shown that for the Hong Kong group the highest "success rate", of 79%, is for accessing monolingual L2 entries alone. In addition to the information source, we would also have to consider what types of learners might typically access this type of information and feel no need to refer to bilingual dictionaries only a click of a mouse away. We would also need to ask which types of word are different types of information typically sought. As far as learner type profiles are concerned, these might include information about L2 language proficiency, skill at using dictionaries, and degree of conscientiousness as a language student. Word types may be categorised according to degree of prior familiarity, of perceived relevance to the learner, according to whether they are abstract or concrete, and for different parts of speech. No information is available regarding any of these factors which may affect learner dictionary use and acquisition rate.

This study, then, does give an indication of the high levels of acquisition that can be achieved when learners focus intensively on learning a small set of words with the aid of lexical information of the type found in dictionaries. It points towards the effects of different types of information or multiple lookups, but does not provide sufficient data to do more than this. Finally, as far as the effect the technology may have on dictionary user behaviour is concerned, the study does demonstrate the large number of lookups that can be made in a short space of time and how students may choose to access and use various combinations of information types according to their purposes, motivation and proficiency levels.

2.4 Discussion

Vocabulary acquisition through dictionary use is a shared concern for the studies reviewed above. They have, basically, the same goal of investigating what the effect of looking up an L2 word may have on the retention of that word. They also experience the same challenges and obstacles to investigating L2 vocabulary acquisition in this context. Almost all the studies face two challenges found in other areas of vocabulary acquisition research but exacerbated by the particular circumstances of research into dictionary use: how to investigate a sufficient number of targeted words, given the time that each lookup task may require, and how to find sufficient numbers of language learners willing to take part in these time-consuming studies.

In many other respects, the studies vary: in the specific issues they address, in their learning environments, in the methods they employ, and in the materials and tasks that form the context in which dictionary use and vocabulary acquisition are investigated. In this section, we will consider the studies reviewed above together in order to gain a clearer understanding of what they have in common and of the respects in which they differ. We will review the shared and individual aspects of the studies by addressing the following questions that have arisen through the commentaries provided above:

- Under what experimental conditions does dictionary use takes place? How much dictionary use is there? How is dictionary use recorded?
- What methods are employed to measure word knowledge? What types of word knowledge are being measured? How sensitive are these measures? How do studies measure word knowledge?
- Are post-test differences between groups attributable to experimental conditions?
- How does the number of participants and of targeted, looked up, and learned words affect the value of the studies?

2.4.1 The use and recording of dictionary use in experiments

All the studies reviewed, with the partial exception of Seibert (1930) and Fischer (1994), investigate L2 vocabulary acquisition through dictionary use. These can be divided into two main types: those in which the researcher's stated purpose for the participants is intentional vocabulary learning (Grinstead, 1915; Seibert, 1930; Black, 1986; Fischer, 1994; Iwai 2000) and those in which whatever vocabulary acquisition takes place is seen as incidental to a different stated purpose for the learners (the first of Aizawa's studies, 1999; Bogaards, 1992; Hulstijn et al.; 1996; Knight, 1994; Krantz, 1991; Laufer and Hill, 2000; Luppescu and Day, 1993). Although, as we have seen, the division between these types is not always clear-cut, especially from the participants' perspectives, it is a good starting point for considering how dictionaries are used in the studies. For what we may term the incidental acquisition studies, there is not usually an assumption that all targeted words will be looked up during the task. In some of these studies dictionary use is recorded, either electronically (Krantz, 1991; Knight, 1994;

Laufer and Hill, 2000) or after the studies (Bogaards, 1992; Hulstijn et al., 1996; Iwai, 2000). In others (Luppescu and Day, 1993 and the first reported from Aizawa's 1999 thesis) no record of dictionary use is made, and the only 'evidence' of dictionary use is through differences between experimental groups in test results and time taken to complete the task. Where recorded, the extent of dictionary use in these studies ranges between an average of under two out of 17 targeted words looked up (Hulstijn et al., 1996) and an average of 2.5 lookups per targeted word (Laufer and Hill, 2000). Cited factors affecting dictionary use include the length of the reading passage, the proficiency levels of the participants, the type of dictionary being used, the highlighting of targeted words, and the extent to which participants see the goal of their task as learning the targeted vocabulary.

For the intentional vocabulary learning studies (Grinstead, 1915; Seibert, 1930; Black, 1986; Fischer, 1994; Iwai 2000), there is an assumption, or requirement, that participants will look up all target words in whatever form of dictionary is made available to them. Perhaps as a result of this, there is no record in these studies of dictionary use. The dictionaries, or dictionary information, used in these studies may be in the form of whole dictionaries, sets of dictionary entries in a booklet, or sets of parts of dictionary entries, such as definitions or example sentences.

2.4.2 Measurement of word knowledge

The most widely used test of vocabulary knowledge, employed in nine of the studies, consists of a simple list of the targeted words for which participants are asked to give the meaning. In most studies, participants are required to give the meaning of targeted

L2 words in their L1 but for some studies L2 synonyms or definitions, or even pictures, are counted as acceptable. In one study, the words were read aloud and the testee responded orally. Otherwise, all the tests were conducted in written form.

Tests requiring participants to give the meaning of isolated words clearly demand a high level of accuracy and confidence about the meaning of the target words, most of which are rare words only encountered once in a text or only looked up once. This type of test is especially demanding where there is a strict rating of answers, such as only accepting dictionary-listed translation equivalents as answers for test words. In an effort to increase the sensitivity of the test, more lenient standards of rating may be applied, but this creates its own problems. For example, if for a test item the participants were required to give an L1 equivalent of *swan*, raters would need to decide whether *big white bird* would be acceptable, or *big bird*, or just *bird,* or even *animal* since, arguably, all are correct if not complete.

Perhaps the greatest weakness of word lists as tests is that they may be insufficiently sensitive to record the often small changes in word knowledge, or in confidence about word knowledge, that take place as a result of single encounters with unknown words in texts or dictionary entries. This may explain why the next most widely used test type, used in five studies, is the multiple-choice test. This type of test measures recognition knowledge of the words, usually a less demanding type of knowledge than the production of equivalents for words in a list. As with the word lists, almost all the multiple choice tests in these studies focus on word meanings, with one correct answer for each target word and three distracters. In three of these tests, there is also a *don't* *know* option available to the testees. In all but one of the studies, the answers and distracters are given in the testees' L1.

Multiple-choice tests have two main attractions: their relative sensitivity, as mentioned above, and the apparent ease with which they can be constructed. This second point is largely illusory since, as Wesche and Paribakht (1996) point out in their critique of multiple-choice vocabulary tests, they require extensive field-testing in order to establish their reliability. There is no record of prior field-testing of items having been undertaken in any of the studies reviewed here.

In addition, in these studies, the following problems have been observed with regard to their use of multiple-choice tests: the use of misleading or incorrect answers and distracters (Luppescu and Day, 1993); the failure to use effective distracters (Knight, 1994); the wide variation between answers and distracters for different test items (Black, 1986); failure to account for the 25% chance of guessing the correct answer where there is a forced choice of one of four answers (Black, 1986); and the problem with interpreting results when a *don't know* option is included (Luppescu and Day, 1993; Knight, 1994; Aizawa, 1999).

A further four test types are used in these studies. There are tests requiring participants to recall the target words or their translation equivalents after seeing the original context (Seibert, 1930; Hulstijn et al., 1996) and a test requiring evaluation of confidence about stated word knowledge (Black, 1986). There is a Vocabulary Knowledge Scale test (Wesche and Paribakht, 1996, used by Fraser,1999) in which participants would state how well they knew each word and give the meaning or write a sentence demonstrating that they could do this, and a test requiring participants to write a sentence for each target word then to translate the sentences (Fischer, 1994).

The use of the context of words (Seibert, 1930; Hulstijn et al., 1996) is a reasonable method for increasing the sensitivity of word tests, and has been used surprisingly little in investigations of learning through dictionary use. This may be in part because it could be argued that needing to see a word in context to be able to give its meaning is not evidence of acquisition but of ability to infer meaning or of remembering the encounter with the word rather than its meaning. Conversely, since words are typically encountered and produced in contexts, their being tested in this way may be seen as having greater validity than testing words in isolation. Black's (1986) use of a measure of confidence about lexical knowledge is a promising direction for research, except that no data are provided for this. Finally, while writing sentences as evidence of knowledge is valuable in that it provides a means for the testing of more than just knowledge of word meaning, it does have a number of drawbacks. These include the high level of confidence and lexical knowledge required to perform this task, the time required for each item and, as a consequence, the small number of items that can be tested, and the problems in achieving interrater reliability when evaluating these sentences.

2.4.3 Are post-test differences attributable to experimental conditions?

There are two main methods used in the studies to ensure that any post-test differences between groups can be attributed to the experimental conditions of the studies: administering a pretest to identify prior knowledge of targeted words or using experimental groups whose comparability is assumed or has been established. Both approaches, at least as they were implemented in these studies, have their weaknesses. With pretests, except where the pretest included large numbers of items (e.g. 148 in Krantz's (1990) study or 80 in Knight's (1994)), there is a danger that the presentation of the targeted words in the pretest may significantly affect the extent to which these words are noticed, looked up or learned. This would apply, for example, to Laufer and Hill's study (2000), in which only the 12 targeted items were included in the pretest. Another problem with pretests in which participants are asked to give the meaning of targeted words is that the form of test used as a post-test may be a different, or more sensitive measure, of vocabulary knowledge. These include the participants doing the pretest on computer and the post-test as a paper test (Laufer and Hill, 2000) or having a larger number of items in the pretest than in the post-test (Knight, 1994). From the participants' perspective, too, even identical tests may be seen and answered differently: the pretest being a survey of word knowledge and the post-test being a check of knowledge of items encountered in the preceding reading or task.

As for establishing the comparability of experimental groups, the various researchers employ a range of approaches. In Luppescu and Day (1993), the first of Aizawa's experiments in this review (1999) and, possibly, Bogaards (1992), pre-established classes of students were used as experimental groups. In Bogaards' study, the validity of using pre-established groups is backed up by students' grades for previous assignments which suggest that the groups are comparable in terms of proficiency in the target language. For the other two studies, however, there is an assumption that classes with students of the same level will share the same extent of lexical knowledge for a set of

L2 words. As we have seen, especially where targeted words are grouped around a particular semantic field, there is a clear risk that knowledge of the targeted words would vary substantially from class to class. Other methods include the random division of participants into groups (Hulstijn et al, 1996; the second of Aizawa's studies, 1999; and Iwai, 2000) and the use of student grades with the aim of ensuring that the groups are comparable in terms of language proficiency (Bogaards, 1992; Knight, 1994; the second of Aizawa's studies, 1999; Iwai, 2000; and Laufer and Hill, 2000).

Two other methods are used to circumvent the problems associated with using pretests or forming comparable groups: asking participants to report previous lexical knowledge after the learning session (Hulstijn et al, 1996; Iwai, 2000) and the rotation of groups among the different learning conditions (Seibert, 1930; Black, 1986). Post-experiment self-reporting avoids the danger of raising awareness of targeted words during the experimental period. It also helps overcome the problem of under-reporting lexical knowledge that may occur in a pretest as a result of not recognising known words. This may be especially relevant to participants whose first language uses a different script from a second language, such as Chinese or Arabic learners of English. A possible disadvantage of this method is that cultural differences between national groups of learners would mean that some groups would over-report or under-report their prior word knowledge. On the other hand, this argument could also be applied to pretests; in some L1 cultures learners may be more willing than others to guess when not sure, while learners in other cultures may be less able to recognise known words presented without a context.

The rotation of groups among the different learning conditions (Seibert, 1930; Black, 1986) helps ensure that subject responses for different conditions are balanced, especially where there are three or four conditions being examined at the same time. This method has three potential weaknesses. One is that switching from one stipulated learning condition to another every few items is very unnatural learner behaviour unless, as in Seibert's 1930 study, there is a series of separate learning sessions with one condition for each session. Another problem is that analysis of results for individual items or participants becomes less straightforward; in fact, in both studies in which rotation is used, only raw figures for results are used. A third potential problem with this method is that there will be only a small number of participants for each condition for each item, and with a small number of target items we should not expect results to show significant differences between learning conditions. In Black's study (1986), this means that, despite having a total of 24 participants, for each of the 24 items there are only six responses for each of the four learning conditions; this is reflected in the lack of significant differences in results for the different learning conditions. Seibert overcomes this problem by using many more participants, 60, and a total of 48 items, producing a total of 720 responses as compared to only 144 for Black. This means that for each learning condition there are 180 responses in Seibert's study, compared to only 36 in Black's.

We will now go on to consider further the effects of the different numbers of targeted words and participants on the results of the studies.

2.4.4 Numbers of targeted words and participants in each study

There is a very wide range in the numbers of targeted words and of participants in these studies. The range of target words is between 12 and 148 items, with the vast majority of studies focusing on between 12 and 24 items. Obviously, there are constraints affecting the number of items that can be treated in the studies, principally of time available but also of possible test-taking fatigue when there are too many items or too much time is spent on testing. Where dictionary use, extensive reading, and writing and translating sentences is involved, for example, each item may require as much as ten minutes of participants' time. This is one factor accounting for the small number of items in some studies. Yet with only 12 items, for example, previous knowledge of just one of the items would account for over 8% of the total score, and in many studies differences between experimental groups are little more than this (Black, 1986; Krantz, 1990; Hulstijn et al, 1996; and the first of the studies in Aizawa, 1999). With small numbers of target words, where parity between groups has not been established specifically in terms of lexical knowledge, differences in previous knowledge of targeted words could easily be a significant factor affecting post-test results.

A further point worth considering is whether 12 targeted words, for example, are meant to represent the target language's lexicon, an aspect of the lexicon, or neither of these. This question is important since some types of words, such as action verbs or concrete nouns, are easier than others to understand, to remember, or to guess from context. When we add the dimension of dictionary use, some definitions will be easier to understand than others, some words will have one sense and others will have more, and some of the targeted words will be for the first or only sense in a dictionary entry while others may be for the second, third or fourth sense. A small number of targeted words is unlikely to be able to represent a balance of all of the above word types. This is not, surprisingly, a concern voiced in many of the studies reviewed. Yet this factor is important; were a different set of words used in these studies, this alone could easily produce significantly different results. While this is a sufficient argument for using larger numbers of target words, various pragmatic considerations limit the possible number of items, as we have seen. The alternative approach of focusing on one type of word, such as verbs or single-sense words seems to have been largely ignored, perhaps because this would limit the scope of any claims that could be made based on the findings of the studies.

Constraints affecting the possible number of participants in a study include the number of suitable, willing, and reliable language learners available to the researchers, the availability of resources such as computer terminals or even paper dictionaries, the time needed for rating participants' responses, the availability of competent raters, and the time available to the researchers for a given study. There may also be financial considerations if honoraria or gifts were given to participants or raters for their assistance, or if it were necessary to buy a dictionary for each participant's use. The importance of these different factors will depend on the nature of the study. For example, reliable student attendance may be a major factor with a longitudinal study and hardly an issue where a study is completed in one session.

There appears to be a general belief that the greater number of participants in a study, the better that study will be. Tono (2000: 63-64), for example, interprets increasing participant numbers in studies as one aspect of the maturing of research in the field as it becomes more scientific in the way in which it investigates empirical questions. In many of the studies reported here, the use of a greater number of participants would result in more reliable data with a greater likelihood of significant differences between results of experimental groups. On the other hand, the use of relatively large numbers of participants, as in Luppescu and Day's study (1993) or Aizawa's (1999), may produce data showing a significant difference between experimental groups while, at the same time, masking serious design errors in the study. Further, in the case of reading an L2 text with or without the use of a dictionary, the largest difference between groups is often the time needed to complete the task. It is true that the use of large groups may make it possible to identify small average differences in word comprehension or retention, but where dictionary use doubles the time taken to complete the task, small differences between groups would be largely irrelevant in terms of the efficiency of the learning environment. Statistical significance is important but it should not be confused with educational value.

A further consideration is the type of study being conducted. Typically, intensive longitudinal case studies such as Grinstead's (1915) can be conducted with just one subject because of the demands on the time of the researcher and the subject. Where raters are needed to evaluate participants' written production (Fischer, 1994; Bogaards, 1992), this consideration alone will restrict the number of participants that can be dealt with. This is even more of an issue where learning sessions are followed by individual interviews, as in Fraser's (1999) study.

2.5 Conclusion

As noted above, and in the individual reviews, there has been a wide range of research into vocabulary acquisition through dictionary use. While this research has revealed various valuable insights into the effect dictionary use has on L2 vocabulary acquisition, much of the research has been flawed in some way. In many cases, the most fundamental fault is the small number of targeted words; either small in comparison with the number of word types in the reading texts from which they were drawn, or small in relation to the learner's L2 lexicon.

The studies reviewed also show little concern for the different types of words that are encountered in text or looked up other than in terms of frequency in the reading texts; there is no focus on different parts of speech, for example, and no planned comparison of words with polysemous or monosemous dictionary entries. Many of the studies reviewed here, based on only a dozen or so targeted words, make claims about vocabulary acquisition through dictionary use that go beyond the limited scope of their data. Future research may be well-advised to make a more specific focus, such as on a particular part of speech or type of dictionary entry. In doing so, findings from such research will carry more weight and claims based on the findings can be made with greater justification.

Behind the failings listed above is a general unwillingness to consider the nature of vocabulary acquisition, whether with regard to the lexicon as a whole or in terms of changes in individual word knowledge. If, for example, we consider that much vocabulary development is understood to be incremental, especially incidental acquisition through reading, we may ask why the most widely used experimental designs consist of only a pretest, experimental treatment, and immediate post-test.

The studies described in this chapter have greatly increased our understanding of the effect of dictionary use on L2 vocabulary comprehension and retention. At the same time, we are still left with three central questions relating to future research in the field:

1. Would more careful application of the methods employed in the studies reported in this chapter be able to produce valuable information about L2 dictionary use and vocabulary acquisition?

2. Will more focused research, such as on different parts of speech or specific types of dictionary entry, give us a clearer picture of how words are learned in general and, specifically, how differences in dictionaries in these respects may affect vocabulary acquisition?

3. Do we need to recognise that vocabulary acquisition, whether through dictionary use or not, is very often not a clear-cut matter of know/not know? If this is so, shouldn't the means by which we approach vocabulary acquisition through dictionary use be qualitatively different from much of the research in this field that has been conducted to date?

The rest of this thesis will be devoted to addressing these questions.

Chapter Three: A Replication

3.1 Introduction

It is one thing to write a critical review of a study, such as those in the Literature Review, and quite another thing to conduct and write up a study oneself. Through conducting a replication of one of the studies reviewed in Chapter Two, we may gain a fuller appreciation of many of the challenges faced by research into L2 vocabulary acquisition through dictionary use. Among these challenges are the need to find suitable participants in sufficient numbers and to create similar experimental groups of participants, the conflict between increasing numbers of target words and the increasing demands on the time of participants, and the tasks of finding or creating suitable learning materials and sets of target words. In the replication reported in this chapter, and in the studies detailed in Chapters Four, Five and Six, the main aim is to see what can be learned about vocabulary acquisition through dictionary use largely using the same instruments and methods as those employed in previous studies. Put otherwise, we are asking whether the problems and limitations already noted are intrinsic to research conducted in these ways, or whether the problems may be overcome by making adjustments to the methods and instruments employed.

We shall now go on to the replication of the study by Ute Fischer, reported in her 1994 paper, "Learning words from context and dictionaries: an experimental comparison". There were two main reasons for choosing to replicate Fischer's study. One was that it is in many ways typical of recent studies into learning from dictionaries: it involves, basically, a dictionary use task preceded and followed by a test of word knowledge, it focuses on a small number of target items, and it is interested in single encounters with unknown words. The other reason for replicating this study was that it is more ambitious than many studies: aiming to investigate the effect of dictionary use on productive use of the target words in addition to knowledge of word meanings.

3.1.2 Outline of this study

In this replication of Fischer's 1994 experiment, the aim was to investigate and compare language learners' comprehension and learning of a set of previously unknown L2 (English) words from two different sources: from monolingual dictionary entries and from their use in an extract of a novel. Being a replication, the procedures for conducting this study were basically the same as Fischer's, as was the coding of the data. These are described below, and where the replication differs from Fischer's study, reasons for the differences are explained.

This replication shares the same goals as those addressed in the original experiment; namely, to investigate the following questions:

- Which of the two sources of information was most helpful to learners in their comprehension and use of the target words.

- How participants with both sets of materials (dictionary entries and text) used the materials.

- What strategies were employed by the participants from the three different groups in using the target words.

One further important goal, set out as one of the purposes of Fischer's research but not reported in her paper, is how use of the two sources of information may affect retention of the target words. In the discussion section of this chapter, we will consider, mainly, the comprehension and retention of the target words, and the effect of the materials provided to the experimental groups on word knowledge. We will also consider the relationship between comprehension and retention of previously unknown vocabulary.

3.2 The study

In this study, one group of participants received set of dictionary entries for the target words (Appendix 3.2), a second group received the extract of a novel in which the same target words were embedded (Appendix 3.1), and a third group received both dictionary entries and text. Participants in a fourth group, serving as a control group, were given the text but with the target words deleted. Fischer's reason for having this control group was to ensure that the text in which the target words were embedded would be sufficiently comprehensible for participants with access only to the text; if this were so, it would suggest that the context of the target words within the text would also be comprehensible and so aid comprehension of the meanings of the target words.

Each learning and testing session lasted a total of 90 minutes and consisted of three successive parts: a word recognition pre-test; a learning phase; and a word recognition post-test. As there was insufficient time in this study to conduct the whole experiment in one 90-minute class, the post-test was conducted one week later than the pre-test and learning phase. It differed in this respect from Fischer's study, in which the post-test was conducted immediately after the post-test. It was felt that this period of one week between the learning phase and the test of vocabulary recall would also provide a better reflection of vocabulary retention than an immediate post-test. We will consider what effect this delay may have had on retention test results in the discussion section of this chapter.

For the learning phase, the participants in each of the three experimental conditions were asked to read carefully through the information they had received and then to use each target word in an English sentence. After about 45 minutes, as participants indicated that they had completed this task, they were asked to write a Japanese translation for each of their English sentences (see bottom of Appendix 3.4; for the study this was given to participants on a separate slip of paper). Prior to this instruction there had been no mention of the translation task. Participants in the control group, who had received the version of the text with the target words deleted, were required to write a summary of the story in Japanese.

3.2.1 Participants

A total of 69 second year English major Japanese university students, aged between 19 and 21 years old and with about seven years of formal instruction in English, assisted with this study from beginning to end. In Fischer's study the participants were German high school students. The reason for this difference in age and academic level, apart from the issue of participant availability to the researcher, is that Japanese students of English are typically less proficient in English than German students of English of the same age. Especially with regard to reading, the lack of cognates in Japanese and the use of a different writing system render a typical reading text more difficult for Japanese learners of English than for their German counterparts. In consequence, the use of MLD entries, an authentic text, and a creative writing and translation task such as in Fischer's study would be beyond the ability of most Japanese high school students.

3.2.2 Learning materials

Even with university students, materials for this study were deliberately chosen for their greater accessibility than those used in Fischer's study. For the reading text, a passage was taken and adapted from the popular writer Sidney Sheldon's *Memories of Midnight* (1990) (Appendix 3.1) as opposed to the more intellectually demanding John Fowles' *The Collector* (1981). As in Fischer's study, the text was adapted to only contain simple syntactic constructions and to contain no difficult vocabulary other than the 12 target words. As the text was already written in an accessible style with little difficult vocabulary, relatively little editing was required.

In addition to the adapted extract of a novel in which the target words occur, the other source information about the target words was monolingual learner dictionary entries. The dictionary entries (see Appendix 3.2) were taken from the *Collins COBUILD English Dictionary*, 2nd Edition, (Sinclair et al., 1995). None of the participants were regular users of a monolingual learner dictionary. *COBUILD* was chosen because the participants had some familiarity with the layout and style of its entries; it had been used a few weeks prior to the study to introduce them as students to monolingual dictionaries. It was also judged to be easier to use than the *Oxford Advanced Learner's Dictionary* (Hornby, 1980) used in Fischer's study.

For the twelve words, only two of the definitions contained information about the context or register in which the words are typically used (*noisome*: formal; *diverting*: old-fashioned), although a total of eleven of the entries indicate semantic restrictions (*If you describe an army or sports team as invincible... Stagnant water is...*). We will return to this question later, but it is worth considering what the word *stagnant*, for example, means without reference to water, and what a test requiring recognition of these words in isolation may be asking of testees.

Entries for all words provided grammar codes, such as ADJ-GRADED or ADJn, indicating acceptable syntax for the word and its lexical environment. As with all grammar codes in monolingual learner dictionaries, however, there is no guarantee that participants would have understood the codes or made use of them when writing. For eleven of the twelve target words, the dictionary entries also contained example sentences or phrases illustrating typical word use.

3.2.3 Target words

In this study, the target words in this experiment were all adjectives rather than a mixture of nouns, verbs, and adjectives as used by Fischer. The main reason for focusing on just one part of speech was that there was no apparent rationale for having a collection of assorted words other than that, superficially at least, it might give the impression of being a balanced set of target words. (We will discuss this issue at greater length in Chapter Four.) Fischer also selected target words on the basis of their being difficult for German learners of English to comprehend; in this study there was no such stipulation. The words in this study were selected on the basis of their not being recognized by a group of students similar to the participants or being included in *Eigo Tango 2001* (Uryu et al., 1993), a semi-official list of words used by students preparing for university entrance examinations.

The 12 target words included in the text, all adjectives, were as follows:

> *acrid*, *armed*, *budding*, *diverting*, *invincible*, *noisome*, *petrified*,
> *relentless*, *stagnant*, *swollen*, *unerring*, *unnerving*.

With the exception of *armed*, all the words were labelled in COBUILD with one or no diamonds: beyond the 6,600 most common words in the corpus used by COBUILD.

In this study, five of the target words were already in the original text (*acrid*, *armed*, *budding*, *stagnant*, *unerring*) while the other seven were added to the text. In Fischer's study, none of the target words were originally included in the text. Although it was impossible to find a text in which 12 suitable words were included, it was felt worthwhile to use one in which at least some of the target words were not in artificially created contexts. We will consider difference between results for original words and inserted words in the Discussion section below.

3.2.4 Testing instrument and learning task

Participants' knowledge of the 12 target words was tested twice: once just before the learning phase and then, unexpected by the participants, again one week after. On both occasions participants were given the list of target words and asked to give the meaning of whichever words they could (Appendix 3.3). The instructions, written in Japanese, emphasised that they could respond in either Japanese or English. The purpose of the pre-test was to learn which of the target words were already known by the participants, and of the post-test to indicate what effect the exposure to, and use of, the learning materials may have had on their learning of previously unknown target words.

As for the learning task (Appendix 3.4), the participants in the experimental groups were asked to write sentences of their own using the twelve target words, one sentence for each word. Once finished, they were asked to translate their sentences into Japanese, as evidence that they had understood the meaning of the target words. This latter requirement was made because it is often possible to write an acceptable English sentence using the target word but without knowing its meaning; use, alone, is not evidence of comprehension.

3.3 Rating the data

There were three sets of data which needed to be evaluated: control participants' summaries of the texts, coding for acceptability of the English sentences produced by the experimental group participants, and the coding for accuracy of the Japanese translation*s* of these sentences and of the translation equivalents for the target words. In addition, raters judged which source of data (text or dictionary entries) and which strategies the participants used in writing the English sentences.

3.3.1 Evaluation of control participants' summaries

Two native speaker teachers of English independently made up a list of important points that they felt should be included in a summary of the story. Eight facts were considered important by both teachers. They then acted as raters and evaluated how many facts from this list were stated in the summaries of the control group participants. Each fact was counted as one point; partially recorded facts were counted as 0.5 points. The raters differed in their judgment of the texts by one point or less for 86% of the summaries.

Overall, 55% of the control participants mentioned five or more of the eight facts deemed to be important in the narrative. Of the remaining eight participants, six mentioned more than 50% of the important points. These results suggest that participants would have had a fair understanding of the story. The implication of this level of understanding, following Fischer's reasoning, is that the text would have provided a largely comprehensible context for the unknown words for the participants who were provided with it.

3.3.2 Rating the English sentences

The same two native speaker teachers of English independently rated the use of the target words in the English sentences in the following way: whether the target word was used in an idiomatically meaningful way, whether its usage was questionable, or whether it was used in idiomatically unacceptable contexts. Sentences were each given a code number and contained no indication regarding which experimental group the writer belonged to. Interrater reliability was .68; disagreements between raters were settled through discussion.

3.3.3 Rating the Japanese translations

A native speaker of Japanese evaluated how well the participants' translations matched a standard, namely the Japanese equivalent of a target word as stated in bilingual dictionaries: *Kenkyusha's English–Japanese Dictionary for the General Reader* (Matsuda et al., 1992) and *An Encyclopedic Supplement to the Dictionary for the General Reader* (Matsuda et al., 1994). The accuracy of the translations was judged to be a match, a near-match, a far-match, or a no-match of the English target word. A monolingual Japanese dictionary, *Kojien Dainihan Hoteiban* (Shinmura et al., 1976), was also used to determine the adequacy of participants' translations. After the rater had completed the rating, a second rater, also a native speaker of Japanese, was asked to check the judgements of the first rater. The raters agreed in 75% of the instances. Disagreements were resolved through discussion.

If a translation was equivalent to a standard, it was called a match. A translation that was not semantically related to the standard was a no-match. A translation was rated as a near-match if it was a superordinate of the standard. For example 汚い水 ("kitanai mizu": *dirty water*) as a translation for *stagnant (water)*. If the meaning of the standard shared part of the meaning of the translation, or vice versa, and their semantic relation could not be characterised in terms of hyponymy, then the translation was classified as a far-match; for example, *noisome* rendered as ひどくつまらない ("hidoku tsumaranai": *terribly boring*). In this case, both words were expressing an unpleasant quality, although of a different kind.

3.3.4 Coding the strategies

The same procedure was employed for coding the perceived writing strategies as for the rating of the Japanese translations. The first rater coded the strategies that the participants in the experimental conditions seemed to have employed, and the second rater checked the ratings. Participants' strategies were inferred from their translations of the target words and the learning materials they had seen. The raters agreed 80% of the time. Again, all disagreements were resolved through discussion. The strategies types identified in Dictionary group participants' writing were as listed in Table 3.1.

Table 3.1
Dictionary group comprehension strategies

Strategy	Description	Example	Dictionary entry
Complete string substitution	Translation utilises complete definition	*I was petrified because I saw a ghost.*	If you are petrified, you are extremely frightened, perhaps so frightened that you cannot think or move. *I've always been petrified of being alone. Most people seem to be petrified of snakes.*
Substring substitution	Translation uses part of definition	After exercising a soccer for a long time I am always petrified.	
Modelling	Sentence is modelled after definition or example	She seem to be petrified of ghosts.	
Unfinished substitution error	Sentence includes part of definition (but not the target word)	Our team will not *unbeatable* in the game.	**invincible** *=unbeatable* [given as synonym in entry]
False positive	Translation equivalent is given for a word with typography similar to the target word	Since it has been 50's the earth is being influence by acrid rain. (*acrid* mistaken for *acid*)	**acrid**

The purpose of focusing on comprehension strategies is to gain a greater understanding of the ways in which the participants made use of the different information in dictionary entries and in contexts of target words to understand and use the target words.

Dictionary group

The occurrence of a substitution strategy was noted whenever the translation of a target word was clearly based on all or part of the definitional information. These were classified as complete string substitutions and substring substitutions, respectively. Responses in which participants directly incorporated information from either definitions or illustrative phrases were termed as copying or modelling.

Copying was coded when an English sentence involved part of the definition or an example. Beside using information verbatim, participants also modelled their sentences after an example or the definition. A false positive was coded as a strategy when the translation (e.g. うるさい – "urusai": *noisy*, for *noisome*) was arrived at by the mistaken linking of the target word with a similar known English word. This type of strategy is different from the one in Fischer; as there are no real English-Japanese cognates there was no confusion in this area. On the other hand, it was not unusual for there to be confusion between similar sounding English words, whether known directly through English or as loanwords in Japanese, so this became the 'False Positive' category. Unfinished substitution errors referred to English sentences that failed to include the target word but included part of its definition instead.

Text group

Table 3.2 lists the strategies for the Text group:

Table 3.2
Text group comprehension strategies

Strategy	Description	Example	Text
L+S	Sentence incorporates lexical context and translation accords with lexical and schematic context (よどんだ - 'yodonda' = stagnant)	*The docks stank of the stagnant mass of dead cats.*	*The docks stank of the stagnant mass of dead cats and dogs.*
L	Sentence incorporates lexical context and translation accords only with lexical context (間違いのない- 'machigai no nai' = without mistakes)	*She is an unerring typist.*	*...Tony's boxing skills and unerring killer instinct.*
S	Sentence and translation accord only with schematic context	*He knows of the unerring way of guns.*	
P	Sentence and translation accord only with part of the schematic context	*I think he went on relentless exercise every day.*	*...the school itself was a relentless battle ground.*

The English sentences written by participants with the text as source were rated according to whether they incorporated the lexical context in which the target word occurred. It was noted whether a target word's context was adopted verbatim or whether participants' sentences were modelled on the sentences in which the target words occurred. In addition, translations were judged according to whether they suited only the lexical context of a target word or whether they also accorded with their schematic context. If an English sentence did not incorporate the lexical context of a target word, it was judged whether the translation of the target word was consistent with all or only part of its schematic context. Translations that seemed unrelated to either the lexical or the schematic context were coded separately as unexplainable.

Regarding Mixed group participant strategies (Table 3.3), participants were assumed to have focused on the text when they used a target word in a construction similar to its lexical context in the text. Participants were judged to have considered both sources of information when they translated a target word in accordance with its definition and used it in an English sentence that was analogous to the text.

Table 3.3
Mixed group strategies

Strategy	Description	Example
Dictionary as source	Translation of definition, and/or copying and modelling from dictionary	*I believe the soccer team is invincible.*
Text as source	Sentence incorporates lexical context from text	*The soccer game made me invincible.*
Both sources	Translation accords with definition, and sentence similar to text: or two translations matching both sources	*The members think themselves invincible.*

3.4 Results

We will begin by looking at the overall results for the participants' retention of the target items, as recorded in the post-test. Following this, we will report prior knowledge of the individual test items as recorded in the pre-test. We will then go on to the two sets of results obtained through the learning task phase of the study, in which participants were required to produce an English sentence for each target word and then to translate the sentence into Japanese. Finally, we will report the sets of data concerned with observed participant strategies with regard to the use they made of the dictionary entries to write the English sentences.

3.4.1 Post-test results

Table 3.4 shows mean numbers of target words known in the post-test by participants in the three groups. Figures are also shown excluding results for the target word armed, since a relatively large number of participants correctly identified this word in the pre-test.

Table 3.4

Mean number of post-test scores (max = 12)

	Dictionary group	Mixed group	Text group
For all 12 words	2.47	2.56	1.18
Minus *armed*	1.53	1.81	0.53

As we can see, there is a clear difference between the figures for the Text group and those for the Dictionary and Mixed groups. We will consider what these results may tell us in the Discussion section below.

With the exception of one word (*armed*), the percentage of participants who gave an appropriate English synonym or Japanese translation to the target words in the pretest was very low; no other word was known by more than one subject in any group. Figures for pre-test knowledge of the target items are shown Table 3.5.

Table 3.5
Number of participants familiar with target words prior to the study

Word	Dictionary gp (N=16)	Mixed gp (N = 15)	Text gp (N = 17)
acrid	0	0	0
armed	7	3	3
budding	0	0	0
diverting	0	0	0
invincible	0	0	0
noisome	0	0	0
petrified	0	0	0
relentless	0	0	0
stagnant	1	0	0
swollen	1	0	0
unerring	1	1	1
unnerving	0	0	0

3.4.2 Adequacy of usage and comprehension of target words

Table 3.6 shows the percentages of appropriate and inappropriate uses of the target words in the English sentences that were obtained for each experimental group. As the table shows, overall, participants in the Dictionary and Mixed groups performed markedly better than participants in the Text group.

Table 3.6

Percentage of omissions, idiomatically incorrect, questionable, and correct uses

Adequacy	Dictionary group	Mixed group	Text group
Omissions	1	2	1
Incorrect	10	7	25
Questionable	25	28	34
Correct	63	60	40

Table 3.7 summarises how accurately participants in each experimental condition were judged to have translated the target words.

Table 3.7

Percentages of omission, no-, far-, near-, and matching translations

	Dictionary group	Mixed group	Text group
Omissions	1	2	1
No-match	17	11	63
Far-match	6	13	12
Near-match	29	23	10
Match	48	51	14

For each subject, the number of matching, near-, and far-matching translations were collapsed into one score, and labelled “correct translations”. The type of information that was available to participants influenced their comprehension. As shown in Table 3.8 and Figure 3.10, participants in the Dictionary and the Mixed groups gave, on average, 9.9 and 10.4 correct translations, respectively. The mean number of correct translations for the Text group was 4.3.

Table 3.8

Numbers of correct translations (max = 12)

Group	Dictionary group	Mixed group	Text group
Mean score	9.9	10.4	4.3
S.D.	1.5	2.2	2.3

For the English sentences, omissions and questionable or unacceptable English uses were collapsed into one category – bad usage. Table 3.9 and Figure 3.10 show that participants in the Dictionary and the Mixed groups gave, on average, 7.7 and 7.1 adequate English sentences, respectively, while the mean figure for the Text group was 4.6 adequate sentences.

Table 3.9

Numbers of adequate uses (max = 12)

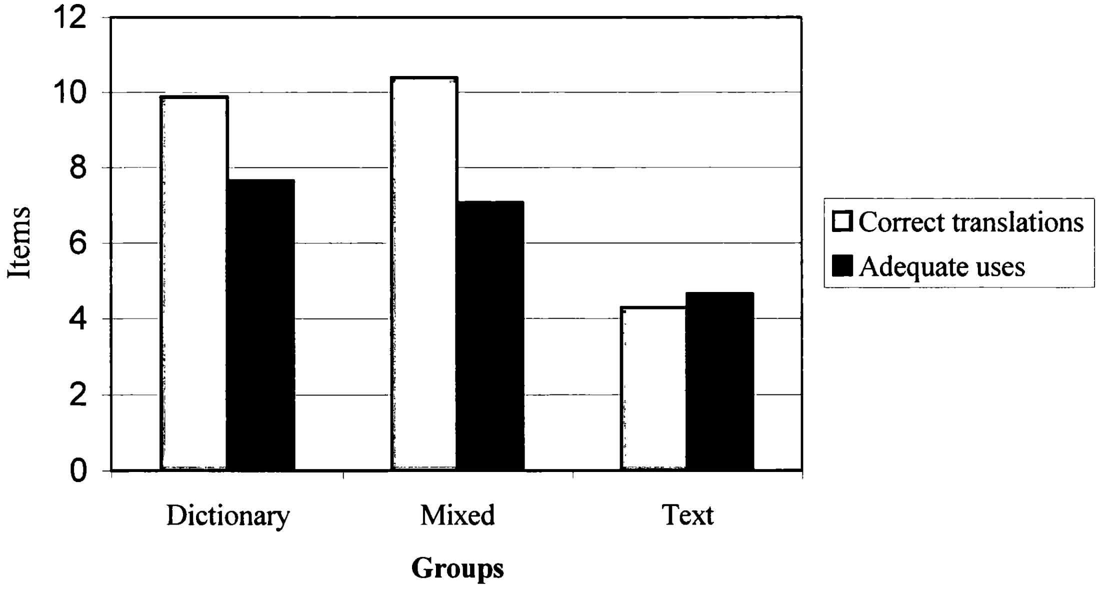
Group	Dictionary group	Mixed group	Text group
Mean score	7.7	7.1	4.6
S.D.	2.4	1.6	3.0

A one way analysis of variance was conducted on the ratings of the Japanese sentences per word for every subject within each group. For this analysis, matching, near-, and far-matching translations were collapsed into one score, as above, as "correct translations", with omissions and no-matching translations counted as "incorrect translations". For the translations, a one way analysis of variance showed that there was a significant difference between the groups [F (2,49) = 47.38, p<.001]. A Tukey test confirmed that the Text group performed worse than the two other groups, but that there was no significant difference between the Dictionary and Mixed groups.

A one way analysis of variance on the ratings of the English sentences per word was conducted for every subject within each group. Similarly, in this analysis omissions and questionable or unacceptable English uses were collapsed into one category – "unacceptable uses". Adequate uses were "acceptable uses". This analysis of the English sentences showed a significant difference between the groups [F (2,49) = 7.63, p<.001]. For the English sentences as well, Tukey tests confirmed that the Text group performed worse than the two other groups, but that there was no difference between the Dictionary and Mixed groups.

Figure 3.10

Average numbers of correct translations and English sentences (max = 12)

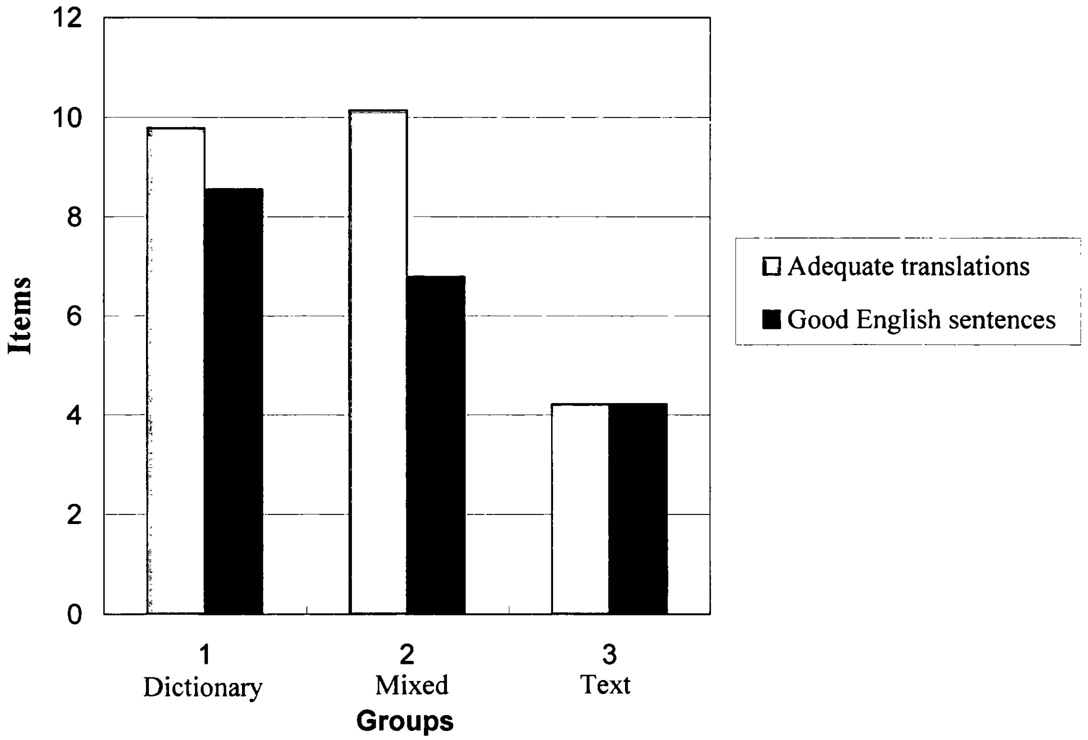


In order to eliminate the effect of words known by participants prior to exposure to the materials, the few participants who did know one or more of the words in the pretest were excluded from the calculations. With these revised groups, participants in the

Dictionary, Mixed, and Text groups gave, on average, 9.78, 10.14, and 4.21 correct translations, respectively. This is shown in Figure 3.11. For the English sentences, participants in the Dictionary and the Mixed groups gave, on average, 8.55 and 6.79 good English sentences, respectively, while the mean figure for the Text group was 4.21 good sentences.

Figure 3.11

Revised groups' adequate translations and acceptable English sentences (max = 12)



When a one way analysis of variance was performed on these new groups, there is still a significant difference between the groups' results, both for the translations, [F (2,34) = 27.95, p<.001], and for the English sentences: [F (2,34) = 10.21, p<.001]. Tukey tests with the revised groups' results confirm that the Text group performed worst of the three and that there was no significant difference between the Dictionary group and the

Mixed group.

Table 3.12 shows percentages of correct answers given in the post-test, taken one week after the participants had encountered the target words, written example sentences for the target words and then translated their example sentences.

Table 3.12

Correct identification of target words in the post-test

Word	Dictionary group (N=16)	Mixed group (N=15)	Text group (N=17)
acrid	5	3	1
armed	12	13	10
budding	1	1	0
diverting	1	0	1
invincible	1	3	0
noisome	3	3	2
petrified	0	0	0
relentless	1	1	1
stagnant	3	4	0
swollen	7	3	1
unerring	1	3	3
unnerving	3	2	0

3.4.3 Participants' strategies

In the Dictionary group, 87% of participants' responses were accounted for by the observed strategies, with participants most commonly adhering to a substitution strategy. More importantly, participants focused 50% of the time on all available definitional information. Substring substitutions, which indicate an inadequate understanding of a definition, was observed in 35% of responses. Copying, which always co-occurred with other strategies, was noted in 7% of responses, with modelling observed in 4% of participants' responses.

The strategies that were discerned for the Text group can explain 35% of the participants' responses. Some participants in the Text group use both the lexical and the schematic context of a word to infer its meaning, but this was evident in only 10% of sentences. In 12% of their sentences, participants preserved the sequence of a target word and adjacent words, while only 3% of their sentences accorded with the schematic context of the target words. A total of 64% of sentences could not be assigned to any of these categories.

Overall, 98% of the responses of the participants in the Mixed group could be explained in terms of the strategies identified in the coding scheme. Of their responses, 80% could only be traced to information in dictionary entries. They focused on the text only 2% of the time, and 16% of their responses were based on both dictionary and text. For the sentences in the Mixed group where they used the dictionary entries, the frequencies of complete string substitution and substring substitution are comparable to the ones observed for participants in the dictionary-only group: 54% and 35% respectively.

We will now go on to consider the implications of the results, focusing specifically on issues relating to the comprehension and retention of previously unknown L2 words. We will also consider how, and why, these results compare with those in Fischer's study.

3.5 Discussion of results

We will begin by looking at two sets of results: the scores for English sentences using the target words and the translations of these sentences for the three experimental groups, since these are the data focused on in Fischer's study. We will compare these results with those for Fischer's study and also consider what these data may mean in terms of lexical knowledge. This will lead us to the data relating to word knowledge retention in our study; here we will consider retention rates in relation to the comprehension rates for target words as demonstrated through English sentences and translations.

3.5.1 Comprehension

As Figures 3.10 and 3.11 show, the two groups of learners with access to dictionary entries in this study clearly produced the largest number of accurate uses of the target words: on average almost twice as many as the text only group. As for acceptable translations of the sentences, which may be a more reliable indication of comprehension than the sentences themselves, the contrast between the groups with dictionary entries and the text only group is even more striking; the average number of acceptable translations for this latter group was less than half that of the two other groups. It is not surprising that the groups with dictionary entries should perform better in terms of comprehension of previously unknown words. One central purpose of dictionaries is to provide comprehensible definitions for unknown words or senses, while a single written context for an unknown word may often be insufficiently informative for the reader to guess at the meaning even in his or her mother tongue.

In Fischer's study, there was no great advantage for comprehension of having access to dictionary entries; for the English sentences, the group with access to both dictionary entries and the text wrote, on average, the largest number of acceptable sentences of the three groups (39%), with the Dictionary and Text groups about equal (32%). As for translations, in Fischer's study, the two groups with access to dictionary entries managed significantly higher scores (42% and 39%) than the text only group (29%).

The most probable reason for the uniformly lower scores for Fischer's participants may be found in the deliberately difficult vocabulary items that she chose as target words for her study. We may discount as unlikely the alternative reason that the German participants in Fischer's study may have had lower English proficiency levels than the participants in this study. As for the lack of significant difference between the numbers of acceptable sentences for the three groups in Fischer's study, this may be because for the text used by Fischer, as all the target words were inserted into the text, they could be presented in a clear and comprehensible context. In the text used in the replication, five of the words were already a natural part of the selected text. In terms of results, this set of original words did not stand out except for *armed*, which was known by many more participants than any other of the targeted words. In the Text group, for example, the average combined number of participants giving matching or near-matching equivalents for the original words was 4.75, with *armed* excluded from the calculation. This compares with an average of 3.0 for the inserted words. However, the original words in the text did serve as a standard for the inserted words, ensuring that the context of the inserted words was not made exceptionally clear or informative.

Another factor accounting for the relatively strong performance of the two groups with dictionary entries in the replication may be the difference between the two sets of dictionary entries. Two major differences between the replication and Fischer's study are the presence or absence of indications of semantic restriction for the words, and of example sentences or phrases in the dictionary entries. In the set of twelve entries for the replication, eleven provided clear indications of semantic restrictions (*Stagnant water is... An acrid smell or taste is...*) and eleven also included at least one, and often two or three, example sentences or long example phrases. In Fischer's case, using the 3rd Edition of the *Oxford Advanced Learner's Dictionary*, semantic restrictions are harder to identify and there are fewer of them: just two or three out of twelve definitions. Seven of the entries for words in Fischer's study contained an example phrase, but these were usually only two or three words long. These differences may help to account for relatively high comprehension rates in the replication for the two groups with access to dictionary entries.

3.5.2 Retention

We now come to consider retention rates for meanings of the target items in the replication. As noted before, these data are not reported in Fischer's paper; we will consider why this may be after reviewing the data on target word meaning retention obtained through the replication. Perhaps the most noteworthy aspect of the retention data are the uniformly low rates of correct responses given in the post-test. Excluding data for the already widely known item *armed*, the participants of the various groups were only able, on average, to provide the meaning of between just over half a word and under two words out of the eleven remaining target words. For the Text group, an average of about 0.5 of the remaining 11 words were known, with 1.8 words for the

Mixed group, and 1.5 words for the Dictionary group. These figures, however, do not take account of the fact that comprehension rates for meanings of target words were far from 100%; it is meaningless to talk of retention of word meaning in cases where there was no prior comprehension of the word. Since not all the target words were the meanings understood; it is more meaningful to talk about the retention of the meaning of previously unknown words if the words have been understood.

When we look at the numbers of retained word meanings as average proportions of target words for which comprehension was displayed, they are as follows: 15.6% for the Dictionary group, 17.8% for the Mixed group and 12.6% for the Text group. Because of the small numbers of items involved, and the relatively small numbers of participants in each group, we cannot expect to find significant differences between the three groups. However, there are clear differences in comprehension rates between the text only group and the two groups with dictionary entries, and these are reflected in the overall retention rates for previously unknown target words. While the data obtained through this replication confirms that monolingual learner dictionary use is more likely to lead to comprehension of unknown words than single encounters with the unknown words in a text, there is no evidence to suggest that retention rates for word meaning knowledge from these two sources differ greatly; in other words, the more words that participants understood, the more words for which they were able to retain meaning knowledge.

The low retention rates in the replication, on average about 14% of words for which comprehension was demonstrated, may give us a clue as to why word retention data may have been excluded from Fischer's paper. The correct translation rates for the groups in her study averaged between about four or five words per subject. If the retention rate of these words was similar to that in the replication, we might expect average retention rates for all groups to have been well under one word per subject: a figure that might, understandably, be omitted from the report of an experimental study.

3.6 Problems with the research

We will now go on to consider in some detail three problems with this kind of research as revealed in Fischer's study and through this replication: the low vocabulary retention rates, problems with inter-rater reliability, and the variability of the target items in various respects.

3.6.1 Low retention rates

In this replication, and in a number of other investigations of vocabulary retention resulting from dictionary use (e.g., Hulstijn, Hollander and Greidanus, 1996), retention rates for words looked up in dictionaries or consulted in dictionary entries have been remarkably low. This is especially worthy of note when, as in the replication, comprehension rates were high for the participants who used dictionary entries. Possible explanations for this may be found in two or three aspects of the nature of the L2 learner's lexicon and of the experiments undertaken and instruments employed. The first relates to the instruments employed. In this study, and in various others, evidence of retention was taken to be the difference between items for which participants gave a correct meaning in the pre-test and in the post-test. This is a test in which the items are presented in isolation, without any context, and for which the participants are required to provide a meaning of the target words. The participants only had one or two encounters with the target words through the experiment. While we may expect this to result in some incremental growth in word knowledge, the participants may still not have gained sufficient knowledge, or confidence, to provide evidence of productive knowledge of the target words. That they were able to do so for some words may indicate that they did learn a lot about the word through the study: perhaps because the nature of the tasks, writing a sentence using the words and then translating the words, involved paying intensive attention to each word for a number of minutes. Alternatively, it may indicate that the participants had some prior knowledge of some of the target words, and that the pretest was insufficiently sensitive an instrument to identify this prior knowledge.

A further possible reason for the low vocabulary retention rates may be the incidental nature of the acquisition that took place. As noted above, the tasks involved in Fischer's study and in the replication did involve intensive attention being paid to the target words and the materials in which they were presented. The focus of the tasks, however, was not to learn the target words; it was to use the words in English sentences then to translate the sentences. Although the distinction between incidental and intentional acquisition may not be as distinct as some researchers may like to believe (for example, Hulstijn, 1992 or Laufer and Hill, 2000), especially for research undertaken in a language learning environment, the lack of a focus on vocabulary learning in these studies may provide an additional explanation for the low retention rates. In other words, as there was no requirement for the participants to try to learn the words, and since they were unaware that the tasks would be followed by a test, a week later, they did not make any effort to learn the words. A more explicit focus on vocabulary learning may well result in higher levels of retention than in this replication.

3.6.2 Interrater reliability

One major problem with Fischer's study and with this replication is the low level of interrater reliability, especially as regards the assessment of the acceptability of the English sentences or the accuracy of the translation equivalents provided for the target words. Both for Fischer's study and the replication, for the English sentences and for the translations, raters' assessments differed for between one third and one quarter of the participants' responses. In both cases, the problem is that judging the acceptability of an English sentence or the accuracy of a translation equivalent is largely a subjective matter. To take a couple of examples from the replication, *noisome* was variously translated as unpleasant, boring, painful, unwelcome, or noisy, while *acrid* was used to describe bread, tomatoes, water, or tobacco smoke. The only guidelines the raters had were a brief set of instructions together with monolingual and bilingual dictionaries; these did, naturally, give meanings or equivalents of the words but, very often, did not provide much guidance as to restrictions regarding meaning or referents for the words.

One partial solution to the low levels of agreement between raters may be to give clearer, more detailed guidelines for rating the participants' responses. This does not, however, fully address the problem that such judgments are, to some extent, inherently subjective. We will see a further aspect of this when we consider the variability of different items from various perspectives. In terms of use or meaning, for example, we can see how some target words, such as *stagnant*, have very restricted meanings, while others such as *relentless* or *diverting*, have much wider uses; this may affect typical interrater reliability retention rates and, as we shall consider below, apparent retention rates.

3.6.3 Variability between items

In Fischer's study and in this study, there are four main factors affecting how comprehension and retention of the target items may differ from word to word: i) the context of the words in the written passage; ii) the dictionary entries for the words, iii) the nature of the words themselves, and iv) the words as seen from the perspective of the participants' first language. We will go on to consider how each of these factors may affect learners' comprehension of L2 words in the two contexts under investigation.

3.6.3.1 The context of the words in the written passage

Depending on the context surrounding the target words, the extent to which the text aids comprehension will differ from word to word. Two examples from the text used in the replication may illustrate this:

> a) The docks stank of the *stagnant* mass of dead cats and dogs.
>
> b) ...he was not ready for Tony Rizzoli's boxing skills and *unerring* killer instinct.

The context a) for *stagnant* may only detract from any previous knowledge that the participants may have had for the word, except to confirm that it means something unpleasant. In contrast, context b) for *unerring* is much more enlightening. This is confirmed by the widely differing comprehension rates among Text group participants for the different words, with numbers of acceptable translations from the 17 participants ranging between only 1 (for *invincible* and *budding*) and 10 (for *acrid*). As for retention of the items, although cognitive effort involved in understanding, using, and translating the target words may be a factor affecting retention, by far the greatest factor in this study appears to be whether or not the word is understood from the context.

3.6.3.2 The dictionary entries for the words

For the English monolingual dictionary entries, comprehension of the target words was obviously still a factor affecting rates of retention, but as comprehension rates were generally high for the two groups of participants using dictionary entries, we can see that various other factors may have a significant effect on the retention of the target words. First, though, we will consider what factors may be involved in assisting or hindering comprehension. Factors that may assist comprehension include the following: a) the clarity of the definition; b) expression of the meaning more than once, in different ways; c) the avoidance of words unknown to the reader; d), the provision of illustrations in the definition; e) the length of the definition; f) clear indication of semantic restrictions; and g) example sentences that help clarify the use and meaning of the word. Examples of these from the replication are illustrated in the dictionary entries for *armed*, *unerring*, and *acrid*:

a), d), e): Someone who is **armed** is carrying a weapon, usually a gun.

b): If you describe someone's judgement or ability as **unerring**, you mean that they are correct and never mistaken.

c), f), g): An **acrid** smell or taste is strong and sharp, and usually unpleasant.

The room filled with the acrid smell of tobacco, The plant has an unpleasant odour and an acrid taste.

Other factors affecting L2 word recognition include the varying effort for different words involved in understanding the meaning and use of the word, and the number of facets of the words such as collocations, semantic and syntactic restrictions, or register, that are encountered through the dictionary entries. Many factors other than those listed above, although apparent in dictionary entries, relate more properly to the nature of the words themselves. We will consider these below.

3.6.3.3 The nature of the words themselves

Factors intrinsic to individual words that may affect comprehension and retention of a previously unknown word include its part of speech, the number of senses, the abstractness of the word or sense, its meaning as a "stand-alone" item, the length of the word, its spelling and the ease with which it can be pronounced. In Fischer's study, the set of twelve target items was made up of nouns, verbs, and adjectives, while in the replication only adjectives were used. Focusing on a mixture of words does facilitate the finding or adapting of a reading passage. On the other hand, with only three, four or five words for each part of speech, it does not seem a reasonable assumption to suggest that the twelve words are in any way representative of words unknown to the participants for a particular part of speech, especially given the wide range of other possible factors affecting comprehensibility and retention listed above. By using a set of twelve adjectives, the study is more likely to tell us something of value about the participants' learning of previously unknown adjectives.

The numbers or senses of a word, and the frequency of the word overall and of each sense, will be factors that affect the learning of words in a natural L2 learning environment, whether in a foreign or second language learning context. For Fischer's study and in the replication, this issue was avoided by either using single sense words or only presenting, in the dictionary entries and the text, the one sense of the words as used in the passage. Some research has suggested that the degree to which a word is concrete or abstract will affect retention (Hatch and Brown, 1995: 186-7). With such a small number of target items, this factor, too, may be a significant factor in the retention of unknown words. Related to this is the degree to which a word has meaning in isolation. This may be especially true of adjectives, so that the meanings of *rancid*, *acrid*, or *stagnant* may be hard to understand and recall without reference to butter, smoke, or water.

3.6.3.4 L2 words seen from the perspective of the participants' L1

The ability to recall or recognize a word will be affected by its length and the regularity of its spelling. This issue may be of particular significance to native speakers of languages with writing systems different from that for English, or for languages that are, compared to English, orthographically shallow languages (Nakamura, 2001). We will now go on to consider language-specific factors that may affect vocabulary comprehension and retention through dictionary use.

We have already noted that the regularity of an L2 word's spelling and its length will affect its recognition and retention to differing extents, depending on the L1 of the learner. Other factors concerning the learner's first language will also affect recognition and retention of unknown L2 words. These include "distracters" in the L1: cognates or loanwords with meanings other than those for the words in English, or that produce cross-associations with unrelated L1 words (Nation, 1990: 45-47), the presence or absence of phonemes that exist in the L1, and the existence or absence of fairly direct translation equivalents of the L2 words in the L1. The existence of distracters for particular words may mean that learners would have to unlearn the wrong links with L1 words prior to learning correct meanings and associations; this would involve more effort and time than for words without this prior mistaken knowledge.

Phonological considerations may also affect ease of retention for particular words, depending on the L1 of the learner. For words where a number of the sounds of the L2 words do not exist in the L1 (for example the *f* and *r* sounds in *furious* for Japanese speakers), it may be hard to retain a phonetic image of the word in the learner's mental lexicon to assist recognition of the word. The existence or absence of L1 translation equivalents will also affect retention. This may be especially evident with users of monolingual L2 dictionaries; a bilingual dictionary will usually provide a translation equivalent even if it is slightly inaccurate or hedged with semantic or other restrictions.

We have looked at some of the many ways in which previously unknown L2 words may vary in the ease with which they may be learned, recognised, or produced. Much L2 vocabulary acquisition research tends to treat target words as blank tokens, with the implication that what may be revealed about the learning of any given twelve or twenty words will be relevant for the vocabulary of that L2 in general. Especially where dictionary use is involved, we can see that this standpoint is far too simplistic, and that results may all too easily be affected by an unwitting bias resulting from the choice of the target words. This will be especially true where only a small numbers of target words are used, and where recorded gains in vocabulary knowledge are small, as in Fischer's study and in this study.

3.7 Conclusion

Having looked at some of the problems with the research conducted both in this study and in Fischer's original experiment, we will now go on to summarise the findings from this study. Here we will focus only on the results for comprehension and retention of target words, the main purpose of this study, making comparisons Fischer's data where relevant.

In the replication one clear finding is that access to monolingual dictionary entries made an enormous difference to overall comprehension rates for the previously unknown words; average comprehension rates for the two groups with access to MLD entries for the target words were more than twice that of the group which only had the text to refer to. In Fischer's study, comprehension rates were lower for all three groups, and differences between groups smaller, indicating how important the choice of target words can be for comprehension. Retention rates largely reflected comprehension rates; when this factor was taken into account, there was very little difference between groups in retention rates for target words which had been understood.

Three problems with the research were particularly evident: low retention rates, low levels of interrater reliability, and variability among target words and the material relating to the target words. Not all of these are necessarily problems in themselves; low retention rates may simply be a normal feature of incidental vocabulary learning involving single encounters with unknown L2 words. On the other hand, it may be that the testing instruments were insufficiently sensitive to record the small increments of vocabulary development that may be typical of vocabulary learning in this environment. The low levels of interrater reliability are definitely weaknesses in the study. Further research would need to be conducted to establish whether a more rigorous approach to this matter would produce greater agreement among raters and more reliable results overall. Finally, variability among the learnability and contexts of unknown words could arguably be said to be representative of unknown L2 words encountered in a natural learning environment. For this argument to be valid, especially for a small set of target words, it would be necessary to select the set of target words very carefully for each different type of word, dictionary entry, or lexical environment to be represented.

This brings us to ask how, or whether, the various problems encountered in this replication may be overcome. One way forward may be to try more sensitive instruments capable of measuring small increments of vocabulary acquisition. Another might be to have multiple encounters with target items so as to increase recordable levels of acquisition. Tests involving recognition of target words, rather than production and translation of sentences, may meet both of these requirements, as well as obviating the need for subjective assessment of participant responses. Such tests may reduce actual levels of vocabulary retention because of the much lighter cognitive demand of recognizing words than of writing and translating sentences containing target words. On the other hand, the use of an experiment with lighter tasks and less demanding tests should make it possible for many more items to be tested. This, in turn, should reduce the problem of variability among items since the more items there are, the greater should be the variety of items and environments tested, and the less impact unwittingly included atypical items would have on overall results.

It is with these considerations in mind that we shall proceed to the studies described in Chapters Four and Five.

Chapter Four: Dictionary Definitions and Contexts

4.1 Introduction

The replication of Fischer's study, described in Chapter Three, was informative in two main respects: in the findings regarding the relationship between dictionary use, comprehension and retention of previously unknown words, and in the problems it revealed regarding methods for investigating vocabulary acquisition through dictionary use. We will begin this chapter by briefly restating the problems encountered through the replication as it is largely from reflection on these that the experiment described in this chapter took shape. Following this, we will give an account of an experiment comparing comprehension, production and retention of previously unknown L2 words as a result of reading dictionary definitions or multiple example sentences. Finally, we will discuss the results of the experiment with reference to three studies which are similar in some respects to the one described here (Black, 1986; Miller and Gildea, 1987; and Nesi, 2000: 84-91). Neither Miller and Gildea's nor Nesi's studies are directly concerned with vocabulary retention, and so are not reviewed in Chapter Two, but as they are concerned with both comprehension of monolingual definitions and the production of sentences, they are very relevant to this study.

4.1.1 Problems with the replication

The replication of Fischer's study revealed four main problems with the means of investigating L2 vocabulary acquisition through dictionary use, problems that are common to other studies in this field.

Firstly, one problem is the small number of target words investigated. As we noted, this is especially problematic when we consider the wide range of types of words, types of dictionary entries, and typical lexical contexts for different words; it is hard to accept claims that results from a set of ten or twelve test items could be typical of each of these types and contexts. With such small numbers of items, the results of a whole study could be significantly affected by data from just a couple of non-typical items.

A second important issue is the selection or creation of suitable learning contexts for the unknown words to compare with learning through dictionary use. In the replication, the learning context employed was single encounters with the target words in a text, with most of these target words artificially inserted into the text. Problems with this may be summarised as follows:

a) One single lexical context may well not reflect typical use of a word;

b) The contexts of words in a text may not be sufficiently rich in clues to enable comprehension of their meaning (Mondria and De-Wit Boer, 1991);

c) Single encounters with the words in a single context may be unlikely to result in much learning.

A third issue is the methods by which target word comprehension and retention were evaluated in the replication. Participants were asked to write English sentences, then to translate the sentences into their mother tongue. Interrater agreement levels for the English sentences and their translations were low, averaging rates of around .7; raters disagreed in their rating of almost one third of participant responses. In this chapter we will investigate means of improving interrater reliability levels.

The final problem is with the measurement of retention of word knowledge. In the replication, pre- and post-tests were used for the measurement of retention; in these tests, the target words were presented in isolation and the participants were required to provide the meaning. The low levels of vocabulary retention recorded may be attributable, at least in part, to the lack of sensitivity of the tests employed.

4.1.2 Addressing the issues

In the experiment described in this chapter, we will set out to address each of the above issues. To increase the numbers of items tested, a less time-consuming learning task for the target words will be introduced. The first part of the task will remain unchanged: to produce an English sentence for each of the target words. For the second part, participants will be required only to give a Japanese translation equivalent for each target word, rather than the whole of each sentence, as in the replication. These changes should allow us to increase the number of items substantially.

The next problem is how to provide participants with sufficient natural and typical contexts for the unknown target words to occur. One way of concentrating the experience of encountering an unknown word in various typical contexts is by using a number of concordances or example sentences drawn from a corpus for each word. While there are drawbacks in using example sentences drawn from a corpus (we shall consider these in the Discussion section below), their use does enable learners to encounter less common words a number of times in a range of typical syntactic and lexical contexts; rarely would this be possible for more than one or two words in any individual text. If a study required language learners to experience multiple encounters with twenty or more low frequency words in unedited texts, we would need them to read many thousands of words of text.

A number of people have advocated the use of concordances or multiple contexts in language learning, both directly and by implication, over the past couple of decades. Craik and Lockhart (1972) have proposed that learners' word retention will increase when more associations or links are available for the learner; concordances may provide this richer learning environment. While also not referring directly to the use of concordances, Sharwood Smith (1994: pp.179-181) recognises the importance of learner exposure to multiple contexts in language learning and rule-building. Earlier, Johns (1991), describing what he terms "data-driven learning", proposed the use of concordances in precisely this role: as a means for learners to learn by identifying typicality and developing rules from the examples of use provided in concordances. This, too, is the approach employed by van Daalen-Kapteijns et al. (2001) with their focus specifically on deriving meaning from multiple contexts.

One objection to using concordances rather than one individual text is that any idea of investigating incidental vocabulary acquisition through reading is lost. If incidental vocabulary acquisition were the object of this study, then this objection might be justified. However, as for both Fischer's study and the replication, this experiment is concerned with the intentional, conscious use of previously unknown English vocabulary by students of English (see Hulstijn (1992) for a discussion of this issue). There is, in fact, one sense in which the experiment may be said to be testing incidental vocabulary acquisition. The object of the task is for the participants to understand and demonstrate understanding of the target words; retention is not set out as a goal and although comprehension is a prerequisite for retention of meaning, it cannot be assumed that the participants would have had any intention to retain any knowledge gained about the meaning of the target words beyond the task itself. As Wingate (2002: 15) points out, the condition upon which judgments of incidental learning should depend is the presence or lack of unintentionality on the part of the learners. Especially for a study of language learners in a classroom environment, in which learners' intentions may not be known and may be varied, it would be unwise to claim categorically, as many studies continue to do, that the learning that takes place is either incidental or intentional.

As for retention, whether incidental or intentional, the researcher's experience in the replication, and also suggested by the lack of retention data in Fischer's study, was that only low levels of word knowledge retention were recorded. As this may be partly due to the use of an insufficiently sensitive instrument, we will test the use of more sensitive evaluation methods to measure retention of word knowledge. In addition, we will aim to record participants' recognition of word contexts rather than production of word meanings without reference to contexts. Remaining theoretical or practical issues will be addressed through the experiment or in the Discussion section of this chapter.

4.2 The study

This experiment involved taking a group of Japanese learners of English, randomly dividing the group into two, and giving one group a set of monolingual English dictionary entries for each of a set of unknown English target words and the other group a set of corpus-drawn example sentences for each target word. The participants' task, using the resources provided, was to make their own sentences for each target word, then to write a Japanese translation equivalent for each target word. A sub-group of participants were also available, five weeks after this learning session, to take a type of word knowledge retention test of the target words. In this test, the participants were asked to match the target words with the same contexts as those in which they had seen these words in the learning session.

4.2.1 Hypotheses

There are three basic hypotheses underpinning this research:

1) Participants receiving dictionary definitions will be more likely to understand the meanings of the target words and thus better able to give accurate Japanese equivalents for the words.
2) Participants receiving concordances will be better informed as to the target words' syntactic and lexical environments and so produce a greater proportion of acceptable English sentences using the target words.
3) Participants receiving dictionary definitions will gain clearer, more focused, knowledge of the target words and this will be demonstrated in better retention of the target words.

The reason for the first hypothesis is that the purpose of monolingual learner dictionary definitions is to convey the meaning of unknown L2 words to language learners, while, as Schatz and Baldwin (1986) have noted, contexts are not typically very informative as to word meaning. Related to this is the reason for the third hypothesis; although the effort involved in seeking to understand the word meanings from definitions and contexts may be similar, the higher rate of comprehension for participants with definitions would result in higher levels of word retention.

4.2.2 Participants

Fifty-four second year English-major Japanese university students were recruited as participants for this experiment. All of these students were native speakers of Japanese, aged between 19 and 21. All were intermediate learners who had received about seven years of formal instruction in English. Twenty-five of the participants received dictionary entries from two dictionaries for 24 words, and twenty-nine of the participants received sets of three sentences for each of the words drawn from a 50-million-word general corpus of English (COBUILD Direct). Eleven participants per learning condition were also available, five weeks after this learning session, to take a word knowledge retention test.

4.2.3 Target words

There were, eventually, 24 target English words, all verbs. Words from a single word class were focused on to avoid the lack of focus that would result from using a varied set of target words. The words were selected by a process of exclusion according to various criteria:

1. More common verbs were excluded (those included in the most frequent 3,400 words according to the *Collins COBUILD English Dictionary, 2nd Edition* (COB2). It was also confirmed that none of the words were

contained in the 8,000-word JACET 8000 list (see Mochizuki, 2003, for an explanation), as these, too might have been known to the participants.

2. Words that were judged to be too infrequent or obscure were excluded; if the words were too uncommon, subjects might not be able to recall (or might not even know) Japanese equivalents. Words for which there were fewer than 25 occurrences in the COBUILD Direct corpus (i.e. less than 1 occurrence per 2,000,000 words) were excluded; not only were they rare, there might also be too few occurrences in the corpus to provide typical usable examples as part of the stimulus material.
3. Polysemous verbs, as identified in COB2, were excluded, with the exception of verbs with a noun sharing the same basic meaning (e.g. *cackle*, *sulk*). This was to facilitate selection of typical sentences from the corpus and to avoid confusion for raters.
4. After 1, 2, 3, and 4, a list of 108 remaining verbs were presented to a group of 17 students similar to the participants. They were asked to identify and give the meaning of any of the verbs on the list which were familiar to them. Verbs correctly identified by more than one student were excluded from the experiment.
5. A list of the 52 remaining verbs was presented to the participants. Verbs correctly identified by any participants were excluded, as were words that were commonly mistaken for other words.
6. Words left from the list after 4 and 5 that did not have fairly direct semantic equivalents in Japanese were excluded.

Twenty-eight words were selected from the 33 words left on the list after 4, 5, and 6. After a pilot test revealed that 28 items would make the experiment too long for the time available, four more of the words on the list were excluded. The final 24 target words were:

> *abbreviate, amputate, appal, blab, bode, cackle, cadge, canoodle, chomp, coerce, dilate, elope, feign, flunk, jilt, perspire, pooh-pooh, raze, sulk, suss, trounce, waft, whinge, wince*

4.2.4 Learning materials

The study involved two different sources of information about the target words. For one group, the Dictionary group, each participant received two monolingual dictionary entries for each target word, stripped of example sentences (Appendix 4.1). The other group, the Example Sentences group, received three example sentences for each word, drawn from the COBUILD Direct corpus (Appendix 4.2). More information is provided about these materials below.

Dictionary group

This group received two dictionary definitions for each of the target words, taken from two popular English monolingual learner dictionaries: the *Longman Dictionary of Contemporary English 3rd Edition* (Summers et al., 1995) (LDOCE3) and *Collins COBUILD English Dictionary 2nd Edition* (Sinclair et al., 1995) (COB2). The order in which the definitions appeared on this group's materials alternated from word to word between these two dictionaries. The dictionary entries for the words were stripped of any example sentences or phrases but included the definitions and any grammatical information or information about synonyms or antonyms. There were two main reasons for including definitions from two dictionaries for each word. First, it was hoped that providing two definitions would reduce the variability in accuracy and clarity from word to word that occurs in dictionaries. Secondly, a learning condition comparable in time and volume to that for the Example Sentences group was required. A single definition from one dictionary would have been too short, but two definitions were very close in terms of length and reading time to those of the example sentences made available to the other group.

The reason for excluding example sentences from the dictionary entries was to provide a less complex source of information for the target words to compare with the multiple contexts of the example sentences described below. This approach was also employed by Black (1986). A further validation of this decision is that some recent electronic dictionaries do not display example sentences for entries unless specifically requested, with the result that many users content themselves with the definitions alone, as in this study.

Example Sentences group

This group received three example sentences for each of the target words, drawn from the 50-million-word COBUILD Direct corpus. The frequency of the target words in the corpus ranged between 1 per 250,000 words and 1 per 2,000,000 words. This meant that the example sentences for each target word were selected from between 25 and 200 concordance lines. Sentences were chosen on the basis of their displaying typical syntactic patterns and collocations, as judged against the instances of occurrence provided by the corpus. For some words, only one pattern occurred very frequently or only one or two words collocated highly with the target word; in these cases, two or three of the selected example sentences for the word would display the same pattern or collocate. For example, things typically only *bode* *ill* or *well*, and people are typically seen *canoodling* *with* somebody else, and this is reflected in the example sentences chosen. These are shown in Figure 4.1.

Figure 4. 1
Sample of data used for target words for the Example Sentences group

> The perilous state of their economy does not *bode* well for their future.
> These developments *boded* ill for the religious peace within the armed forces.
> Coupled with last year's hot summer, this *bodes* well for British tourism.
>
> Her former boyfriend was seen *canoodling* with 34-year-old Paula Yates.
> In the audience Paula Abdul *canoodled* with actor Emilio Estevez.
> She was shocked when she saw him *canoodling* with Robin Givens.

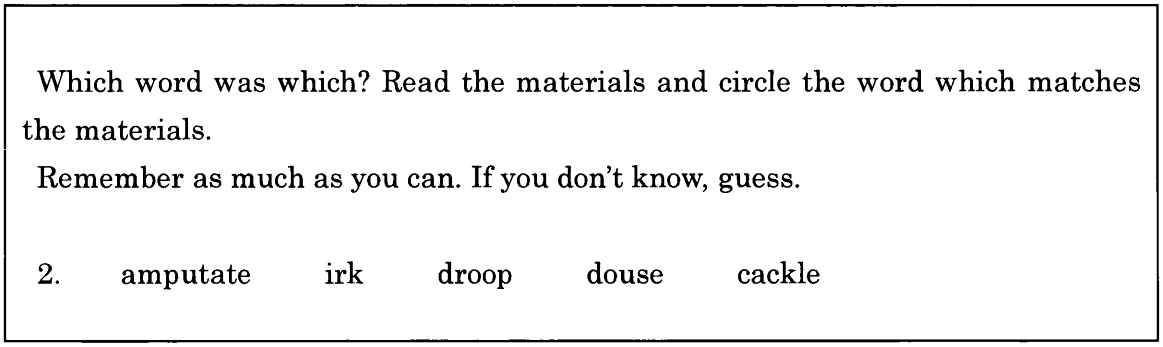
4.2.5 Procedure

The participants were randomly divided into two groups. Both groups were given an answer sheet with 24 presumably unknown English verbs written on it (Appendix 4.3). With the answer sheet, the Dictionary group was also given a set of definitions for each of the target words (Appendix 4.1) while the Example Sentences group was given a set of three example sentences for each target word (Appendix 4.2). The participants were instructed, in Japanese, to study the materials provided and to write their own sentences using the information provided. As they finished this task, they were, individually, instructed to write Japanese equivalents for each of the English words. On average, the sentence writing took the participants about 50 minutes, with the writing of the translation equivalents taking a further 20 minutes.

4.2.6 Retention test

A retention test was conducted five weeks after the main vocabulary learning session to measure retention of the target words. As it appeared that low post-test scores in previous studies could, in part, be attributed to the insensitivity of tests that required the production of the target word without reference to the context, a more sensitive test was devised. A total of 22 participants from the original group (11 from each experimental group) made themselves available for this retention test. They were given the same materials as they had had during the learning session except that the target words were deleted from the definitions or example sentences (samples of both are provided in Appendix 4.4). The participants were given answer sheets (Appendix 4.5) on which a choice of five words was given and were required to match the correct word with the set of gapped definitions or example sentences for that word, as shown in Figure 4.2.

Figure 4.2
Instructions and extract from the retention test for Example Sentences group.



Which word was which? Read the materials and circle the word which matches the materials.
Remember as much as you can. If you don't know, guess.

2. amputate irk droop douse cackle

It was, then, a type of multiple-context, multiple-choice cloze test. It was also a forced choice test; testees were required to select one answer for each item, whether or not they knew the correct answer. The five words for each test item were composed of one correct answer, one other target word, and three other words that the participants had seen previously in a pilot test when asked to identify which words they knew. The gapped materials were, apart from the gaps, identical to the materials seen five weeks previously.

4.2.7 Rating of responses

A major problem with the study described in Chapter Three was the low level of interrater reliability for the evaluation of the English sentences and their translations. Two measures were employed to improve this: for the sentences, more stringent, specific rater guidelines with examples were provided. For demonstration of comprehension, we only required translation equivalents for the target words as opposed to translations of the English sentences.

The revised rater guidelines (Appendix 4.6) for the English sentences gave clear instructions, together with illustrative examples taken from the participants' responses, for the three categories of Correct, Questionable, and Unacceptable. For the Correct category, the following issues were addressed: incorrect subject-verb agreement; spelling errors; errors in another part of the sentence; acceptability of verb subjects and objects; transitivity; and evidence of mistaken meaning.

Despite the detailed revised guidelines for the English sentences, there was little effect on interrater reliability, which improved only marginally from .68 in the study reported in Chapter Three to .69 in this study. Using a changed format for the translation equivalents was, however, successful in significantly improving rater agreement: from .75 in the replication to .85 in this experiment.

4.3 Results

There were three sets of results from this experiment: for the English sentences written by the participants, for their translation equivalents of the target words, and for the retention test scores for the smaller group. We will report these individually and also compare ratings for the English sentences and translation equivalents.

4.3.1 English sentences

For the English sentences produced by the participants, raters' categories of omissions and questionable or unacceptable English uses were collapsed into one category – "unacceptable usage". Other uses were "acceptable usage". The results for the two groups were very similar, with both groups producing acceptable sentences for between 55% and 60% of the items, as shown in Table 4.3.

Table 4:3
Acceptable English sentences for target words (maximum = 24)

	M	S.D.
Dictionary group (N = 25)	13.8	4.73
Example Sentences group (N = 29)	13.38	3.86

A t-test performed on the English sentences ratings for the two groups confirmed that there was no significant difference between the groups [t=0.35, p>.1].

4.3.2 Translation equivalents

Scores for matching and partially-matching translations equivalents were then collapsed into one score as "adequate translations", with omissions and non-matching translation equivalents counted as "inadequate translations". Using these collapsed scores, a t-test was conducted on the ratings of the Japanese translation equivalents for the participants in the two groups. The result of the t-test (t = 4.67, p<.001) confirms that there is a significant difference between the translation equivalents results for the two groups.

Table 4.4
Adequate translation equivalents for target words (maximum = 24)

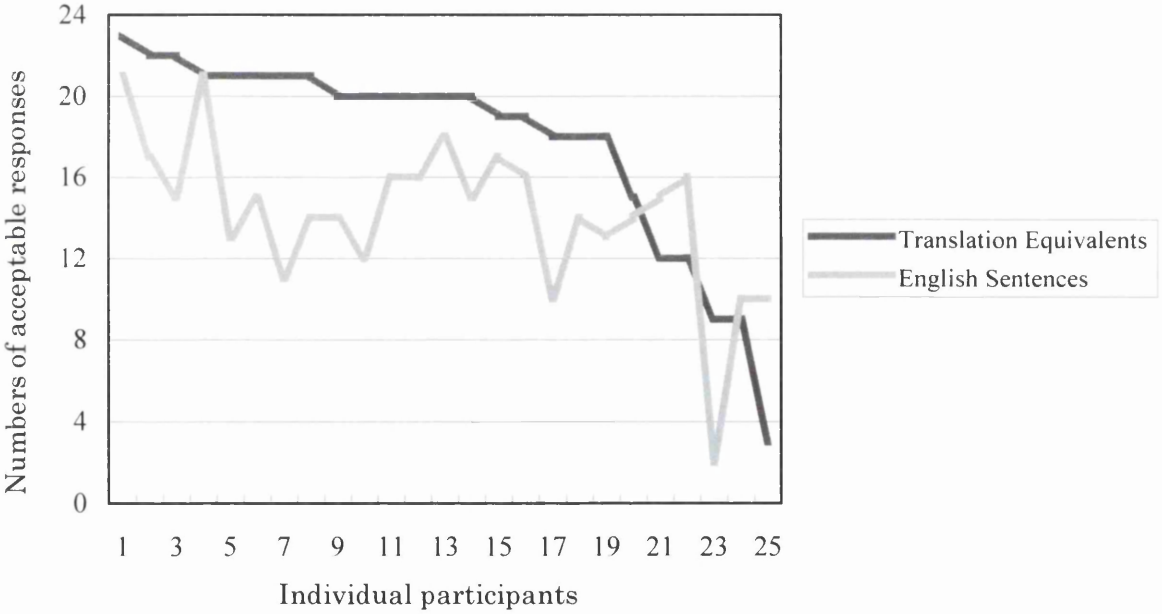
	M	S.D.
Dictionary group (N = 25)	16.20	8.94
Example Sentences group (N = 29)	7.79	1.21

Apart from the difference in mean scores, the large differences between standard deviations for the two groups is also worthy of comment. The Dictionary group's high level of standard deviation indicates a wide range in scores between the individual participants, suggesting that some of the participants were able to understand almost all of the dictionary definitions while for others definitions for a large proportion of target words were too difficult to understand. For the Example Sentences group, the situation is quite different. On average, the participants were able to correctly guess the meaning of only one third of the target words from the contexts in which they appeared. The number of correctly guessed words did not vary widely from participant to participant. Worthy of note, however, was that there was a very wide range in numbers of correct responses from target word to target word. This is confirmed when we look at the data for individual words. For almost a third of the target words no more than four of the twenty-nine participants correctly guessed the meaning, while for three of the target words twenty or more participants gave the correct answer. We will consider the implications of this data in the Discussion section below.

4.3.3 Comparison of results for English sentences and translation equivalents

Before going on to retention test scores, it is worth noting the differences between the two groups regarding the relationship between scores for English sentences and for translation equivalents. These are shown in Figures 4.5 and 4.6 below.

Figure 4.5
Translation equivalents and English sentences for Definitions group

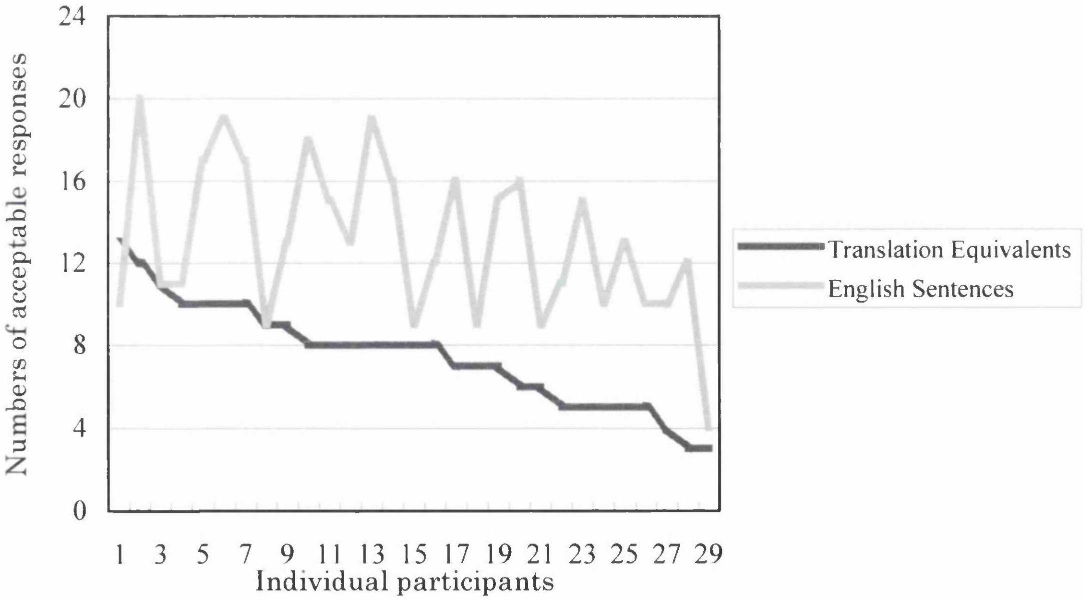


For the Definitions group, almost all participants produced more acceptable translation equivalents than good English sentences. This means that they were more able to understand the meanings of the target words than to use them correctly.

For the Example Sentences group, the opposite was true: for almost all participants the number of acceptable example sentences exceeded the number of acceptable translation equivalents. Most participants in this group, then, were able to use the target words correctly in sentences more often than they were to accurately guess the meaning of the target words. In other words, Example Sentences group participants often produced acceptable sentences for target words which they were unable to understand correctly.

Figure 4.6

Translation equivalents and English sentences for Example Sentences group.



4.3.4 Retention test

The results of the word knowledge retention test for the smaller groups (11 in each group) are shown in Table 4.7 and Figure 4.8. The raw figures and percentages shown here do require some interpretation, as the test was a forced choice multiple choice test, with the correct answer and four distracters to choose from for each of the 24 test items. This means that blind guessing of the answers would result in a score averaging 4.8 points, or 20% of the total. A more likely situation, but still simplified, would be one in which a participant recognizes 40% of the target words, and guesses the answers for the remaining 60% of items. Assuming a one in five success rate for the guessed items, this would result in a total score of 52%.

Table 4:7
Word knowledge retention test results (maximum = 24)

	M	S.D.
Dictionary group (N = 11)	11.9	3.5
Example Sentences group (N = 11)	9.1	3.0

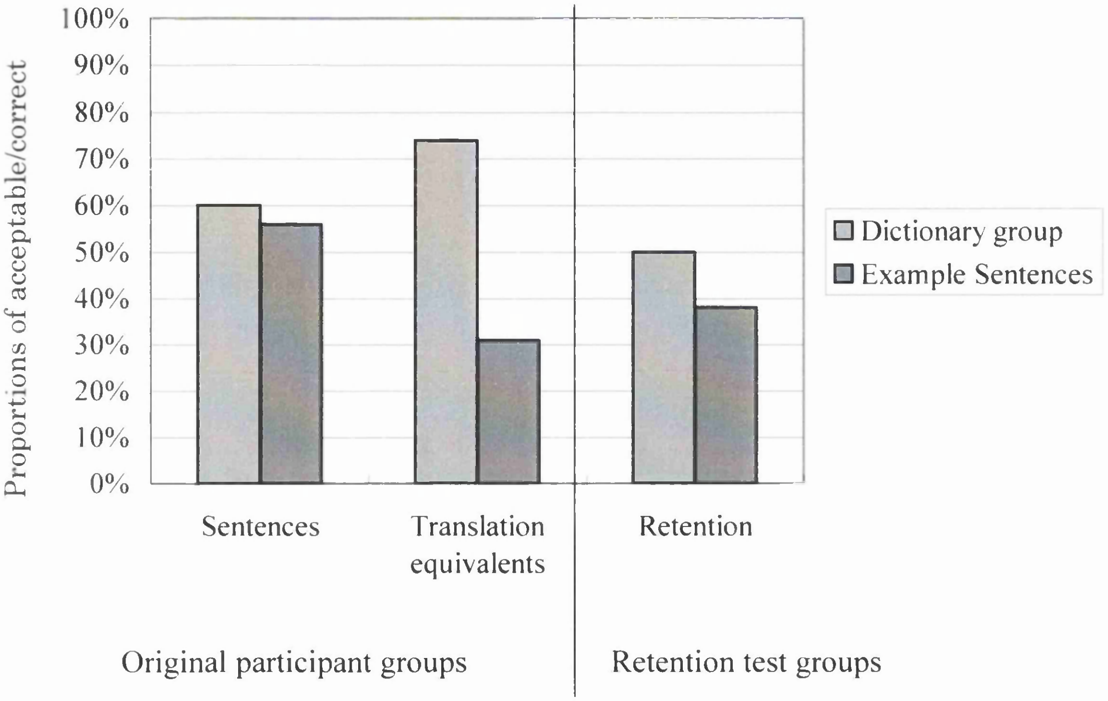
The raw and revised figures and graphs show a marked difference between the results for the two groups, with the Dictionary group performing on average 12 percentage points better than the Example Sentences group. A one-tailed t-test conducted on the raw scores ($t = 2.01$, $p<.05$) indicated that the difference between the results for the two groups is significant, despite the relatively small numbers of participants involved in this part of the experiment. We will discuss these results in greater detail below.

4.4 Summary of results

The main findings of this study are summarised below and presented in Figure 4.8.

Figure 4.8

Acceptable English sentences, translation equivalents, retention scores



In brief, the findings were that:

1. Both the dictionary definitions and the example sentences resulted in the participants' producing a large proportion of acceptable sentences for previously unknown words (almost 60%).
2. The Dictionary group participants produced, on average, more than twice as many acceptable translation equivalents as the Example Sentences group: an average of over 70% acceptable equivalents for the Dictionary group as compared with just over 30% for the Example Sentences group.

3. Retention of lexical knowledge, after five weeks, was relatively impressive for both groups: for the Dictionary group, almost 50% and, for the Example Sentences group, just over 35%. Even with allowance made for random guessing in a multiple choice test, the figures are still high when compared with those in similar studies.

4.5 Discussion

We now consider the findings summarised above with reference to the hypotheses proposed earlier. In doing so, we will also consider these findings in the light of those from other studies which address similar issues to those addressed here.

4.5.1 English sentences

We will begin by looking at two aspects of the results regarding the English sentences:

a) the proportions of acceptable English sentences produced by the two groups;

b) the similar numbers of acceptable English sentences produced by participants in the two groups.

4.5.1.1 Proportions of acceptable sentences produced

Participants in both groups produced on average well over 50% of acceptable sentences for the target words, which were all previously unknown low frequency English verbs. Whether this figure is high or low may only be judged by comparison with similar studies. In Miller and Gildea's (1987) study, figures were similar to this with the exception of their definition only group, for which only 36% of the sentences were judged acceptable. In their study, however, participants tended to copy from dictionary entries or example sentences. In the study reported here, however, this does not appear to have been the case; participants were asked to produce their own sentences using the target words (literally "free composition") and it is evident that they often did this, seeing it as an exercise in communication rather than just English practice. It might be argued that the participants simply modelled their sentences on those in the materials they were given. This is self-evidently true, and was in some respects the aim of the exercise; the materials were meant to provide information about word meaning and typical use of the words. No sentences, however, were copied word for word, and in only a few instances was there obvious literal copying of parts of sentences, such as noun phrases.

A combination of at least three reasons may account for the difference in results for sentences between this study and Miller and Gildea's (1987) study. First, the participants in Miller and Gildea's study were younger school children, and the vocabulary and dictionary use under investigation were from the children's mother tongue. The age difference between the researchers and participants in their study, and their lack of teacher-student relations, may explain why the children may not have thought of making the activity communicative in any sense. Secondly, the fact that the activity in this study was described as "free composition" in a foreign language may, too, have masked the research aspect of the activity and so encouraged originality in a way that Miller and Gildea's failed to do. Thirdly, in this study, dictionary entries were written with language learners in mind, deliberately avoiding difficult vocabulary or grammatical structures. For the Example Sentences group, three sentences were presented for each word, with the sentences selected on the basis of typicality and clarity. This is in contrast with single example sentences, either made up or from the *New York Times* in Miller and Gildea's study. With a single sentence or definition, the children may have been unsure how free they were to diverge from the model provided. In contrast, more than one example sentence or definition may provide learners with a greater sense of freedom to write within the parameters of the words' meanings and use.

A related issue, raised by Miller and Gildea (1987) and also by Nesi and Meara (1994) in relation to L2 learners, is concerned with whether participants were able to understand the words for which they produced sentences. We will consider this below.

4.5.1.2 Comparison of groups' production of sentences

The second hypothesis proposed with respect to this experiment was that Example Sentences group participants would produce accurate sentences for a greater proportion of the target words than participants of the Dictionary group. In fact, the two groups produced approximately the same number of acceptable sentences. It may appear a simple matter to explain this: the amount of useful linguistic input was equal for the two groups and this resulted in similar numbers of acceptable sentences. This similarity in numbers of acceptable sentences does, however, mask fundamental differences between the linguistic data to which they were given access.

In the case of the Dictionary group, the L2 definitions provided usually did enable the participants to understand the meanings of the target words; on the basis of this understanding they created their sentences. Where participants were unaware of the specific syntactic behaviour of a given word, this may have been an impediment to producing an acceptable sentence for some words but not for others. In any case, for one of the dictionaries used (COBUILD), the definitions are written so that the target words are used in the definition in a syntactically typical context. This does appear to have a significant effect on the production of the sentences. As stated above, for the Dictionary group, two definitions were provided for each target word, with the order of the definition alternating between that from LDOCE3 and COB2.

For the 12 words where the COB2 definition appears first, an average of 16.6 of the 25 participants provided acceptable sentences, as compared with an average of 13 participants for the 12 words for which the LDOCE3 definition appears first. This confirms the dictionary editors' purpose for such definitions, to "hold up models that would be of assistance to learners in encoding English" (Hanks, 1991: 116-136), as opposed to only decoding L2 text. The advantage gained by these encoding-friendly definitions helps explain why, overall, the Dictionary group participants managed to produce a proportion of acceptable sentences very similar to that produced by the Example Sentence group participants. It would also explain why it differs in this respect from the results in Miller and Gildea's (1987) study, in which the definition only group produced a smaller number of acceptable sentences than either of the example sentence groups in their study. Nesi's study (2000: 84-91) in which she specifically compared the effect of different definition types on sentence production by language learners, did not reveal any significant effect in this respect, but this may because her the set of target words used in her study were a combination of verbs, nouns, and adjectives. We might expect to definition types to assist production differently for different parts of speech; this issue will be addressed further in the following chapter, which focuses on adjectives.

For the Example Sentences group, as Figure 4.6 clearly shows, although the proportion of acceptable sentences approached 60%, demonstration of correct understanding of the word was much rarer: only one word in three, on average. This means that although for many words the meaning was incorrectly interpreted, many of the participants were still able to produce acceptable sentences. They would do this by using the sentences provided as syntactic models and noting the nature of the subjects or objects of the verbs. This discrepancy between scores for sentence-writing and comprehension suggests that the LUCAS (Miller and Gildea, 1987) methodology is basically flawed. Just as a parrot can imitate sounds, so can a dictionary user very often write sentences, using target words without being able to understand their meaning. There is an implication, made explicit by Wesche and Paribakht (1996) in their Vocabulary Knowledge Scale, that sentence writing is a demonstration of comprehension, or that it shows more than just comprehension. Where examples of the target word in context are provided, this is very often not true.

4.5.2 Translation equivalents

There was a wide variation in proportions of acceptable translation equivalents for the two groups. Put simply, the Dictionary group were able to identify or guess the correct meaning of the target words for over two out of three words, while for the Example Sentences group under one in three words were identified correctly. Even when three typical, relatively clear examples of use of unknown words were presented to the learners, the chances of their understanding the word were much lower than for learners who read the dictionary definitions. Or, rather, for the Example Sentences group, the chances of misinterpreting the information were much higher. This contrasts with the general beliefs regarding the wordlearnings, whether by children or for older L2 learners, expressed by Miller and Gildea (1987) as follows: "What they need is not definitions, but lots of examples of how the word is used". The high levels of misinterpretation of words from the sentences may be partly because the three sentences used in this study may not be the "lots of examples" that Miller and Gildea had in mind.

The results of this study also conflict with the data, admittedly limited, from Black's (1986) study of advanced L2 learners in which those with access to example sentences were better able to understand words in a text than learners with access to L2 definitions. Four factors may account for the difference in results for the Example Sentences group in the two studies. Firstly, since in Black's study, the participants are advanced level English learners, more of the information in the example sentences would be accessible to them than in this study. Secondly, there were four or five example sentences for each word in Black's study, and only three per word in this study. Thirdly, it is possible that verb meanings are less easy to understand from context than adjectives, for example. This does need further investigation, and will be considered in more detail in Chapters Five and Six. Finally, Black's example sentences were made up, and were pregnant in clues as to meaning whereas the corpus-drawn example in this study demonstrated typical use but often provided little accessible information regarding word meaning. Sets of example sentences from the two studies, shown in Figure 4.9, illustrate this.

Figure 4.9
Example sentences from Black's (1986) study and the present study

"I don't like these new-fangled washing machines" my grandmother always says.
"It's just as easy to wash the clothes by hand, and the clothes last longer"
We need better teachers, not new-fangled ideas of education!
To old Ned, even cars were new-fangled contraptions.
(Black, 1986)

I can probably cadge a lift later.
On the rare occasions that Liam was in the house when we were working, he would cadge cigarettes off my labourers.
None of them will walk anywhere if they can cadge a ride.
(This study)

If we look at the material for a couple of words for which the number of correct equivalents is low, we may gain a better understanding of what may have lead the learners to the wrong conclusion. We will start with data for the word *cadge*, for which the example sentences are shown above and the definitions are shown in Figure 4.10.

Figure 4.10.
Definitions provided to Dictionary group for *cadge*

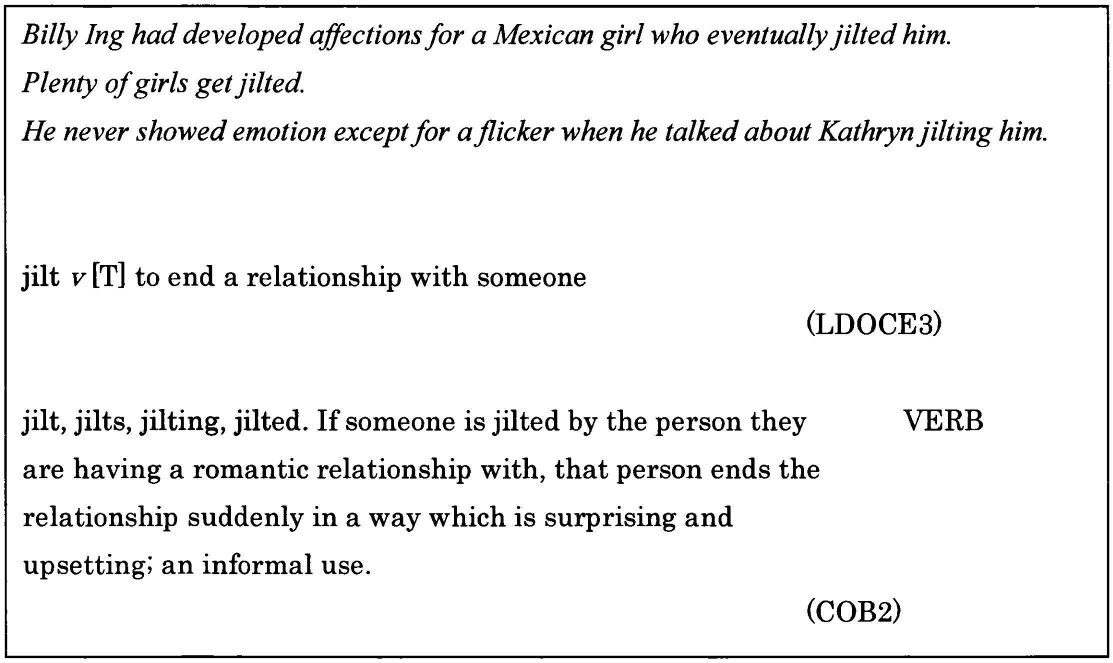
cadge *v* [I,T] *BrE informal* to ask someone for food or cigarettes because you do not have any or do not want to pay; MOOCH AmE: cadge sth from/off
(LDOCE3)

cadge, cadges, cadging cadged. If someone cadges food, money, or help from you they ask you for it and succeed in getting it; used mainly in informal British English. VERB
(COB2)

For this target word, only two Example Sentence group participants were able to give even partially correct translation equivalents. The Dictionary group did better, with 8 matching equivalents and 11 partially correct equivalents. The poor scores for the Example sentences group may be due to the words *lift* and *ride* both being used in senses that would have been unfamiliar to the participants; words that are only known in senses other than those in the definitions may be positively unhelpful in providing information as to word meaning. For the word *jilt*, the situation is slightly different. The example sentences and definitions are shown in Figure 4.11 below. For this word, only two of the Example Sentences participants managed partially correct equivalents, while for the Dictionary group three participants gave correct equivalents and 20 gave partially correct equivalents. In this case, the example sentences do not contain much difficult vocabulary; rather they provide very little information at all. The majority of answers from this group were concerned with having, rather than ending, a relationship.

As the Dictionary group's scores indicate, there are problems with the definitions too; the first is too general while the second is long and complicated. It is not difficult to imagine how many of the Dictionary group participants only derived a vague, partial meaning for *jilt* such as to part or separate; they may have read the first (LDOCE3) definition, looked at the second (COB2) for more precision as to the word's meaning, then given up. This would not be so much of a problem for skilled dictionary users, especially those with larger vocabularies. They would be able to scan the entries quickly, and draw information from both definitions or from whichever definition is more informative.

Figure 4.11.
Example sentences and definitions provided for *jilt*



Billy Ing had developed affections for a Mexican girl who eventually jilted him.
Plenty of girls get jilted.
He never showed emotion except for a flicker when he talked about Kathryn jilting him.

jilt *v* [T] to end a relationship with someone

(LDOCE3)

jilt, jilts, jilting, jilted. If someone is jilted by the person they are having a romantic relationship with, that person ends the relationship suddenly in a way which is surprising and upsetting; an informal use. VERB

(COB2)

4.5.3 Retention test scores

With few exceptions (Iwai, 2000; Laufer and Hill, 2000), studies of L2 vocabulary acquisition through dictionary use conclude that little measurable acquisition of word meaning was recorded, even with immediate post-tests. In Hulstijn, Hollander and Greidanus' (1996) study, retention of various types of word knowledge were tested for, including production of word meanings without the aid of context, production of word meanings with the aid of context, and word recognition; not surprisingly, the highest figures were obtained for recognition of previously unknown words. In this present study, it seemed reasonable to expect that ability to recognise previously unknown words might be a type of word knowledge that would be retained through the single, but intensive, encounter with the target words of the participants in the study described in this chapter: writing sentences using the target words and then trying to provide L1 translation equivalents for these words. The results of the retention test confirm this.

Both the Dictionary group and the Example Sentences group participants in this study recognised a sizeable proportion of the target words, although, as explained above, these results do have to be treated with caution. The Dictionary group gained an average raw score of 12 out of 24 items. If we assume that for unknown items, one in five would be guessed correctly, the 12 items answered correctly can be broken down as 9 items known and 3 items out of the remaining 15 guessed correctly. If we apply the same reasoning to the Example Sentences results, the raw score of just over 9 items can be broken down as follows: 5 items known and 4 out of the remaining 20 items guessed correctly. These figures compare favourably with those obtained in similar studies; it seems that by asking less of the participants in terms of word knowledge, we were able to get a more meaningful and satisfying result.

A critical reaction to this type of test could be to ask what value there may be in simple recognition of the context of a word without knowledge of the word's meaning. There is some justification to this, but there is even more to be said in defence of testing for this type of knowledge. Since vocabulary acquisition, especially incidental acquisition, is often perceived as incremental, we need to think how evidence of initial increments may be observed. Short-term retention of word knowledge might be understood to be an indication of early incremental knowledge, and this has been tested for widely in vocabulary acquisition studies. The matching of target words with their superordinates, too, might be used to reflect learners' partially formed word knowledge. As for recognition of word forms, Hulstijn, Hollander and Greidanus (1996) describe this as "incidental vocabulary learning in its most modest form". Recognition of word contexts is similar to this, and may be said to share its lowly status. Yet, very often, without such humble indicators of vocabulary learning, it would be impossible to identify any learning at all. This applies especially, as in this study, when the retention test is conducted a number of weeks after the participants' single contact with the target words.

4.6 Conclusion

The study reported in this chapter had three main purposes related to language learners' use of monolingual learner dictionary definitions with typical corpus-drawn examples: to investigate which source of information was more effective in enabling learners to produce English sentences for previously unknown words, to investigate which was most likely to provide learners with an accurate understanding of the previously unknown words, and to investigate which resulted in the highest levels of word knowledge retention. In this conclusion we will begin by pointing out areas in which this study was successful. We will then indicate various problems with the study. Finally, we will consider how to proceed with further research in this area.

4.6.1 Achievements of the study

This study provided useful data in a number of areas. First, in terms of testing methods, the study achieved its goal of gaining a substantial increase in the inter-rater reliability rate for the evaluation of the translation equivalents: from .75 in the replication reported in Chapter Two to .85 in this study. Secondly, the study indicated significant differences between the two groups in terms of comprehension of the target words; the Dictionary group produced, on average, more than twice as many correct translation equivalents as the Example sentences group. Another valuable finding is that, despite this difference in comprehension rates, there was little difference between the two groups in terms of the production of acceptable English sentences for the target words. Related to this, the relative superiority of COBUILD-style definitions was also apparent for the production of sentences for the target words. Finally, the use of an instrument requiring recognition of lexical contexts for the measurement of retention of word knowledge does appear to have been successful in identifying low levels of retention, but with some reservations which are discussed below.

4.6.2 Problems with the study

There were three main problems with this study, all concerned with the testing instruments employed. The first is that the interrater reliability rate for the English sentences produced by the participants in this study remained low, despite measures taken to improve this figure. An agreement rate of .69 means that raters' evaluations of the participants' sentences differed almost one time in three. This adds further evidence to the doubts expressed about the use of sentence writing as a measure of comprehension or meaningful production.

The other two problems concern the retention test. One was the lack of a significant difference between the retention test results for the two groups. Although there appears to have been a sizeable difference between the two groups' test results, the widely ranging scores among the small number of participants for this part of the study meant that the difference between them was not statistically significant. The other problem with the measurement of vocabulary retention was the use of a multiple choice format for the Retention test. With a choice from five answers for each item, interpretation of the results is far from straightforward, as can be seen in the discussion of the scores above. With no knowledge at all, random guessing would still result in an average score of 20%. What, then, does the Dictionary group's score of close to 50% mean? How much of this score may be accounted for by correctly guessed items? In addition, how do we know whether or not items were half guessed? The use of a multiple choice format with this evaluation method leave us with as many questions as answers and needs to be reviewed.

4.6.3 The way forward

One important, but unexpected, outcome of the study is the effect of dictionary definition types on sentence production and how this might differ according to the part of speech under investigation. This suggests that it would be valuable to investigate sentence production, comprehension, and retention of a set of unknown words other than verbs, such as adjectives or nouns. Some problems with the study described above, however, suggest that changes should be made to the evaluation methods employed in this study. Specifically, there are problems with two of the instruments used for evaluation in this study: the participants' production of sentences for the target words and the use of a multiple-choice format test for the retention test. In Chapter Five we will consider what viable methods may be employed as alternatives to these instruments.

Chapter Five: Focusing on Adjectives

5.1 Introduction

In this chapter we will investigate the learning of English adjectives through monolingual learner dictionary use by Japanese learners of English. As in Chapter Four, we will compare the effect of learning from monolingual learner dictionary definitions with learning from authentic example sentences drawn from a corpus. The main reason for focusing on adjectives in this chapter is to investigate whether sentence production, comprehension, and retention of unknown words from contact with these materials is observably different for different parts of speech. The focus on adjectives in this chapter is also relevant to various other aspects of the research. These include how dictionary definitions may differ for different parts of speech in terms of defining style, definition length, and comprehensibility to L2 users. We will also consider how informative as to meaning natural example sentences may be for different parts of speech. Regarding the evaluation of participants' responses, we will look at how accuracy or levels of agreement in rating participants' sentences or translation equivalents may depend on the part of speech of the target words.

Three main factors have led to this study on adjectives. They were that:

a) In much research into dictionary use and L2 vocabulary acquisition (see, for example, Black, 1986, Luppescu and Day, 1993, and Laufer and Hill, 2000), a "balanced" set of target words is used, composed of nouns, verbs, adjectives and, in some cases, prepositions or adverbs. One implication of this is that if all the target words were verbs or nouns, for example, the sample would be unbalanced and the results would not reflect how words are typically used. This may imply the belief that learning and retention are different for different parts of speech. It may also reflect researchers' fears that a focus on one part of speech would limit the conclusions regarding vocabulary acquisition in general that could be drawn from their research.

b) Opinion remains divided as to whether L2 learners' comprehension and retention of previously unknown vocabulary does occur in different ways or at different rates for different parts of speech. Research on this topic has been very limited, both in the number of studies conducted and in the scope of this research. Specifically, no studies that the author is aware of have set out to investigate part of speech differences in the comprehension and retention of unknown words through dictionary use.

c) With regard to the effect of dictionary definition type on sentence production, contrasting results for verbs investigated in Chapter Four as compared with results for a mixed set of target words in a study reported by Nesi (2000: 84-91) suggest that part of speech may be a factor worthy of attention. Further research on this topic may provide data that would have a bearing both on research into dictionary use and on choices made by dictionary makers regarding definition styles.

We will continue by reviewing research into the question of whether language learners' acquisition of previously unknown words is affected by their part of speech. Following this, we will give a description of the experiment into the comprehension, use, and retention of previously unknown adjectives. We will then consider the three sets of data obtained through the study: for the English sentences produced by the participants, for the translation equivalents the participants gave for the target words, and for the participants' retention test scores. We will look at these results from the following perspectives:

i) According to the part of speech of the target words, as we compare the results of this experiment with those for verbs reported in Chapter Four;

ii) According to the materials for the target words received by the two groups;

iii) According to the definition type appearing first for the words in the materials provided to the Dictionary group.

We will then summarize the important findings of this experiment, review the methods and instruments employed here in the light of the results, and propose how further research on this topic may be conducted.

5.1.1 Part of speech and vocabulary acquisition

We will now go on to briefly review research investigating whether part of speech is a factor in the learning of words. There has, in fact, been relatively little research focused on this particular issue, and rather more commentaries or interpretations of this research. Rodgers' (1969) study involving the learning of Russian and English word pairs concluded that nouns are easiest to learn, followed by adjectives and then verbs. Philip's (1981) investigation into the nature of word difficulty in L2 word learning also found that nouns were easier to learn than verbs or adjectives. Further, Gentner's (1982) study of children learning their L1 concluded that nouns are learned first and so, logically, with greater ease than words in other word classes. In Ellis and Beaton's studies (1993, 1995), too, part of speech appeared to be a significant indicator of learnability of L2 words, although they do give a word of caution (Ellis and Beaton, 1995) by suggesting that the relative imageability of words in a given word class is a more pertinent factor to consider than merely the word class of the words under investigation.

The conclusions of the earlier of these studies were initially widely accepted, with Nation (1990: 48), for example, simply repeating Rodger's findings, and confirming that in guessing from context, nouns and verbs are usually easiest. Since then, some researchers have challenged the findings of the studies, with Laufer (1997) pointing out that in Rodger's (1969) study, the particular word forms of the Russian verbs presented to the subjects may have distorted the findings. As for Gentner's (1982) findings, Ellis and Beaton (1995) observe that the first nouns that children learn are concrete nouns and proper nouns, and that these cannot be said to be typical of their respective word class in general. This challenges the implications that have been drawn from Gentner's study about the learnability of words from different word classes, as the data given for different word classes are derived from non-typical sets of words.

As Singleton (1999: 142) notes, Ellis and Beaton's (1995) comments lead us forward to consider that the words in some word classes may be typically more easy to learn because they are typically more imageable. There are also other aspects of words that may typically be found in words of a particular class and so affect learnability. These include the concreteness or abstractness of words. As De Groot (1993) and Hatch and Brown (1995: 220) observe, concrete nouns tend to be learned earlier, remembered better, and translated faster than abstract nouns. Such evaluations may not be restricted to nouns but also to other word classes.

If we think in terms of the learning burden of words in different word classes, we may find that there are generalizable differences here too. These differences may also vary widely from language to language. Laufer (1997), for example, points out the extra load in learning an English verb such as *break*, with its lemma *break*, *breaks*, *breaking*, *broke* and *broken* as compared with an adjective such as *large*. She suggests that while it may be misleading to say that, for example, verbs are more difficult than adjectives, the intrinsic element of morphological complexity may well lead us to the same conclusion. In some languages other factors may affect the learning burden; Hatch and Brown (1995: 256), for example, refer to the learning difficulty of adjectives in languages where they change their form in agreement with the gender and singular or plural form of nouns. From a connectionist perspective (see, for example, MacWhinney, 1997: 118-122), the number or strength of semantic, collocational, associative or other links which a word may have within the mental lexicon may, typically, be different depending on whether the word is a noun, a verb, or an adjective. We would expect this factor to affect word retention, especially over longer periods.

Brown (1993) points to the issue of the saliency of words in the context in which are encountered. She draws no conclusions regarding typical relative saliency of different parts of speech but this is a factor that may affect whole word classes in different degrees. Saliency is a factor that depends on the context in which words are encountered and, as we are considering word learning from different contexts, the variable effect of context on words in different word classes is certainly a factor that should not be overlooked. Related to this, and as noted above, Nation (1990: 48) observes that nouns and verbs are easier to guess from context than adjectives. This would have direct implications both for dictionary use and vocabulary of words according to their word class. Finally, as observed in Chapter Four when production of sentences for verbs was compared with Nesi's (2000: 84-91) research with a mixed set of words, it also appears that the effect of dictionary definition type on productive use by language learners may depend on the word class of the L2 words involved.

This brief overview of literature in the area points both to the limits of research so far and to the growing interest in aspects of words that may variably affect comprehension, production, and retention of words in different word classes. Many general questions remain unanswered, while specific areas such as the effect of word class on learning from context and dictionary use have yet to receive serious attention. We will seek to understand these issues better as we go on to describe a study of adjective learning and use through encounters of unknown words in multiple contexts and in monolingual dictionary entries.

5.2 The study

This experiment, like that reported in Chapter Four, investigates the comprehension, production, and retention of previously unknown L2 words. In Chapter Four, the focus was on verbs while the study reported here focuses on adjectives. Otherwise, the methods employed were largely the same. For a set of previously unknown English adjectives, one group of intermediate level Japanese learners of English were given a set of monolingual learner dictionary entries while a comparable group were given a set of corpus-drawn example sentences for each word. As before, the participants' tasks were to compose a sentence for each target word then to provide a translation equivalent for each target word. Three weeks later, the participants were given a vocabulary retention test, a kind of gap-fill exercise in which they had to match the example sentences or definitions with the correct target word.

5.2.1 Hypotheses

There are four hypotheses proposed regarding the results of this study. Justification for these hypotheses is provided below. The hypotheses are as follows:

1) There will be higher proportions of acceptable sentences than in the study reported in Chapter Four;
2) Example Sentences group participants will be unable to identify the meaning of many the target words from their contexts;
3) There will be lower average retention rates as compared with the study with verbs, both for participants using MLD definitions and for participants using example sentences;
4) For the Dictionary group there will be little or no difference in the proportions of acceptable sentences according to whether the first definition is taken from the COBUILD (COB2) or the Longman dictionary (LDOCE3).

The justification for the first hypothesis is simply that the syntax for English adjectives is generally simpler than for verbs. As a result, a greater number of participants should be able to produce acceptable sentences for a larger proportion of target words.

The second hypothesis is proposed because of the relatively marginal circumstances of adjectives in a sentence as compared with the central role that verbs typically occupy in English sentences. This would result in a large difference between the two groups' results for translation equivalents.

The lower retention rates are proposed in the third hypothesis for different reasons for the two experimental groups. For the Dictionary group, this would be because the simpler syntactic environments of adjectives would require less effort to understand and produce sentences for the target words; this would result in lower levels of retention. For the Example Sentences group, the lower retention rates would be due to the less informative and less memorable contexts for the adjectives than were provided by the example sentences for verbs.

Finally, various reasons lead us to the fourth hypothesis that definition type will have little or no effect on Dictionary group participants' success in producing acceptable English sentences for the target words. While the study on verbs reported in Chapter Four showed a small, but significant, advantage for the full-sentence style COBUILD definitions, Nesi's (2000: 84-91) research on the effect of different definition types for a mixed set of target words produced no significant difference between definition types. Wingate's study (2002: 193-219) using the two definition types in German for Chinese learners did find a difference for more advanced learners of German but not for lower level learners, and no conclusions could be made regarding part of speech.

5.2.2 Participants

Thirty-two Japanese university students were recruited as participants for this experiment. A slightly smaller group of 24 of the original participants also took part in the retention test conducted three weeks after the learning phase of the experiment. All the participants were in their second year at university, aged 19 – 21, majoring in English at a middle-ranking Japanese university. They had all received about seven years of formal instruction in English. Although individual TOEFL scores for the participants were not available, the great majority of TOEFL ITP scores for 2nd year English Department students at this university range between 410 and 480.

5.2.3 Target words

There were 20 target words, all adjectives and all judged to be unknown to the participants. The words were selected according to the following criteria:

1. To exclude words that are too rare or too likely to be known, only words in the COB2 (Sinclair et al., 1995) 'two diamond' frequency band were included for consideration: from the 3,400th to the 6,600th word. In addition, only words not listed in the JACET 8000 word list were included.
2. Any words with more than one sense recorded in the dictionaries were excluded.
3. Only words for which there were at least 30 occurrences in the COBUILD Direct 50 million word corpus were included.

4. A list of 85 adjectives compiled according to the above guidelines was presented to the participants as a pretest (Appendix 5.1). Words correctly identified by any participants or often mistaken for other words were excluded from consideration.
5. The 20 target words were chosen from the remaining 54 words on the list. The aim of selection at this stage was to include a variety of adjectives: adjectives only followed by nouns (e.g., *illicit*), adjectives only preceded by link-verbs (e.g., *afoot*), adjectives for which both of these main usages is possible (e.g., *colossal*), adjectives followed by a preposition (e.g., *averse*), and adjectives with other syntactic restrictions (e.g., *galore*, which only follows plural nouns or mass nouns).

The number of target words was set at twenty to allow enough time in a 90-minute class period for participants to complete the tasks as required. The final 20 target words are as follows:

afoot, *akin, averse*, *bereft*, *blatant*, *callous*, *colossal*, *defunct*, *eerie*, *fleeting*, *furtive*, *galore*, *gaudy*, *illicit*, *inviolate*, *lenient*, *morbid*, *obese*, *poignant*, *quaint*

5.2.4 Learning materials

As described above, the participants were randomly divided into two groups, the Dictionary group and the Example Sentences group, and used one or other of two different resources for the target words to complete vocabulary tasks.

Dictionary group

This group received two dictionary definitions for each of the target words, taken from the Longman Dictionary of Contemporary English, 3rd Edition (LDOCE3, Summers et al., 1995) and the Collins COBUILD English Dictionary, 2nd Edition (COB2, Sinclair et al., 1995). These are shown in Appendix 5.2. The order in which the definitions for the target words appeared first in this group's materials alternated between the two dictionaries. The main typical differences between the definitions for the two dictionaries are in the defining styles and typical definition lengths. The COBUILD dictionary uses whole sentence definitions containing the headword, while for the Longman dictionary the definitions are usually in phrases and the headword does not usually appear in the definition. Partly as a result of the defining styles, definitions in COB2 are, typically, longer than those in LDOCE3.

For this study, only the definitions were used for the materials; no illustrative sentences, grammar codes or other information from the dictionary entries were included. The reason for this was to be able, as in Black's (1986) study, to investigate the role of definitions alone in the comprehension and retention of unknown words. A further justification for this is that one feature of many full-content handheld dictionaries is to not show illustrative sentences, unless specifically requested; with these, MLD users may increasingly seek to understand the meanings of looked up words without reference to anything other than the definition.

Example Sentences group

This group received three example sentences for each of the target words (see Appendix 5.3), drawn from the 50-million word COBUILD Direct corpus. Sentences were selected by an experienced lexicographer to display typical syntactic patterns and collocations, and for their comprehensibility. Wherever possible, sentences were taken directly from the corpus without changing them in any way. In a few cases, however, parts of sentences were deleted if the sentences were too long but otherwise ideal. Where more than one grammatical pattern for a word occurs frequently, this is reflected in the choice of example sentences.

5.2.5 Vocabulary retention test

A vocabulary retention test was conducted three weeks after the vocabulary tasks in which the above materials were used. For this test, participants were given a multiple choice answer sheet for the target words (Appendix 5.4), for which the participants had to identify the word which matched either the example sentences or the definitions. They were given the same materials as three weeks previously as shown in Figure 5.1, except, that the target word was deleted from the definitions or example sentences and the test items were randomly reordered. More samples are also shown in Appendix 5.5.

Figure 5.1

Gapped example sentences and definitions as used in the retention test.

A. haphazard quaint defunct idyllic

[For Example Sentences group:]

All the shops are closed due to a _______ Roman tradition.

I am aware of a number of _______ pastimes that are performed in rural parts of Britain.

Fingleton, in one of his many book, made it clear how ______ he found all this stuffiness.

[For Definitions group:]

Unusual and attractive, especially in an old-fashioned way.

Something that is _______ is attractive because it is unusual and rather old-fashioned.

The four choices for each item consisted of the correct answer, one other target word, and two words that had been encountered by the participants in the pre-test of 85 adjectives.

The form and focus of the retention test was chosen for two main reasons. Firstly, in comparable research, including the study reported in Chapter Three, where a retention test requires the production of the meaning of isolated target words, retention rates have generally been very low, even when such tests were conducted immediately following the encounters with definitions or example sentences. In this study, where evidence of longer term retention was sought, the recognition of context was chosen as the subject of a more sensitive test of word knowledge retention.

Secondly, this retention test did not depend on participants having used the target words correctly or understood the words' meanings, but only to recognize the contexts for the words which they had sought to understand and use in sentences. In the case of a test of word meaning retention, participants would first have to have understood the meaning of the target words. Comprehension rates were fairly low for the Dictionary group in this study and even lower for the Example Sentences group; to focus on retention of word knowledge by participants who had understood the words would have resulted in a focus on very limited data.

5.2.6 Rating

Two experienced native speaker teachers of English rated the English sentences produced by the participants and two highly proficient Japanese teachers of English rated the participants' translation equivalents for the target words. For the rating of the sentences, a detailed set of guidelines (Appendix 5.6) were used to help the raters determine what would count as an acceptable sentence. In addition, a set of concordance lines for each target word was provided in case raters needed more information about word use to reach a decision. For the raters of the translation equivalents, entries from two monolingual dictionaries other than those used in the experiment were provided for each word: *Oxford Advanced Learners Dictionary, 5th Edition* (Crowther et al., 1995) and *Cambridge International Dictionary of English* (Proctor et al., 1995). The raters also consulted the widely used *Kenkyusha English-Japanese Dictionary for the General Reader* (Matsuda, 1999).

The raters of the English sentences judged each sentence as Acceptable, Unacceptable, or Questionable. Interrater reliability for the sentences for all the items was .83. All differences were resolved at a joint meeting, at which the Questionable category was also discarded because of the very few items rated in this category; items initially rated as questionable were rerated as either Acceptable or Unacceptable.

The raters of the Japanese translation equivalents of the target words rated each translation as Correct, Partially Correct, or Incorrect. Rater guidelines are in Appendix 5.6. Here is one example of how the equivalents were judged: For *obese*, translations meaning 'very fat' or 'too fat' were judged correct, while meanings approximating to 'fat' were judged to be partially correct. Partial superordinates for *obese*, such as equivalents meaning 'big' or 'unhealthy' were rated as incorrect. Overall interrater reliability was .78. Differences of opinion were resolved at a meeting of the two raters.

5.3 Results

There were three main sets of results: for the participants' English sentences, for their Japanese translation equivalents for the target words, and for the retention test. These are presented one by one in Tables 5.2 to 5.4, and shown together in Figure 5.5.

5.3.1 English sentences

In this analysis, as in Chapter Four, omissions and questionable or unacceptable English uses were collapsed into one category: not correct sentences. As can be seen from Table 5.2 and Figure 5.5, there was a large average difference between the two groups as regards their production of acceptable English sentences. A t-test conducted on the participants' scores for the English sentences for the two groups confirmed that there was a significant difference between the two groups (t = 2.6 p<.02).

Table 5.2
Acceptable English sentences (maximum = 20)

	M	S.D
Dictionary group (N = 16)	10.19	6.46
Example Sentences group (N = 16)	6.31	3.12

5.3.2 Translation equivalents

The participants' Japanese translation equivalents for the target words were rated as matching the English word, partially matching, or not matching. Where no answer was given, this was included in the not matching category. Matching and partially matching categories were collapsed into one category of acceptable translation equivalents. As shown in Table 5.3 and Figure 5.5, there was a very large difference between the two groups for their production of acceptable translation equivalents. A t-test conducted on these results (t = 17.89, p<.001) confirmed that there was a significant difference between the two groups' results.

Table 5.3
Acceptable translation equivalents of target words (maximum = 20)

	M	S.D.
Dictionary group (N = 16)	14.3	0.94
Example Sentences group (N = 16)	4.1	2.11

5.3.3 Retention test results

Twenty-four of the original participants, 12 per group, made themselves available for the retention test conducted three weeks after the initial part of the study. In this test, the participants' task was to match the target words with the gapped contexts (definitions or example sentences) in which they had encountered the target words during the learning task. Retention test results are shown in Table 5.4 and Figure 5.5.

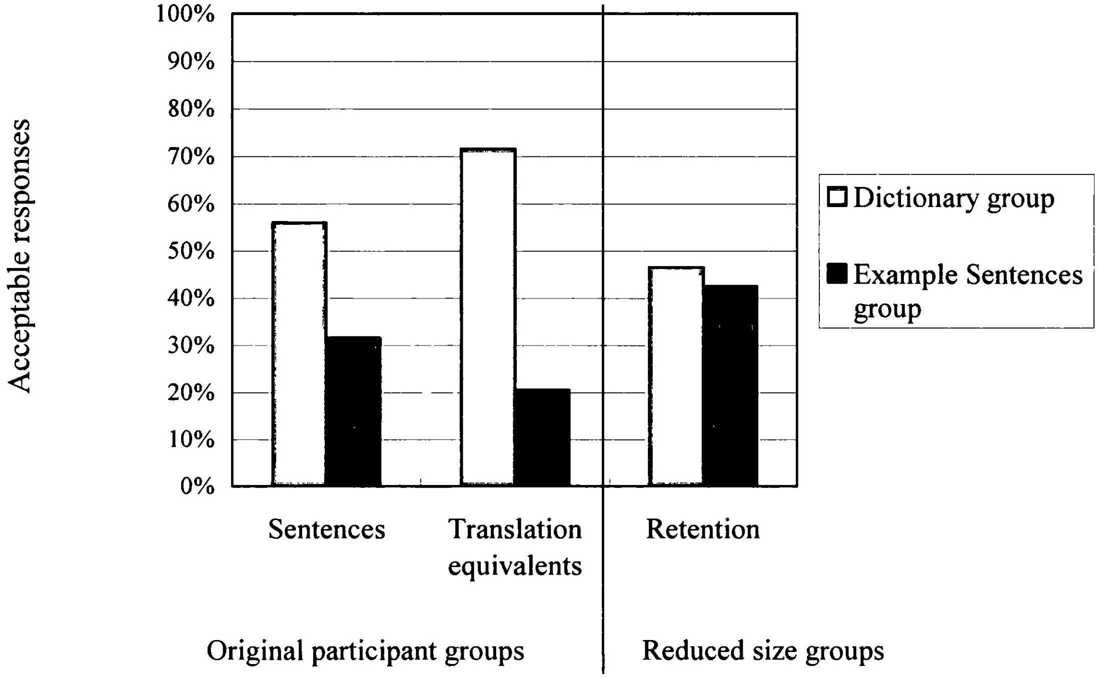
Table 5.4
Retention test scores (maximum = 20)

	M	S.D.
Dictionary group (N = 12)	9.3	3.5
Example Sentences group (N = 12)	8.5	4.89

There was little difference between the mean results for the two groups, and standard deviation levels, especially for the Example Sentences group, were high. A t-test conducted on these results confirmed that there was no significant difference between the two groups' results.

Figure 5.5

Mean % of acceptable English sentences, translation equivalents, retention scores



5.4 Review of results

We will briefly review the main results of this experiment, before going on to the Discussion section. Data relating to the first, second, and third hypotheses are shown together in Figure 5.5 above, while comparisons of data for English sentences and translation equivalents are provided in Figures 5.6 and 5.7 below.

1. For the English sentences, the Dictionary group participants produced an average of just over ten acceptable sentences for the 20 target words, while the Example

Sentence group participants produced a little over six acceptable sentences, on average. This means that participants in the Dictionary group were typically able to produce reasonable sentences for just over half the target words, compared to just under one third by the Example Sentences group participants.

2. The Dictionary group participants produced, on average, just over 14 acceptable translation equivalents for the 20 target words (71%) , compared to little over four (21%) for the Example Sentence group participants. As is shown in Figure 5.6, the Dictionary group participants almost all produced more acceptable translation equivalents than acceptable English sentences, while Figure 5.7 shows that for the Example sentences group the opposite was generally true.

3. For the retention test, there was very little difference between the two groups. Both got average raw scores of around nine correct answers out of 20 and, while the Dictionary group scored marginally better, there was no significant difference between the two groups' results. Although the raw scores suggest correct answers for almost 50% of the test items, we must bear in mind that as this was a forced choice multiple-choice test with four options per item, a large proportion of the result may be attributable to guessing, rather than knowing, the answer. However, even when we take this factor into account, the test did provide some evidence of word knowledge retention for both groups.

Figure 5.6

Acceptable translations and English sentences for Dictionary group

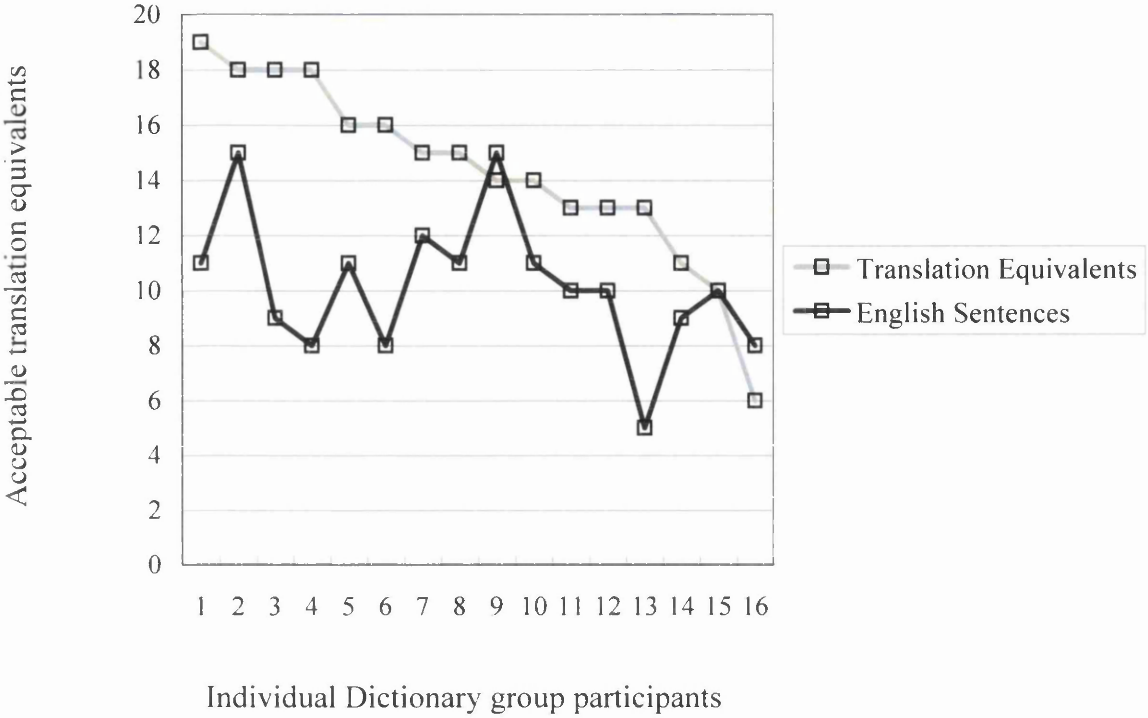
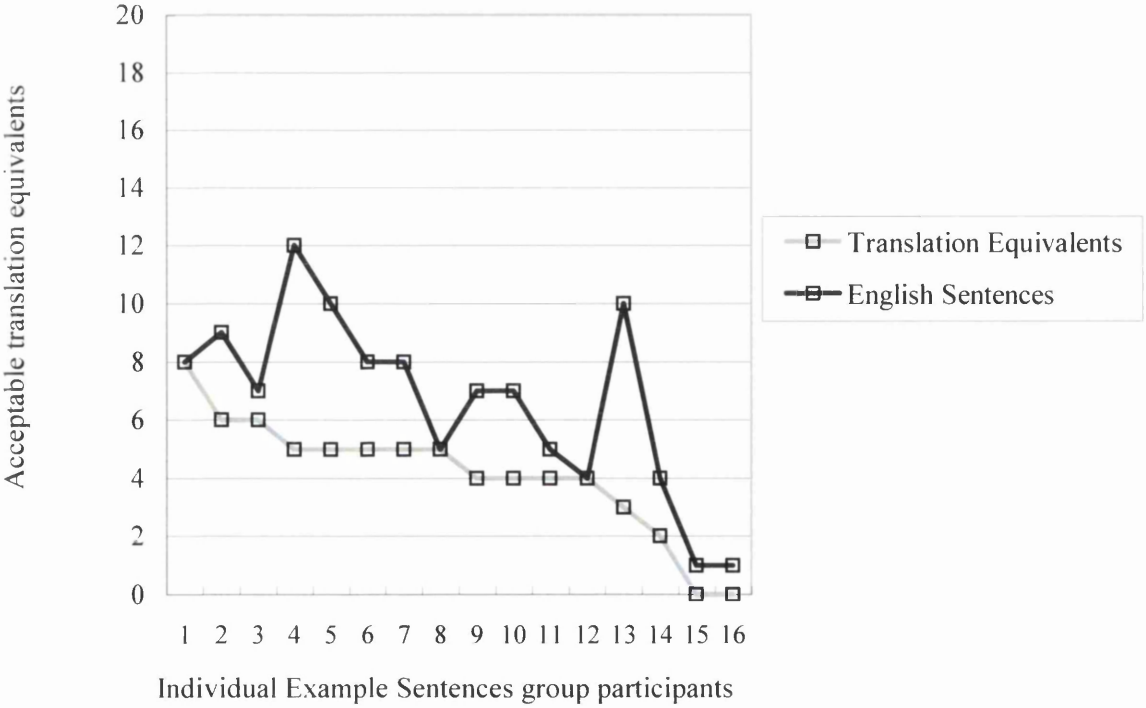


Figure 5.7

Acceptable translations and English sentences for Example Sentences group



5.5 Discussion

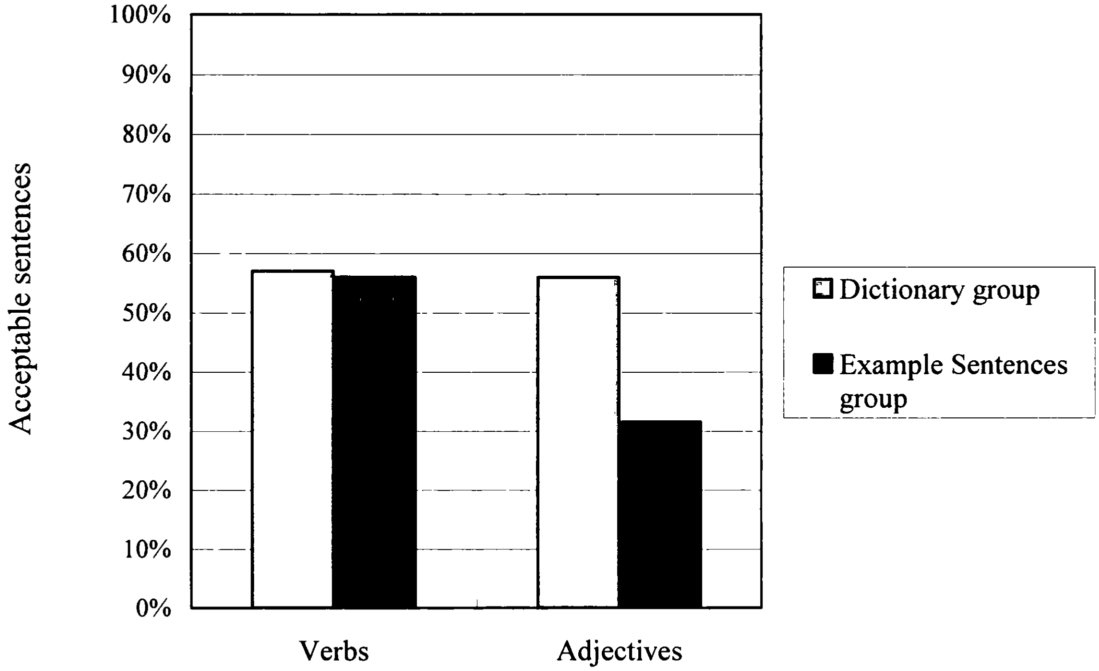
In this section, we will discuss the three sets of data in relation to each other and in relation to the data from the experiment for verbs reported in Chapter Four. We will also consider the value of the instruments employed for the evaluation of word comprehension, use, and retention. We will begin by considering the production of English sentences, and will at this point discuss the effect of different definition types on production. Following this, we will go on to discuss the issue of the comprehension of unknown words through dictionary definitions and limited multiple contexts. Here, we will focus in detail on differences between definition types for different parts of speech, and the effect that these differences may have on comprehension of the definitions. Finally, we will consider the retention test results in more detail: how they may be interpreted, what they may tell us both about retention of previously unknown word knowledge and about testing methods for identifying word knowledge retention.

5.5.1 English sentences

One important finding from this study was that average sentence production rates were especially low for the Example Sentences group. As Figure 5.8 shows, this figure is low both in comparison with that for the Dictionary group and when compared with the scores for the two groups in the study on verbs described in Chapter Four.

It is not immediately clear why these differences exist or should be so pronounced. One clue as to a possible factor may be found in the relationship between participants' scores for target word comprehension and scores for English sentences. In the study on verbs, Example Sentence group participants were often successful in writing English sentences for words which they had failed to understand. For adjectives, this was only rarely the case; where Example Sentence participants were unable to guess an adjective's meaning, they were also, usually, unable to write an acceptable sentence for the word.

Figure 5.8
Mean % of acceptable sentences for verb and adjective target words



One possible reason for the relatively low success rate in sentence production for Example Sentence participants in the study focusing on adjectives is the centrality, both syntactically and semantically, of the verbs in the example sentences as compared to the adjectives in their example sentences. For the target word verbs, participants could usually make a reasonable guess as to meaning, even if it was often an inaccurate guess. For these words, they were also usually able to produce acceptable sentences on the basis of their guesses and with the syntax of the example sentences to guide them. For the adjectives, however, few participants were able to write acceptable sentences for words where no reasonable meaning could be guessed at from the contexts provided. There is some confirmation for this difficulty in the markedly higher rate of participants' refusal to write sentences or guess at meanings in this study as compared to that for verbs.

Lexical context is obviously an important factor determining how difficult nouns or verbs may be to comprehend or to learn. In the light of the findings from this study, research to date on word comprehension and acquisition difficulty for different parts of speech may need to be revisited to see what types of context were used and how these may have affected outcomes.

5.5.2 Definition types and word classes

In Chapter Four, we noted that the number of acceptable English sentences produced by the Dictionary group participants seemed to depend in part upon the definition type of the first definition for each word in their materials. Typically, Longman definitions are not full sentence definitions while COBUILD definitions always are. In the experiment described here, however, there was almost no difference between participants' production of English sentences according to which dictionary's definition appears first for the word, as can be seen in Table 5.9. High standard deviations, too, indicate the wide variation in correct numbers of acceptable sentences from word to word. In the Discussion section below, we will discuss various possible factors that may account for these results..

Both for the definitions from LDOCE3 and from COB2, the overall proportion of participants producing acceptable sentences was 50%, although there was wide variation from word to word. For example, only two out of 16 participants produced acceptable sentences for *blatant*, as compared to 13 out of 16 for *colossal*.

Table 5.9
Scores for acceptable sentences for Dictionary group according to definition

	Total	M (max. 16)	S.D.
LDOCE	80	8.0	5.31
COBUILD	83	8.3	4.32

These results do confirm the second hypothesis, that the source or type of definition, would have less effect on participants' sentence production using adjectives than it did for verbs. The reason for proposing this hypothesis was based on the findings from Nesi's study (2000: 84-91) of sentence production, in which she found no significant difference attributable to definition type. Since the study reported in Chapter Four did reveal a difference for verbs, it was assumed that this would not be evident for adjectives. However, we have suggested no reason for this expected lack of difference relating to the nature of dictionary definitions themselves. Only through a closer look at the definitions used by the two dictionary publishers, Longman and COBUILD, does a likely explanation come to light. For the verbs investigated in Chapter Four, with one exception, Longman did not use sentence-style definitions. Yet LDOCE3 does use this style of definition for some of the target word adjectives in this chapter: for two out of ten of the adjectives where the Longman definition appears first and for a further two out of ten where the COBUILD definition appears first. Perhaps more important, however, is the difference between COBUILD definitions for the two parts of speech.

One important quality claimed for sentence definitions (Hanks, 1987: 121-122) is that the definition itself provides a model of typical use of the word; the value of this was confirmed the participants' sentence production as reported in Chapter Four. In the case of a large number of target word adjectives, however, the sentence definitions do not provide this model of typical usage. One important new goal in the second edition of the Collins COBUILD English Dictionary (1995) was to convey pragmatic aspects of the word such as speaker attitude or emphasis through the definition. This resulted in definitions such as the following for two of the target words:

> **Blatant**
> You use **blatant** to describe something bad that is done in an open or very obvious way in order to emphasize your shock or surprise that it is done in such an open or obvious way.
>
> **Furtive**
> If you describe someone's behaviour as **furtive**, you disapprove of them behaving as if they wanted to keep something secret or hidden.

In neither of these definitions, nor in four other of the COB2 definitions for the target adjectives, does the definition show the word being used in a typical context, or in a true syntactic context at all. Five of the six definitions of this type appear first for the target word, representing half of the definitions appearing first for that dictionary in the Dictionary group participants' materials. In addition, by highlighting pragmatic information, these COBUILD definitions tend to be longer than they would otherwise be. For the six target words where this is the case, the definitions average almost 24 words per definition, as compared with an average of 9 words each for Longman definition for the same words. One likely consequence of this, especially since these longer definitions are also more complex, is that the participants would ignore the COBUILD definitions and go directly to the shorter and more accessible Longman definitions. There is some limited identifiable evidence of this for some of the target words. For example, the expression "bereft of hope" which occurs only in the second-placed Longman definition for the word *bereft* is used by 20% of participants in this group. For the word *blatant*, there is direct evidence in only a couple of the participants' sentences but the meaning 恥ずかしい ("hazukashii": *embarrassed, ashamed*) given for this word by at least three participants can be traced to the words *embarrassed* and *ashamed* which only appear in the Longman definition.

The above factors help to explain why, for the Dictionary group, there was little or no difference in the English sentences produced attributable to the source or style of the dictionary definition appearing first in the participants' material. This analysis has also brought to light an unexpected difference between learning words through dictionary use for words from different word classes. On the basis of, admittedly, limited data, we can see that the comprehensibility and usefulness for production, of definitions from different monolingual dictionaries may depend to a large part on the word class of the word being looked up. This may mean, for example, that students of English with the choice available to them would be well advised to look up verbs in their COBUILD dictionary and adjectives in their Longman dictionary.

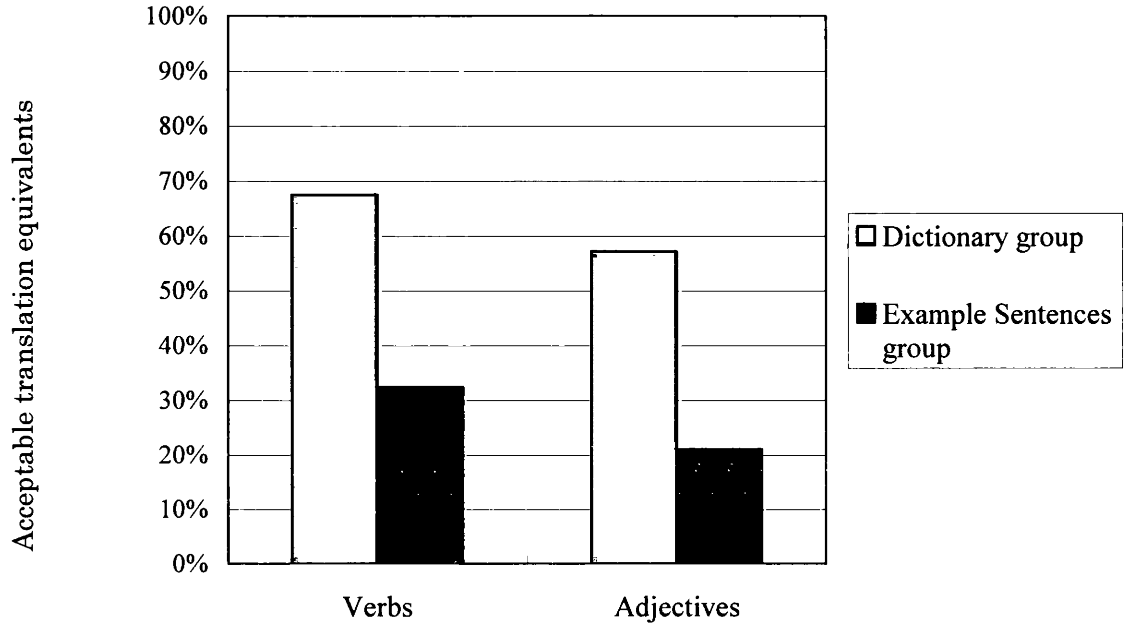
Wingate (2002: 130-134), in her investigation of comprehension of dictionary definition types, notes how the actual form of COBUILD-style defining sentences is largely determined by the part of speech. She suggests that this factor may affect comprehension of the definitions. She reports the findings of her research into dictionary use by Chinese learners of German (2002: 197-207) comparing comprehension rates for COBUILD-style full sentence definitions with those for traditional definitions of the type typically found in monolingual dictionaries written for native speakers of the language. However, with only three or four words for each part of speech, we should not be surprised that she found no significant results regarding the effect of part of speech on comprehension of sentence-type MLD definitions.

5.5.3 Translation equivalents

Regarding translation equivalents for the target word adjectives, most striking is the small number of words for which Example Sentence group participants were able to produce acceptable translation equivalents. This is shown in Figure 5.10 below. We have briefly touched on reasons why this may be, suggesting that, both syntactically and semantically, adjectives are not central to the sentences in which they appear, at least for sentences presented in isolation as in this study. An alternative argument is that adjectives add meaning to sentences; if an adjective is not known, that part of the sentence may become devoid of meaning to the reader. If this is the case, the strategy of trying to infer meaning from context may often be a pointless activity for unknown adjectives.

Figure 5.10

Mean % of acceptable translation equivalents for verb and adjective target words



A couple of examples of sets of gapped sentences may best illustrate the challenge faced by the participants in their effort to infer meaning of unknown adjectives, since a totally unknown word may be equated with a gap in terms of the extent to which it contributes to comprehension. The reader's attempt to guess the deleted words, or at least the meanings of the deleted words, may illustrate the problems faced by the participants of this study when encountering unknown words:

Set 1

Ali described a ________ moment when she was a few years old.
It is a ________ love story starring Juliet Aubrey and Robert Carlyle.
The remainder of his story is both ________ and disturbing.

Set 2

Naomi Wallace's play is ________, scary, and full of pent-up emotional violence.
They paused, hearing an ________ sound echoing through the woods.
Her works as a young adult were dark, ________ abstract paintings of people.

Few adult native speakers of English would be able to guess at the meaning of the missing words, and fewer still to identify the words themselves: *poignant* for the first set of sentences and *eerie* for the second. In fact, most of the six sentences above appear to be complete even with the target words deleted. This helps us to appreciate the peripheral state of many adjectives, compared to verbs, in the contexts in which they are typically found. These gapped sentences also help us understand why Example Sentence group participants were able to arrive at the meaning for so few of the target words in this study.

As we can see in Figure 5.10, while the Dictionary group produced almost three times as many acceptable translation equivalents as the Example Sentences group, at well under 60% this figure is still low. If this figure reflects in any way the general comprehension rates of MLD entries by intermediate level learners of English, we should not be surprised at their reluctance to use MLD and their preference for bilingual dictionaries. As a response to these figures, we need to consider what obstacles to comprehension there may be for so many of the target words: obstacles concerning the target words themselves, the definitions, the participants' English ability, or their dictionary use skills.

If we consider the definitions from the perspective of user difficulty, one factor is their length. Although, as Wingate (2002: 42) notes, redundancy may often assist comprehension, the opposite may be true of the increasing syntactic complexity, and sheer volume of reading, for longer definitions. Table 5.11 shows average definition lengths for the two dictionaries and for verbs and adjectives used in the studies reported in Chapter Four and in this chapter. These figures need some explaining before we go on to reflect on what the numbers might tell us about definition difficulty. For LDOCE3, much information regarding the register or variety of a word is abbreviated, italicized, or otherwise codified and was not counted as words in the definitions. For COB2 entries, however, this information is usually appended to the defining sentence: either at the beginning (e.g., *In informal British English...*) or at the end (e.g., *: a formal word.*). The figures in the row "Full definition" in table 5.11 include these phrases, while in the row labelled "Main body", these phrases are excluded.

Table 5.11
Average definition length for target words in studies in Chapters Four and Five

	Verbs		Adjectives	
	LDOCE3	COB2	LDOCE3	COB2
Full definition	14.0	22.6	9.5	20.2
Main body	14.0	19.6	9.5	18.7

As noted before, the length of the definition, especially when comparing different definition styles, is not necessarily a measure of word difficulty. What is of interest, though, is the large difference between lengths of Longman definitions for verbs and adjectives, and that this difference is much less apparent for the COBUILD definitions. This is partly due to the redundancy that is an intrinsic part of sentential definitions. A very simple example of this is for the word *akin*:

> If one thing is **akin** to another, it is similar to it in some way: a formal word.

Even this simple definition uses 18 words, three of which provide information about register. This compares with six words for the LDOCE3 definition for the same word, while the shortest LDOCE3 definition for a target word adjective, *colossal*, is only two words long: "extremely large". Another reason for the greater length of the COB2 definitions, especially for the adjectives, is that many of these definitions provide more pragmatics information than the LDOCE3 definitions: to highlight user attitudes (e.g., for *gaudy*, ": often used to express disapproval and to suggest that it is vulgar.") or emphasis (e.g., for *colossal*, "...you are emphasizing that it is very large.").

The contrasting definition lengths for the two dictionaries, so marked for target adjectives, suggest that the adjectives were easier to define at a basic level than the target word verbs, but that fuller, more accurate definitions may result in longer, harder definitions. This observation may add support for the view (Rundell, personal communication) that with MLDs, more is not always better. Increasing accuracy, and volume, of word definitions and entries may have a negative effect on comprehension, especially for intermediate level users of these dictionaries. Although the monolingual learner dictionary market has long been dominated by the full-size "flagship" dictionaries of the major publishers, these results add weight to the belief that for many language learning dictionary users, big is not so beautiful.

5.5.4 Retention

The figures for word knowledge retention for the twenty target word adjectives are shown in Table 5.4 and Figure 5.5. As in Chapter Four, a forced choice multiple choice test was used to evaluate retention levels of context recognition for the target words. As with the data for verbs, the raw figures for retention need to be interpreted. For both the

Dictionary group and the Example Sentences group, the participants chose the correct answer for between 40% and 50% of the items. Since random guessing of the correct answer from the four choices per item for the whole test would have produced a correct answer rate of 25%, these figures are substantially higher than chance and we can be fairly sure that some retention took place.

Recognition of context alone, which was the focus of the retention test, is not an immediately useful achievement in terms of vocabulary acquisition. It may, however, provide a good indication of attention paid to a word and to its context. The comparable figure for the two groups suggests that level of attention paid to the word form while trying to understand and write sentences for the target words was similar, regardless of differing levels of success in terms of acceptable sentences or accurate translation equivalents.

Comparison between retention rates for verbs and adjectives for these two studies would be largely meaningless as there were different numbers of test items for the two experiments (24 for verbs and 20 for adjectives), and a different number of distracters for each retention test item (4 per item for verbs and 5 per item for adjectives). We will address this question more fully in Chapter Six.

5.6 Evaluation methods

As we review the experiment described in this chapter, we can see that there are various problems with at least two of the instruments employed: with the "write a sentence" task as evidence of accurate productive use of the target words, and with the use of a simple multiple-choice format for the retention test. We will conclude this chapter by listing problems with these methods before considering more appropriate methods or instruments for research into L2 vocabulary acquisition via MLD use.

5.6.1 English sentence production

Although a fairly widely used procedure among researchers into productive dictionary use (see, for example, Miller and Gildea, 1985; Fischer, 1994; Nesi, 2000: 71-116), there are at least four problems with this approach:

i) It is not a familiar activity for language learners to be given a word and told to write an isolated sentence using the word. Some learners will be more proficient at this activity than others, and this ability may well be independent of both general L2 ability or of comprehension of the materials they are given.

ii) It is difficult to make judgments of acceptability of isolated sentences with much objectivity. Despite various measures taken to improve levels of inter-rater reliability, and although these do appear to be better for adjectives than verbs, they have remained low and reflect poorly on the validity of the research.

iii) It takes a long time for participants to write each sentence. As a result, the number of items that can be tested is severely restricted. As the problems of focusing on small numbers of target items become increasingly evident, it is important to use instruments that allow for the evaluation of larger numbers of items.

iv) The production of acceptable sentences is not evidence of comprehension of the target words. In some cases these sentences will demonstrate lack of comprehension but, in many cases, a correct sentence cannot be seen as proof of comprehension of the target word.

5.6.2 Multiple-choice tests

The multiple-choice method for testing word knowledge has various attractions. Above all, they are easy to create, fast to do, and easy to mark. They have also been shown (Nist and Olejnik, 1995) to be easier for learners to give responses for and so, in this respect, may be said to offer a more sensitive instrument for investigating word knowledge than many other measures. However, as Meara and Buxton (1987) point out, there are various problems with using these tests. Two important problems relevant to this research are:

i) Not all distracters are equal. Especially as we increasingly recognize that much word knowledge is partial, the choice of distracters to accompany the correct answer in a test item may often affect the selection of the correct answer, even if it is unknown to the testees. In a test item, if one or two distracters are recognized as not being the correct answer, for example, the chances of guessing the correct answer are substantially improved.

ii) Marking multiple-choice items is quick and easy, understanding the results is not. Random guessing for unknown test items usually results in a sizeable percentage of correct answers: 25% where test items contain one correct answer and three distracters. With this in mind, interpreting, for example, a score of 50% correct answers is not straightforward. We cannot take this at face value, claiming that half the items were known, (as in Black, 1986) but it is also difficult to say more than that the participants are performing substantially above chance levels.

5.7 Conclusion

Despite continuing problems with some of the instruments employed, this study has produced valuable data in an area in which research is still very limited. Among the findings are the markedly lower comprehension rates for adjectives both for dictionary definition users and for example sentence users, and the much lower acceptable sentence production rates for adjectives for the example sentence users. We have also seen how different defining styles in dictionaries for different parts of speech may affect both comprehension of the definitions and production of sentences using the definitions. Clearly, more research is needed on comparing verbs and adjectives in these contexts, and this is the purpose of the study described in Chapter Six.

Chapter Six: Adjectives and Verbs

6.1 Introduction

In this chapter we will look at the fourth in a series of experiments investigating the comprehension and recall of previously unknown English words as a result of contact with either monolingual learner dictionary definitions or corpus-drawn examples of target words used in sentences. The specific focus in this study is on different parts of speech: verbs and adjectives. We will begin by reviewing two important pieces of research in this field before going on to describe the experiment itself. As we do this, we will consider how various aspects of a study, such as the learning materials, target word comprehension and retention, and interrater agreement may be affected by whether the target word is a verb or an adjective. In this chapter, as well as reporting overall average results for different groups and parts of speech, we will also look at the materials for individual words in relation to responses that participants produced for these words.

6.1.1 Focus on verbs and adjectives

The study reported in Chapter Four investigated the use and learning of verbs and that in Chapter Five reported the use and learning of adjectives. Although, in Chapter Five, we compared the results of the two studies, in some ways they were not parallel studies; neither the groups of participants nor all the instruments employed were the same for the two studies. The purpose in this study is to focus specifically on comparing comprehension and retention of English verbs and adjectives. The main reasons for conducting this study, with its clear focus on two different parts of speech, are as follows:

i) In many of the vocabulary acquisition studies focusing on dictionary use, researchers have deliberately chosen a set of target words composed of words from different parts of speech, such as nouns, verbs, and adjectives. This suggests a belief that comprehension and retention rates for words in different word classes may differ. Without knowing what differences there are, a "balanced" set of target words is seen as the safest response.

ii) Research into vocabulary acquisition rates for different parts of speech has been very limited and remains inconclusive. Opinion is divided as to whether vocabulary learning and retention are different for different parts of speech, and there has been little discussion which takes account of the effect of different learning contexts.

iii) Almost no research about word classes has been conducted in the context of dictionary use. As we found in the studies described in Chapters Four and Five, there appear to be typical differences in definitions and example sentences for different parts of speech. Apart from any intrinsic difficulty, these factors may well affect word comprehension and retention for different parts of speech.

As we go on to review two major contributions to this aspect of research into vocabulary acquisition, we will see how much work in this field remains to be done.

6.1.2 Research into part of speech and vocabulary acquisition

In Chapter Five we briefly reviewed research and commentaries relating to the effect of part of speech on vocabulary acquisition. Here we would like to focus in more detail on two influential studies in which this issue has been addressed: Rodgers' 1969 study and Ellis and Beaton's study, reported in papers published in 1993 and 1995. We will consider the findings of the studies, the languages investigated, and the learning contexts involved.

Rodgers' (1969) study of vocabulary learning through oral drilling is often cited in relation to the typical learning rate for different parts of speech. His study focused on the learning of 300 Russian words by English speakers through oral drilling of the words. We are told that the target words were high frequency words but there is no indication as to whether there was a pre-test to determine prior knowledge of the target words, either as cognates of words in known languages or through prior exposure to the words. There are indications that some cognates were included, at least for French words that might well be known by the participants in the study. For example, one target word was 'MYX', meaning *fly*, which is pronounced in a way similar to 'mouche', the French word for *fly*. Within the set of 300 target words, there were words from various word classes and types. Word classes for which a sizeable number of words were tested, together with their average retention rates, are listed in Table 6.1 below. For these figures to be meaningful, we must assume that previously known words were either excluded from the study or were evenly distributed among the word classes. The results indicate that concrete nouns were most easily learned, that retention rates for abstract nouns and adjectives were similar to each other, and that much smaller proportions of adverbs and verbs were recalled.

Table 6.1
Major word classes of target words in Rodgers' 1969 study.

Word class/type	No. of items	Mean retention rate
concrete nouns	84	.58
adjectives	57	.41
abstract nouns	34	.36
adverbs	22	.23
verbs	59	.21

As we reflect on these figures, two issues are worthy of note. Rodgers observed that words which were more difficult for the learners to pronounce were less likely to be remembered; if phonologically difficult words were not evenly distributed among the different parts of speech, or were not typical of their respective part of speech, this may have affected results. The second point, noted by Laufer (1997), is that the form in which some of the verbs were presented may have presented additional difficulties for learning the verbs as opposed to the nouns. Her conclusion is that the results of the study might reflect the form in which the words were presented rather than their part of speech. Although imbalances in the phonological difficulty and in the form of the target words may have affected Rodgers' results to some degree, it is unlikely that these factors alone could account for the widely varying retention rates shown in Rodgers' study in which, for example, concrete nouns were almost three times more likely to be recalled than verbs.

Rather than rejecting Rodgers' findings, a more measured response may be to recognize their potential limitations. His data refer to high frequency Russian words learned by

English speakers through oral drilling. This research does provide an indication of learning difficulty for words from different word classes, but further research is necessary to confirm whether the findings have relevance for other languages, for other levels of word frequency, and for other learning contexts.

We will next go on to a study conducted by Ellis and Beaton (1993, 1995). This study is interesting in that it provides data on the effect of different learning strategies on the learning of a set of target words: by using noun or verb keywords as mnemonic devices, by repeating target words out loud, and by using whichever method the participants wanted. Although the main focus of this study is the use of keywords, it does provide valuable data relating to word class and ease of learning. The mother tongue of the participants was English and the target language German. There were 36 target words, half verbs and half nouns. Overall retention test results showed that with a recall rate of 68%, nouns were recalled more often than verbs at 53%. We are told that the advantage for nouns over verbs was observable in all learning conditions, although figures for each condition are not provided.

Ellis and Beaton (1995) suggest that two main factors may account for the greater general ease of learning for nouns over verbs. These are imageability and meaningfulness. They explain the imageability of a word as the degree to which it arouses a mental image, while meaningfulness is explained in terms of the number of associational relationships that a given concept has with other concepts. They give as examples the verb *to run* compared with *dog*, noting how many more attributes the latter has in comparison with the former.

The study by Ellis and Beaton provides valuable data confirming the ease of learning nouns as compared with verbs. It does, however, have weaknesses or limitations in two important areas. First, the experimental groups were small, with as few as eight or ten participants in three of the experimental groups. This means that although overall figures may be reliable, conclusions drawn from individual groups' results should be treated with caution. Secondly, for all groups the L1 meanings of the target words were given in a word list; there was no linguistic context provided for the target words and no opportunity to learn from context.

The two studies reviewed above, especially when taken together, do provide convincing data confirming the view that some parts of speech are easier to learn than others. They are, however, both concerned with the retention of L2 word meanings that are given to the participants. While using lists of word pairs remains a much practised basis for L2 vocabulary learning, learning from context or through consulting a dictionary is also widespread, and these approaches differ fundamentally from word pair learning in a number of respects. Perhaps most importantly, comprehension of an L2 word cannot be taken for granted when language learners meet the word in context or when they look up the word in a monolingual dictionary. Since word meaning retention depends on prior comprehension of the meaning, this factor will be crucial in determining learning rates of target words. While the figures provided through Rodgers' (1969) and Ellis and Beaton's (1993, 1995) studies may give an indication of expected learning rates per word class for L2 words that have been understood, comprehension rates for each word class may be very different. As for factors affecting comprehension rates, we have already seen in Chapters Four and Five how definition styles may affect comprehension for different parts of speech differently. Regarding how informative contextual clues may be for guessing meanings, Schatz and Baldwin (1986) have pointed out the unpredictability of contexts and Mondria and Wit-de Boer (1991) have considered how the amount of accessible context clues may affect word meaning retention.

In this review, we can see how our understanding of the relationship between word class and learning difficulty has developed in recent years. Rodgers' (1969) study presented, or was accepted as presenting (e.g., Nation, 1990: 48), a simple one-dimensional idea of intrinsic word class difficulty. Ellis and Beaton (1993, 1995) recognized that the difficulty of words from different word classes may depend on the learning strategy employed, and then went on to analyse the multiple causes that may be behind the relative ease or difficulty of learning for words from different word classes. When we consider other word learning environments, we will see how success in both inferring and learning meaning from context may also vary according to the part of speech. Further, levels of success for different parts of speech may also vary according to the type of linguistic context: whether, for example, this is the context of a sentence or the context of a definition in the target language. More specifically, we may even find that different types of definition favour the comprehension, use, or retention of different parts of speech in different ways.

6.2 The study

In the two studies reported in Chapters Four and Five, only verbs were focused on in one study and only adjectives in the other, and the main focus of the two studies was on the effect of dictionary definitions and example sentences on vocabulary comprehension, use, and retention. Although the effect of the two types of learning materials on comprehension and retention of words remains an important issue in this study, we will consider the issue here largely from the perspective of different parts of speech.

6.2.1 Hypotheses

The focus in this study on parts of speech is reflected in the following hypotheses which will be addressed in this chapter:

Regarding parts of speech

1) Regardless of learning materials used, comprehension levels will be higher for the set of verb target words than for the set of adjective target words.

2) There will be a greater difference between comprehension levels for the adjective target words and the verb target words for the Example Sentences group than for the Dictionary Definitions group.

3) For both groups, retention levels will be higher for the set of verb target words than for the adjective set of target words.

4) There will be a greater difference between retention levels for the adjective target words and the verb target words for the Example Sentences group than for the Dictionary Definitions group.

5) Interrater agreement levels for the participants' translation equivalents will be higher for verbs than for adjectives.

Regarding learning materials

6) The Dictionary Definitions group will be more successful than the Example Sentences group in understanding and giving translation equivalents for the target words.

I will now briefly explain the reasoning behind each of the above hypotheses. For 1), little relevant data exist other than the results of the studies reported in Chapters Four and Five. In these, comprehension rates were higher for the study focusing on verbs than on that for adjectives; this was true both for Dictionary groups and Example Sentences groups. The reasoning behind hypothesis 2) is that as inferring meaning from example sentences is generally less successful than using dictionary definitions, there will be more scope for the Example Sentences group to reflect differences relating to part of speech. For hypothesis 3), concerning retention of word knowledge, the anticipated higher comprehension rate for verbs should also lead to higher levels of retention for this part of speech. Hypothesis 4) is proposed following the same reasoning as the second hypothesis: that there will be more scope for variation within the Example Sentences scores. Hypothesis 5) is proposed because the meaning of adjectives often depends on the noun that it is modifying; as such, it may be harder for both raters and participants to pin down a clear meaning for many of the adjectives. This would result in wider variation for adjectives in raters' evaluations of the acceptability of participants' responses. For hypothesis 6), the far higher comprehension rates for the

Dictionary groups in Chapters Four and Five as compared to the Example Sentences groups leads us to expect the same outcome in this modified study.

One issue for which a hypothesis is not proposed is the effect of different definition types on comprehension and retention of target words. This was discussed in Chapters Four and Five and is an issue deserving of further research to build on that undertaken by Nesi (2000: 71-92) and Wingate (2002: 193-219). It will not be discussed further in this chapter since the variation among definitions for the two dictionaries used in this study, and between the two parts of speech under investigation, suggest that no worthwhile data will be available from this experiment. As we consider the above hypotheses, we will now go on to the study itself.

6.2.2 Method

I will now describe how this study was conducted. Following a pretest to ensure that target words were unknown, a group of Japanese intermediate learners of English were given a set of 40 unknown words: 20 verbs and 20 adjectives. Together with the target words, one group (the Dictionary Definitions group) were given a set of monolingual dictionary definitions for the target words. The other group (the Example Sentences group) were given a set of authentic example sentences for the words. More information about these materials is provided below. The participants were instructed to study the materials and write Japanese equivalents for each of the target words. Three weeks later, the participants were given a test of vocabulary retention, a kind of gap-fill exercise in which they had to match the sentences or definitions with the correct word. This experiment differed in a number of ways from the two studies described in Chapters

Four and Five. These are explained, together with reasons, below.

a) Participants in this study were not required to produce their own sentences, as in the previous studies, but only to give translation equivalents for the target words. This was partly because of the persistent problem of low inter-rater reliability rates for the evaluation of the English sentences and partly because of the time-consuming nature of the sentence-writing activity.
b) As there was no sentence-writing part to the test, the total number of target words could be increased from 20 to 40.
c) The set of 40 target words was composed of a subset of 20 verbs and a subset of 20 adjectives.
d) The multiple-choice format used in previous studies was replaced with a test in which the effect of random guessing on scores would be reduced.

6.2.3 Participants

A total of 72 Japanese university students were recruited as participants for this experiment. The participants were first year (N = 27) and second year (N = 45) university students, aged between 18 and 20, majoring in English at a private middle-ranking Japanese university. They had all received between six and seven years of formal instruction in English. Despite this, they were generally intermediate level students and their TOEFL scores were estimated to range between 380 and 510. A smaller group of 57 of the participants were available to take part in the retention test which was held three weeks later.

6.2.4 Target words

There were 40 target words, 20 verbs and 20 adjectives, all unknown to the participants. In the studies described in Chapters Four and Five there were a total of 24 and 20 target words, respectively. The words in this study were largely the same as the ones used in those experiments, and were selected according to the following criteria:

1. To exclude words that were too likely to be known, only words in the Cobuild 'two diamond' 'one diamond' or 'no diamond' frequency bands (Sinclair et al., 1995) were included; words ranked 3,400 and above in terms of frequency. In addition, only words not in the first seven levels of the JACET 8000 list (see Mochizuki, 2003) were included.
2. Basically, any words with more than one sense identified in the dictionaries used were excluded; this was to avoid confusion for participants in the experiment and for raters of the translation equivalents.
3. Only words for which there were at least 25 occurrences in the *Cobuild Direct* 50 million word corpus were included. This condition was set for three reasons:

    i) To ensure that no especially infrequent words were included as target words,

    ii) To ensure that typical syntactic patterns and collocations of the target words could be identified,

    iii) To have a sufficiently large pool of example sentences from which to select three typical sentences for each target word.

4. A list of 80 words conforming to 1, 2 and 3 above was presented to the participants as a pre-test, with words correctly identified by any participants or frequently mistaken for other words being removed from the list.
5. The 40 target words, 20 verbs and 20 adjectives, were selected from the remaining unknown words.

The number of target words was set at 40 to allow enough time in a 90-minute class period for the participants to complete the tasks as required. This is still, however, a larger set of words than in most other studies in the field. The final 40 target words are as follows:

Verbs:

abbreviate, amputate, appal, blab, bode, cackle, chomp, coerce, dilate, elope, feign, jilt, perspire, pooh-pooh, raze, sulk, suss, trounce, waft, whinge

Adjectives:

afoot, *akin*, *averse*, *bereft*, *blatant*, *callous*, *colossal*, *defunct*, *dilapidated*, *eerie*, *fleeting*, *furtive*, *galore*, *gaudy*, *hoarse*, *illicit*, *morbid*, *obese*, *ostensible*, *poignant*

6.2.5 Learning materials

As described above, the experiment required the participants to be divided into two groups and use one or other of two different resources for the target words. The participants were divided randomly into the Dictionary Definitions group and the

Example Sentences group. After data for participants absent from the retention test part of the experiment were excluded from the study, there were 35 participants remaining in the Dictionary Definitions group and 37 participants in the Example Sentences group.

Dictionary Definitions group

Participants in this group received two dictionary entries for each of the target words, taken from the *Longman Dictionary of Contemporary English, 3rd Edition* (Summers et al., 1995) (LDOCE3) and the *Collins COBUILD English Dictionary, 2nd Edition* (Sinclair et al., 1995) (COB2). These are found in Appendix 6.1. The dictionary entries were stripped of any example sentences and grammatical information so that only the definitions for the words were provided. The reason for this was to focus specifically on the effect of dictionary definitions and contrast these with example sentences. There is a valid argument for using whole dictionary entries, given that this is what dictionary users usually see when using a dictionary. In this study, however, it was felt that the composite nature of monolingual learner dictionary entries, containing definitions, example sentences, grammatical and other information, would have rendered it impossible to know which elements within the entries might have contributed to the participants' success or failure in understanding and learning the target words. A further justification for this decision may be found in the format of many popular hand-held electronic dictionaries in which, initially, only a word's definitions are shown on the small screen; the example sentences are only displayed when requested by the user (Koyama and Takeuchi, 2004).

Example Sentences group

This group received three example sentences for each of the target words, drawn from the 50 million word COBUILD Direct corpus, as shown in Appendix 6.2. Sentences were chosen to display typical syntactic patterns and collocations, and for their comprehensibility. Wherever possible, sentences were taken directly from the corpus without changing them in any way. In a few cases, however, parts of sentences were deleted if they were too long but otherwise ideal. Where more than one grammatical pattern occurs frequently, the choice of examples reflect this.

There have been various approaches to types and sources of example sentences used in dictionary entries. Basically, two types or sources of example sentence have been used in monolingual learner dictionary entries over the past few decades (Rundell, 2006). One is purely made-up examples, written to demonstrate the use and meaning of the word. This approach has usually resulted in what Mondria and Wit-de Boer (1991) describe as pregnant contexts for the headwords; not reflecting natural use but giving many contextual clues as to the meaning of the words. A contrasting approach has been to use examples sentences taken directly from a corpus of the language; these aim to illustrate typical use but may be difficult to understand. In addition to these, a third approach has been to use corpus data to inform example sentence writers about typicality; these may be easier than authentic corpus-drawn sentences but lack much of the "pregnancy" of purely made-up example sentences.

The above arguments concerning the value of the various types of example sentences for learner dictionaries provide a context for considering the comprehensibility of isolated sentences in a language learning context. We should bear in mind, however, that the arguments are not directly relevant to this experiment. The example sentences used in this experiment are not intended to be seen as a part of a dictionary entry (although the findings might have implications for dictionary design). Rather, the purpose of the example sentences is to create a learning environment for unknown words similar to that experienced by language learners through multiple encounters with unknown words through L2 reading.

6.2.6 Vocabulary retention test

A test of vocabulary retention was conducted three weeks after the main vocabulary learning session in which the above materials were used. For this test, the 57 participants were given an answer sheet (Appendix 6.3), together with the same materials as three weeks before, except that the target words were deleted from the materials, leaving gaps in their place. The forty sets of definitions or example sentences were divided into verbs and adjectives and, within their respective part of speech, randomly reordered and put into ten-word sets of materials. See Appendix 6.4 for examples for the two sets of materials for each part of speech. For each item, participants had to choose the matching answer from a set of ten target words of the same word class.

The aim in changing the format of the test from that used in Chapters Four and Five was to reduce the effect that random guessing might have on scores while, at the same time, retaining the sensitivity of the instrument in recording partial word knowledge.

6.2.7 Rating

Two sets of highly proficient Japanese users of English rated the participants' translation equivalents for the target words. To help the raters, dictionary entries from two monolingual English and one English-Japanese dictionary were provided for each target word. The dictionaries were ones not used in the experiment itself: *Oxford Advanced Learners Dictionary, 5th Edition* (Crowther et al., 1995), *Cambridge International Dictionary of English* (Procter et al., 1995) and *Kenkyusha English-Japanese Dictionary for the General Reader, 3rd edition*, (Matsuda, 1999). Rating guidelines (see Appendix 5.6) were provided for the raters.

The raters of the Japanese translation equivalents of the target words rated each translation as Correct, Partially Correct, or Wrong. The target word *cackle* provides a good illustration of how difficult it is at times to judge the accuracy of translation equivalents. If we think of possible partial synonyms of *cackle*, such as *laugh*, *sneer*, *laugh loudly*, *make fun of*, or *laugh at*, we can begin to appreciate the difficulty of the raters' task.

The interrater agreement level overall was .80. Agreement for translation equivalents for verbs was .78, while for adjectives it was slightly higher at .82. A t-test conducted on the number of rater disagreements for the two parts of speech confirmed that these differences were not significant. Differences were resolved at a joint meeting of the two raters.

6.3 Results

There were two main sets of results from the experiment: for the participants' Japanese translation equivalents of the target words, and for the retention test. These results were analysed both in terms of the two experimental groups' scores and according to the part of speech of the target words. These results will be presented one by one then discussed both individually and in relation to each other.

6.3.1 Translation equivalents

The Japanese translation equivalents were rated as follows: Correct –matching the English word; Partially Correct – partially matching the English word; or Wrong – not matching the English word at all. In the few cases where no answer was given, this was added to the last category. The Correct and Partially Correct equivalents were collapsed into one category as 'acceptable equivalents'. There was a large difference between the Dictionary Definitions group and the Example Sentences group for their production of acceptable translation equivalents for the target words. A t-test conducted on the ratings of the Japanese translation equivalents for the two groups confirmed that this difference is significant ($t = 10.36$, $p < .001$). The results are shown in Table 6.2.

Table 6.2
Japanese translation equivalents for target words (maximum = 40)

	M	S.D.
Dictionary Definitions group (N = 35)	20.4	5.51
Example Sentences group (N = 37)	9.00	3.56

6.3.2 Part of speech

The translation equivalents were then considered from the perspective of part of speech; was it easier, as suggested in the hypothesis, to produce translation equivalents for verbs than for adjectives? As Table 6.3 shows, there was an overall difference of about 8% between the mean numbers of acceptable answers in favour of the verbs. A one-way ANOVA confirmed that, overall, participants from the two groups produced a significantly higher number of acceptable answers for verbs than for adjectives (t = 2.04, p<0.044).

Table 6.3

Acceptable translation equivalents of verb and adjective target items (N = 72)

	M	S.D.
Verbs (20 items)	7.94	3.98
Adjectives (20 items)	6.60	3.96

Overall, it was easier for the participants to give accurate translation equivalents for the verbs than for the adjectives. We now need to ask whether this was true for each group, and to consider in what other respects the two groups' results may differ from each other. First, the figures for the Dictionary Definitions group are shown in Table 6.4.

Table 6.4

Acceptable translation equivalents of verbs and adjectives: Dictionary Definitions group (N = 35)

	M	S.D.
Verbs (20 items)	10.91	3.23
Adjectives (20 items)	9.49	3.46

A one-way ANOVA conducted on these figures showed that the Dictionary Definitions group participants did not give significantly more accurate equivalents for verbs than adjectives (t = 2.04, p<.079).

The figures for the Example Sentences group are provided in Table 6.5. The ANOVA showed that the difference between scores for verbs and adjectives is significant for the Example Sentences group (t = 2.65, p<.01).

Table 6.5
Acceptable translation equivalents of verbs and adjectives: Example Sentences group (N = 37)

	M	S.D.
Verbs (20 items)	5.14	2.18
Adjectives (20 items)	3.86	1.95

As we review the figures for acceptable translation equivalents for the target words, the main difference between the two groups is that the Dictionary Definitions group achieved more than twice as many acceptable translation equivalents as the Example Sentences group. This is true for the whole set of target words and for the subsets of verbs and adjectives too. As regards scores for verbs and adjectives, participants in both groups were, on average, able to give more acceptable equivalents for verbs than for adjectives, although this difference was only significant for the Example Sentences group. As for the difference between numbers or proportions of translation equivalents for the subsets of verbs and adjectives, these are very similar for the two groups. In both cases, about 55% of the acceptable equivalents were for verbs and around 45% for adjectives. As we look at the scores in more detail below, we will possible factors between these similarities and differences.

6.3.3 Retention test

The results for the retention test are shown below in Table 6.6. They indicate that there was a small difference between the two groups' overall scores for the 40 target words in favour of the Example Sentences group. However, a one-way ANOVA conducted on the participants' scores showed that this group did not score significantly higher than the Dictionary Definitions group ($t = 1.63$, $p < 0.10$). It should, though, be pointed out that there were various problems with the administration of the retention test which cast some doubt on the validity of both between-group comparisons and comparisons between verbs and adjectives.

Table 6.6
Retention test scores according to learning materials used (40 items)

	M	S.D.
Dictionary Definitions group (N = 27)	4.81	2.40
Example Sentences group (N = 30)	6.10	3.42

We will now go on to look at the retention test scores for the sets of adjectives and verbs for the two groups, as shown in Tables 6.7 and 6.8. For the Dictionary Definitions group, the average retention rate for the set of adjectives is higher than for the set of verbs, but an ANOVA showed that this difference is not significant ($t = 0.86$, $p<0.39$). It is worth noting that scores for both sets of target words are very low, with that for verbs close to that which would typically be obtained through blind guessing of answers.

Table 6.7

Adjective and verb retention test scores: Dictionary Definitions group (N = 27)

	M	S.D.
Adjectives (20 items)	2.59	1.78
Verbs (20 items)	2.22	1.34

For the Example Sentences group, average retention rates for the set of adjectives was higher than for the set of verbs. As can be seen in Table 6.8, however, standard deviations are relatively high, and an ANOVA (t = 0.86, p<0.39) showed that the difference between scores for the two sets of words is not significant. We will now go on to consider these and the other results from this experiment.

Table 6.8

Adjective and verb retention test scores: Example Sentences group (N = 30)

	M	S.D.
Adjectives (20 items)	3.27	2.64
Verbs (20 items)	2.77	1.81

6.4 Discussion

We will now return to the hypotheses proposed regarding this experiment. We will begin with the issue of whether dictionary definitions or example sentences are of most help in understanding unknown English words. After this we will focus on the main topic of this chapter: the relationship between word comprehension or retention and part of speech. In relation to both of these issues, we will look at the actual materials provided for different target words and the type of responses they typically elicited. We will then go on to the retention test: the problems with the administration of the test and the effect of these on the validity of the test results. We will also evaluate the other instruments used, the effect of the specific learning materials, and issues relating to language learner proficiency levels and skills.

6.4.1 Comprehension and learning materials

As in the research reported in Chapters Four and Five, there is a very large difference between the Dictionary Definitions group and the Example Sentences group in the average numbers of acceptable translation equivalents produced for the target words. In fact, the Dictionary Definitions group, with an average of just over 20 acceptable responses per participant, were more than twice as successful as the Example Sentences group, with an average of 9 acceptable responses per participant. The difference is further underlined if we look at proportions of matches and partial matches for the translation equivalents for the two sets of materials. Partially matching equivalents account for well over half of the Example Sentences group's acceptable responses (55.7%) while the Dictionary Definitions group's figure for partially matching equivalents is less than half (46.4%).

These scores presented in terms of averages do, however, mask considerable differences both within each of the two groups and among the 40 target words. In previous chapters we have discussed the differences between the materials in general terms and the participants' scores largely in terms of averages. In this chapter we will see what we can learn both by looking at the ranges of acceptable equivalents among participants and at the range of numbers of acceptable equivalents given for individual words and the various materials and responses for individual words. In these contexts, we will look at the sets of materials for individual words and the answers that were produced in response to these.

As shown in the tables in Appendix 6.5, for 29 of the 40 target words, the dictionary definitions were a more helpful resource for more participants than were the example sentences. For seven of the target words, there was little or no difference between participants' scores in the two groups, while for the remaining four target words, more participants in the group using example sentences were correct than in the group using dictionary definitions. Although average scores might lead us to believe otherwise, it would be inaccurate to say that the dictionary definitions were always a more reliable source of information for the target words than the example sentences.

Regarding acceptable equivalents for participants, in the Dictionary Definitions group scores ranged between 8 and 29 acceptable responses, while Example Sentences group participants' scores ranged between 2 and 17 acceptable responses. Most Dictionary Definitions group participants scored more than any of the Example Sentences group, but the highest scoring Example Sentences group participant achieved a higher score than eight of the Dictionary Definitions group participants. Again, we cannot say that the dictionary definitions were always the most informative or accessible source of information for participants about target word meanings. However, it is likely that those participants from the Example Sentences group who achieved high scores would have achieved even higher scores had they been given the dictionary definitions instead.

While we are looking at ranges of participants' scores for the two sets of materials, it is also worth considering what the two the results of the highest scoring Dictionary Definitions group participants may tell us about the materials and the participants. The best-scoring students in this group managed to provide acceptable translation equivalents for just under three quarters of the target words. Whether explicitly stated in the title of the dictionary or not, full-size monolingual English learner dictionaries are generally seen as dictionaries suitable for advanced learners of English. With estimated TOEFL scores ranging between 380 and 510, most of the participants in this study might be regarded as intermediate level learners, with only perhaps the top 10% really being advanced level learners of English. Yet even the top four or five participants in this group, those with TOEFL scores around 500, were unable to produce acceptable translation equivalents for over a quarter of the target words.

The comprehension rates of the Dictionary Definitions group participants in this study led us to three possible conclusions regarding the definitions, the participants, and the task itself:

a) Many of the definitions were either incomprehensible or were too imprecise;

b) L1 equivalents for the concepts of the target words were unknown to the participants;

c) The participants' task of providing L1 equivalents for the target words was somehow unreasonable.

We will consider a) in more detail as we look at materials for individual words below. It is worth pointing out, however, that the level of definition difficulty will not be uniform within a given dictionary; for example, less frequent or harder words will often have harder definitions, the assumption being that someone looking up a rare word should be familiar with the less rare words used in the definition. For b), this issue is concerned both with the limitations of participants' knowledge about their mother tongue and with cultural differences, since the presence of a familiar and widely used term in Japanese may greatly assist learners to guess at the meaning of a target word. Both of these are undoubtedly factors, for example, contributing to the verb *elope* being the most highly scoring of the 40 target words for the Dictionary Definitions group. As for c), especially for words for which there are no familiar word-for-word equivalents, the request for participants to provide L1 equivalents may have been an unreasonable task if they felt that they had to provide single word answers as equivalents for the target words. Some apparently did feel this and answered accordingly while others adopted the strategy of explaining some of the target word meanings in short phrases. More specific guidance for participants in future studies may largely address this issue.

For the participants in the bottom 10% of both groups, with scores of just two to four acceptable answers for the Example Sentences group and scores of eight to twelve for the Dictionary Definitions group, other issues are clearly involved. The English proficiency levels of these participants would, generally, be lower than average but such differences cannot be simply attributed to differences in language ability alone. The university entrance examination system in Japan would usually determine that students accepted to a given department of a particular university are of a largely similar level; if they were much better than average, they would have chosen a better university, while if they were much worse, they would not have been able to enter this one. Rather, these differences seem to point to a specific ability that these tests require of the participants: the ability to make reasonable guesses based on limited information.

It is true that a larger vocabulary would render definitions or example sentences more comprehensible, and so the task of guessing easier, but in many cases the participants seemed to have adopted one of two stances regarding the guessing of meanings. These were either "It's like a jigsaw puzzle: some pieces are missing but with what I know I can try to guess what the picture is" or "Without all the clues, it's impossible to guess the correct answer. It's a pointless exercise". This was not only evident in the answers given but also in the time taken to complete the test; participants with the lowest scores generally took all the time available, spent some of the time pretending to sleep, or refused to complete the test. This was especially evident among the Example Sentence participants, understandable since, as we shall see below, for many of the target words there was insufficient information for any of the participants to guess the meanings of the words.

This last point brings us to consider the scores and situation of the Example Sentences participants with the highest scores, in the area of 15 – 17 acceptable equivalents for the 40 target words. As the scores show, these participants clearly did not see guessing the meaning of the words as an impossible task, and they were successful in providing acceptable equivalents for well over a third of the target words. In some respects, the task the participants in this group faced may be viewed as a kind of productive multiple choice test. The information they could glean from the sentences gave them an idea of a number of words or meanings that could fit the contexts of the sentences; from these possibilities they would have to select the meaning they felt would be most likely to be the right answer. The best scoring participants' relatively high scores may be attributed to four related factors: a) their willingness to respond to this challenge; b) their ability to pick up a large number of clues from the context; c) their skill in inferring possible meanings; and d) their good judgment in assessing which of the possible meanings was most likely.

Nation (2001: 247) suggests that L2 proficiency is a major factor in successful guessing. That, undoubtedly, is true in one respect; as mentioned above, the higher the level of proficiency or the larger the L2 vocabulary, the more clues are available to the participant. This is reflected in the answers of the highest scoring participants; the participants that were expected to get high scores generally did. In the case of the lowest scoring participants, however, there had been no indication until this point that they were especially weak in terms of L2 proficiency. Further tests, possibly with interviews into language use outside class, might be necessary to identify what makes L2 learners able, or unable, to make intelligent guesses.

Many activities involving language, both in the L1 and the L2, require intelligent guessing with limited information: whether watching a film, reading a book, or taking part in a conversation. Of these, only reading, however, allows time for reflection on the linguistic context to guess at unknown words or information. Generally, most people in the language group under investigation have substantial experience of watching films or television and joining conversations in their L1, while the amount of L1 reading will vary widely from person to person, even among university students. As for the L2, there will be wide variation among the participants in their experience of listening to English, whether through conversations or film and television, and in their experience of extensive reading in English. It is proposed that these, especially extensive reading, may be crucial factors determining whether L2 learners are good guessers or poor guessers. While Nation does consider the causes of poor guessing (2001: 246-7), such as insufficient context, what may be more pertinent is not what causes successful or poor guessing but what makes good or poor guessers. The research here does point to some possible answers but clearly more research in this field is still needed.

As we look below at some sets of example sentences and the meanings guessed for the target words for these sentences, we will consider further what factors in language students and in L2 contexts may assist or hinder accurate guessing of word meaning.

6.4.2 Comprehension and part of speech

The first hypothesis proposed that participants would be more often able to provide accurate translation equivalents for the set of verbs than for the adjectives. The results of the study confirmed this; participants overall scored around 20% more for verbs than for adjectives. The advantage for verbs was also apparent in the independent results of the Dictionary Definitions group and the Example Sentences group.

We made this hypothesis largely on the basis of test results from the experiments described in Chapters Four and Five in which scores for the set of verbs were markedly higher than for the set of adjectives. The results of this study have confirmed the hypothesis, but now we need to go beyond these results and ask what factors made it easier for the participants to produce translation equivalents for the set of verbs than for the set of adjectives. We begin to do this in two related ways:

i) By looking at how results for the two word classes differ for the two groups of participants;

ii) By identifying types of words that are easier or harder for participants according to the word class and the learning materials they were given.

After these two preliminary stages we will continue by investigating the materials for individual words from the types identified in ii), and considering how to account for the participants' success or failure in comprehending the meaning of the target word. After this, we will return to the question of how definitions or example sentences for different parts of speech may affect comprehension of unknown words.

i) As we investigate how results for the two word classes differ for the two groups of participants, it is worth noting that there was little evidence to support the second hypothesis, that the difference between results for the two parts of speech would be greater for the Example Sentences group than for the Dictionary Definitions group. Numbers of acceptable equivalents differ widely between the two groups but in terms of proportions of equivalents for verbs and adjectives, the two groups were largely the same. For the Dictionary Definitions group, while the least accessible 12 words are divided equally between verbs and adjectives, of the most accessible 12 words seven are verbs and five adjectives. For the Example Sentences group, too, while the least accessible 14 words are also divided equally between verbs and adjectives, of the most accessible 10 words seven are verbs and three adjectives.

ii) The description of the data in i) might lead us to have a one-dimensional picture of word difficulty; we expect the proportion of participants able to work out the meaning of a target word to be different for the two groups and we expect this difference to be largely uniform from word to word. This would mean, for example, that if for a particular word 50% of Dictionary Definitions group participants and 25% of Example Sentence group participants gave acceptable equivalents, we would expect this ratio to be true for other target words. Looking at the data for individual words (Appendix 6.5), we can see that this is not the case. There is wide variation from word to word, and the Dictionary Definitions group did not produce the greatest number of acceptable equivalents for every target word. In fact, from the perspective of meaning comprehension rates for the two sets of materials, we can consider the target words as belonging more or less to one of the four word types:

A) Words which are hard to understand regardless of the materials provided (e.g. *morbid, poignant*, *appal*);

B) Words which are relatively easy to understand with both sets of materials (*suss*, *chomp*, *obese*);

C) Words which are markedly harder to understand with the example sentences than with the definitions (e.g. *blab*, *defunct*, *coerce*);

D) Words which are markedly harder to understand with the definitions than with the example sentences (e.g. *blatant*, *afoot*, *waft*).

As we can see, both word classes under investigation are represented in all four word types. This alone challenges the simple idea of uniform intrinsic word class difficulty, at least in terms of meaning comprehension. As we go on to look at representatives of these four word types in more detail we may gain a clearer understanding of what may constitute word difficulty in this respect and how word difficulty may be said to vary from word class to word class.

We will now look at the materials given to the participants for some of the words from the word types listed above. We will look at one word from each word class for each of the four types A-D listed above. We will start with A: words which appeared to be difficult both for Dictionary Definitions group participants and for Example Sentence group participants. For the adjective *morbid*, there were no participants, from either group, who were able to give an accurate Japanese equivalent. The materials the Dictionary Definitions group had for this word were as follows:

> If you describe a person or their interest in something as **morbid**, you mean that they are very interested in unpleasant things, especially death, and you find this strange or unwise. (COB2)
>
> Having a strong and unhealthy interest in unpleasant subjects, especially death. (LDOCE3)

The first definition is a long single sentence of 31 words stretching over four lines of the dictionary column. It contains four verbs (*describe*, *mean*, *are*, *find*) and four adjectives (*interested*, *unpleasant*, *strange*, *unwise*), some of which may seem contradictory to the learner reader. The second definition, if participants went on to read this, is much shorter but starts with a gerund (whereas Japanese sentences are usually verb-finite), contains two adjectives that may seem contradictory (*strong*, *unhealthy*), and has a referent which is polysemous (*subjects*). Most responses suggested something undesirable, with *bad* or *evil habits*, *unhappy* and *unlucky* being the most common. The "Kidrule" phenomenon (Miller and Gildea, 1985) was also in evidence in some responses, with the responses *be interested* and *be entertained* indicating that some participants mistook the words *interest* or *interested* used in the definitions for the meaning of the word itself.

For example sentences, the situation is in some respects different, in that while a dictionary definition may be called successful or otherwise in the degree to which users understand it, sentences other than those in dictionaries or language teaching materials do not usually serve this overt purpose. The Example Sentences group participants received the following sentences for the word *morbid*:

> More onlookers might have been expected, if only out of morbid curiosity.
> You should get away from all these morbid imaginings.
> He is not one to get morbid.

For someone unfamiliar with the word *morbid*, there are few clues here as to a possible meaning. The vocabulary and structures of the sentences are perhaps easier than the dictionary definitions but the very wide range of answers, including the Japanese words おかしい ("okashii": *strange*), ばかな ("bakana": *stupid*), あぶない ("abunai": *dangerous*), and 大勢 ("oozei": *numerous)*, reflects how little the participants could infer about the word's meaning. Many, perhaps guided by the middle sentence, gauged or guessed that *morbid* means something negative, while others apparently settled for an adjective that can be used in front of *curiosity*. None, however, came close to providing an accurate equivalent.

The situation was similar for the verb *appal*, for which very few participants from either group correctly guessed the meaning. The definitions for this word are not, at first glance, difficult to understand:

> To shock someone by being very bad or unpleasant. (LDOCE3)
>
> If something **appals** you, it disgusts you because it seems so bad or unpleasant. (COB2)

Most participants in this group identified this word as being something bad, but there was a wide range of interpretations: participants gave the following responses: うわさをする ("uwasa o suru": *gossip*) がっかりさせる ("gakkari saseru": *disappoint someone*), 怒らせる ("okoraseru": *annoy*) 悲しい、うつの気持ちを持つ ("kanashii, utsu no kimochi o motsu": *feel sad or depressed*), etc. Perhaps here it may have been the fairly complex syntax of the first definition which confused participants, while the central use of the not widely known verb *disgusts* in the second definition may also have been an obstacle to comprehension.

The sentences provided to Example Sentences group participants for *appal* are as follows:

> He said he was appalled at the way animals were treated.
> Mr Peters said he had been appalled by the conditions in which the housemaids live.
> His ignorance appals me.

Again, a large number of participants' equivalents identify something bad but without quite being able to get to the real meaning, and give: meanings equivalent to *make angry*, *make unpleasant*, *go bad*. For many other participants, however, there again seems to be a variant of Miller and Gildea's (1985) "Kidrule" operating, except with contexts rather than definitions. Example Sentence participants who provided Japanese translation equivalents for *appal* with meanings such as けがをする ("kega o suru": *injure)*, 痛む ("itamu": *hurt*), 人の世話をする ("hito no sewa o suru": *be cared for*) 虐待させる("gyakutai saseru": *be ill-treated*), or 奴隷にする("dorei ni suru": *enslave*) all seem to take the treatment of animals or housemaids described in the sentences as the meaning of *appal*, rather than *appal* being the hearer's response to this treatment. There is no easy answer to the question as to whether these words are intrinsically difficult or whether the particular sets of materials were difficult, but we can see how the example sentences and definitions in these cases do little to aid participant's comprehension. We will return to this question after looking at more word types.

We will now go on to type B words for which relatively large numbers of participants from both groups were able to produce accurate translation equivalents. We will start with the verb *chomp*, for which 28 Dictionary Definition group participants and 21 Example Sentences group participants gave accurate equivalents.

We will begin by looking at the definitions provided for *chomp*:

> To bite food noisily. (LDOCE3)

> If a person or animal **chomps** their way through food
> or chomps on food, they chew it noisily; an informal use. (COB2)

In contrast to the entry for *morbid*, the first definition is short and clear: only four words. The second definition, from the perspective of meaning, does include redundancies in its "…chomps their way through food or chomps on food…" but, as Wingate (2002: 43) notes, this redundancy may well help to reinforce meaning rather than obscure it.

The Example Sentences group were provided with the following sentences for *chomp*:

> I chomped hungrily through the large steak.
> Miguel chomped on his fresh stick of gum.
> He took the ice from his cup and began chomping on it.

Here, the human subjects of the verb and the edible, and familiar, objects – *steak*, *gum*, *ice* – may account for the fact that more than half of the participants in this group provided accurate equivalents.

One adjective that was widely understood by both groups was *obese*. For this, the definitions provided were as follows:

> (Technical,) Very fat in a way that is unhealthy. (LDOCE3)

> If someone is **obese**, they are extremely
> overweight or extremely fat. (COB2)

This time, the first two words of the first definition are sufficient to convey the main meaning of the word, with the rest of the definition just adding to this meaning. The second definition is longer but still simple and clear.

The example sentences provided for the target word *obese* were as follows:

> Fasts and very low-calorie diets do not work for obese people; they put the weight back on.
> In 1980, 6 per cent of men aged 16 – 64 were obese.
> When the filming was over, I asked the producer if I looked obese.

Here, there are few words that the participants would not know (perhaps only *fasts*) but variously widely known words that do inform the reader as to the meaning of the word *obese*: *low-calorie*, *diets*, *weight*, *looked.* Furthermore, most of these are found in the first sentence.

In these two cases examined, it is perhaps easier to affirm that the words *chomp* and *obese* are intrinsically easy to understand. After all, just one word – *eat* or *fat*, respectively – would be enough for a partially correct response for these words. The simple short dictionary definitions confirm this. However, this evaluation may not apply to guessing meaning from context. All three sentences for *chomp* are simple and contain easy to understand clues, but for *obese*, two of the sentences are not informative. If these were presented alone, or even first, participants may have been much less successful in guessing the meaning.

We will now go on to words for which participants in one group were better able to give equivalents than the other group. The verb *blab*, a type C word, was widely understood by Dictionary Definition group participants but by none of the Example Sentences group at all. We can start by looking at the definitions:

> To tell secret information to someone who is not
> supposed to know it. (LDOCE3)

> If someone **blabs** about something secret, they tell people
> about it; an informal word. (COB2)

Again, the main meaning is conveyed very early on, in the first four words of the first definition. This is confirmed in the rest of that definition and in the second definition. The example sentences for *blab* were as follows:

> Her mistake was to blab about their affair.
> But one of the gang has blabbed to police, starting a violent witch-hunt to find the traitor.
> She'll start blabbing it out to the whole class.

The most common equivalents given were つかまえる ("tsukamaeru") meaning *to be caught*, and みつかる ("mitsukaru") meaning *to find out* or *discover*. Other equivalents that participants provided include 気にしない ("ki ni shinai": *to not mind or worry*), 告白する ("kokuhakusuru": *to confess*), or 逃げる("nigeru"*: to run away*). All of these proposed equivalents point to the phrase 'about their affair' in the first sentence, and all are feasible in this context. This reveals another problem that language learners experience both in guessing from context and in reading monolingual definitions; if language learners guess a word's meaning from the first sentence or the first part of a definition, they may take this guess to be the true meaning rather than a hypothesis to be confirmed or rejected by subsequent information in sentences or definitions.

Waft is one word for which more Example Sentences group participants gave equivalents than Dictionary Definition group participants. The example sentences for this type D word are as follow:

> A faint, very aromatic scent of fire smoke wafted towards us.
>
> Soft, romantic music wafted through the luxurious hotel suite.
>
> She complained after smoke from Mr Legg's cigarettes wafted into her house.

These are the two definitions for *waft*:

> To move gently through the air. (LDOCE3)
>
> If sounds, scents, or smoke **waft** through the air,
> or if something such as a light wind wafts them,
> they move gently through the air. (COB2)

Here, the short simple style of the first definition is relatively unsuccessful in conveying the meaning of *waft*. Many Dictionary Definition group participants gave answers relating to movement: とりぬけう("torinukeru": *go through*, *penetrate*) 踊る ("odoru": *dance*), なめらかに流れること("namerakani nagareru koto": *flow smoothly through*). Others gave answers related to sounds or to air: 風がそよぐ("kaze ga soyogu": *the wind rustles*), 排気ガス("haiki gasu": *exhaust fumes*), 響く("hibiku": *resound*), 風("kaze": *wind*) but few participants combined these two elements of the word's meaning. The problem may be partly that the word is complex in some respects; as the example sentences illustrate, only certain types of thing (*scent*, *music*, *smoke*) waft, they always waft somewhere, and they only waft in a certain way – slowly, gently or softly. For large numbers of Example Sentences group participants, however, this complexity was not an obstacle to comprehending the meaning of the word and conveying it in their L1.

After looking at quite a few sets of materials for specific words, together with the type of responses they produced, we may suggest some initial conclusions regarding comprehension difficulty of words looked up in MLDs. Definition length does seem to be both an important indicator and cause of word meaning difficulty At least two other factors may affect comprehension: frequency and currency of the concept in the learner's L1; and the accessibility of the syntax and vocabulary of the definition,

For encounters with unknown words in context, different factors are involved: the pregnancy of the context; the comprehensibility of the context; the lack of distracters leading to wrong conclusions; and the clarity or lack of ambiguity in the contexts. As we have seen with the words *waft* or *chomp*, example sentences may occasionally provide a vehicle for comprehension that is as good or better than definitions. In general, though, as other research has confirmed, guessing from context is a much less reliable medium for comprehension than consulting a dictionary.

The means of comprehension sometimes appears quite similar for verbs and adjectives. With *morbid* and *blab*, for example, the challenge faced by participants was to find a meaning that fitted both syntactically and lexically. For these words, their respective sets of contexts were equally unhelpful. In other cases, however, the contexts of verbs may typically provide more clues as to meaning than adjectives. For transitive verbs such as *chomp*, we are given information about both the subject and object of the verb. For verbs used with adverbs, such as *waft*, for which example sentences contain the 'what', 'where', and often the 'how' of the verb, it may be easier to guess the meaning than for adjectives for which few clues are available.

Dictionary definitions vary widely from word to word, as indicated by the variable success in understanding definitions, and the varying length of definitions. Adjectives may be easier when providing a straightforward description of something, but where pragmatic elements such as user attitude or emphasis are contained in the meaning of the word, these may be difficult to convey and difficult to comprehend. For these, from the perspective of defining style, there may be an argument for keeping the pragmatic element separate from the 'meaning' element: putting it in the extra column if in a COBUILD format or expressing it as a symbol or at the end of the definition.

6.4.3 Retention test

The retention test scores for both groups and both parts of speech are generally low, as found in many other studies of vocabulary retention. Apart from low scores, however, the results of the retention test were not as were expected. The Example Sentences group, which performed much worse than the Dictionary Definitions group in the test which required them to give translation equivalents for the target words, performed significantly better in this retention test. The scores for adjectives in the retention test are higher than those for verbs, while in the comprehension test participants were better able to give equivalents for verbs. These unexpected results may be due, in part, to two specific errors in the conducting of the retention test. We will first address these before going on to consider what we may learn from the retention test results.

The first problem with the administration of the retention test in this study is that not all the participants were able to complete the test in the time allowed. A consequence of this is that, for both groups, scores for the final set of ten verbs were much lower than for the other three sets of words. For the Example Sentences group, participants achieved a combined score of 35 for the final set of verbs as compared with 64, 56 and 53 for the other three sets of words, while for the Dictionary Definitions group, participants' scores for the final set were 26 as compared with 45, 39 and 42 for the other three sets. Scores for the first set of verbs are very close to scores for the two sets of adjectives, which suggests that insufficient time to complete the test may be the main cause behind the difference in retention test scores for the two parts of speech.

The second fault with this test is that some gaps in the definitions and sentences for adjectives are preceded by the word *an*, indicating that only a target word starting with a vowel will fit the gap. This occurred with the example sentences for two target words and with the definitions for one word. While scores for these items are not higher than average, suggesting that few participants noticed this, this may have given a slight advantage to the Example Sentences participants.

This second problem alone is unlikely to account for the difference in retention scores for the two groups. Two or three other factors are likely to be significant contributors to the higher scores for the Example Sentences group:

a) For the Example Sentences group, the target words were always presented to the participants in the context of the sentences, so for them the nature of the retention test was to recall which word had been in the sentences in the spaces now represented by gaps. For the Dictionary Definitions group, although over half of the definitions did contain the target words, most of the shorter definitions, which participants may have focused on, did not. This means that for these, the participants' task is to match the target words with the right definitions, an arguably harder task than matching words with contexts.

b) As has been suggested before, a further reason proposed for these results is that, regardless of their success in identifying an acceptable translation equivalent, they reflect the greater depth of processing required for the Example Sentences group to complete the translation equivalents test (Craik and Tulving, 1975).

c) The Example sentence group participants may have established stronger mental links between the target words and the syntax and lexis surrounding and interacting with them than did the Dictionary Definitions group participants between the target words and the

> definitions. Since a word's collocates and grammar may be seen as part of the word, the surrounding context of a target word could be said to be *part* of the word, while the definitions are *about* the words. Example Sentence group participants' greater identification of target words with their contexts as compared with Dictionary Definition group participants' with the definitions may help account for the high retention test scores for the Example Sentences group participants.

Despite the problems with the retention test, specifically results concerning part of speech, the results are of value when we consider the difference in scores according to the materials used by the participants. They suggest that while dictionary definitions are undoubtedly more reliable as sources of information about word meaning, encountering unknown words in context may be a more memorable experience for many language learners, leading to greater retention of at least some forms of word knowledge.

6.5 Conclusion

The study reported in this chapter produced data regarding the comprehension and retention of unknown verbs and adjectives that is valuable in two important ways: in what it tells us about the nature of L2 word comprehension, especially regarding part of speech, and in the relevance that the findings have regarding L2 learners' encounters with unknown words. We will summarize these first from the perspective of part of speech then in terms of learning materials used. Finally, we will assess the value of the methods and instruments used in this study.

6.5.1 Part of speech

Overall, target word comprehension rates were significantly higher for verbs than for adjectives. This was true for both the Dictionary Definitions Group and the Example Sentences group. There were, however, very wide variations from word to word, with various adjectives showing higher comprehension rates than many of the verbs. Comprehension rates in this study overall were considerable lower for either part of speech than in the results for verbs reported in Chapter Four and for adjectives in Chapter Five. This suggests that the sentence-writing task used in the studies reported in the two earlier chapters may have aided participants' comprehension of the target words.

No reliable retention figures for the 40 target words were available through this experiment regarding verbs and adjectives, since quite a large number of participants failed to complete the final part of the retention test, for a set of 10 verbs. Results from the other three sets suggest that retention rates for verbs and adjectives are likely to be similar to each other, and both to be very low.

6.5.2 Learning materials

Results for comprehension of target words again confirmed the general superiority of the dictionary definitions over the example sentences, although there was wide variation from word to word and a considerable range among the participants in each group. This, too, challenges blanket statements about one source of information about the target words always being superior to the other. Further, we are reminded that comprehension of meaning is only one aspect of vocabulary knowledge, is not necessarily an indication of strengths in other areas of lexical development, and is no guarantee of retention.

Finally, it is worth noting that results for both groups were considerably lower than for their counterparts in the studies described in Chapters Four and Five.

The data from the retention test regarding learning materials showed a significant advantage for users of example sentences. This is in some respects surprising, given the much lower comprehension rates for participants in this group. It does, however, confirm the complexity of the issue of which material is best for learning a word, and the danger of simplistically equating word comprehension with retention.

6.5.3 Instruments employed

The experiment used in this chapter to demonstrate comprehension of the target words departed from the "Look Up Compose A Sentence" (LUCAS) method of demonstrating comprehension and use of the target words. While there are undoubtedly weaknesses to the LUCAS method, results from this experiment compared with those reported in Chapters Four and Five suggest that the completion of the LUCAS task does increase comprehension and retention rates. This should not be surprising since this method requires more time and more mental effort to be focused on each target word than that required for providing translation equivalents. In addition, in this study, the number of target words was increased and the time available per word decreased. Under these circumstances, we should also expect levels of word learning to fall too.

The retention test format employed here was intended to retain the sensitivity of the word context recognition multiple choice tests used in the studies described in the two previous chapters and, at the same time, to overcome the weaknesses associated with simple multiple choice tests. In fact, the test in this experiment, and the administration of the test, revealed its own weaknesses. Allowing sufficient time to complete the test would address one major problem, but the low retention rates recorded here may not only be an indication of lower levels of word retention but also of a lowered level of test sensitivity. Further, the test employed here only aims to measure a very specific aspect of word retention: the recognition of contexts in which the words were previously encountered. This does have its value but it also neglects many other aspects of word knowledge retention.

Despite the problems listed above, both of the testing instruments employed in this study have their own qualities. However, since variations of these instruments have been used in the experiments described in the past three chapters, it may be worthwhile to investigate the possibilities offered by other testing instruments.

6.5.4 Refining our objectives

If we now restate our research goals, we may better understand how to proceed. The aim is to gain a greater understanding of the nature of second language vocabulary acquisition through dictionary use and encounters with unknown words in authentic contexts. In the light of the studies reported in Chapters Three to Six, to further our achievement of this goal, we can see that the following objectives are worth pursuing:

i) To investigate the learning of a very large number of target items is preferable, since word comprehension and retention varies widely from word to word, both within and between different parts of speech.

ii) To follow the development of targeted words over a longer period of time so that measurable amounts of retention may take place.

iii) To use real dictionaries and full texts so as to investigate learning conditions that are closer to those found in natural L2 dictionary use.

In the following chapters, we will investigate how these objectives may be achieved.

Chapter Seven: Background to Case Studies

7.1 Introduction

The studies reported in the four previous chapters have shed light on a number of facets of L2 vocabulary acquisition in the context of dictionary use. They have also confirmed the suspicion, voiced in the literature review in Chapter Two, that some of the most widely used instruments for the investigation of L2 vocabulary development in this context are very limited in what they are able to reveal. In this section we will give an account of four case studies in which we compare the effect of two learning conditions on the English vocabulary of individual learners of English as a foreign language: extensive reading of an English text and extensive reading of an English text with the aid of a monolingual learner dictionary. In preparation for these studies, we need to ask what requirements we might have of instruments for the measurement of vocabulary change in these learning environments. To do this we will briefly consider the specific nature of L2 vocabulary development in the contexts of extensive reading and dictionary use. We will then consider what kind of instrument may be able to meet these requirements and how they may be used.

7.2 The nature of vocabulary acquisition in these contexts

We will begin by considering some of the characteristics of L2 vocabulary acquired through encounters with unknown words in texts and dictionary use. Two related aspects of vocabulary development in the context of extensive reading are that it tends to be incremental, and that these increments are typically small. According to Herman and Anderson's (1985) research into L1 vocabulary acquisition from encounters with unknown words through reading, the chances of gaining full understanding of a previously unknown word through one encounter with the word in context are typically between .05 and .11: between one in twenty and one in under ten unknown words would become known through one encounter in context. Expressed otherwise, the figures .05 to .11 may be used to estimate the average number of encounters required for total comprehension of previously unknown words: between ten and twenty times. Clearly the circumstances of learning from context are not as simple as these averages suggest; Brown (1993), for example, points out how factors such as saliency and conceptual familiarity may affect the rate of acquisition. Further, we cannot assume that words rated as unknown by language learners had not been encountered before, since single encounters may often not provide sufficient learning for a word to be recognised out of context.

Another aspect of L2 vocabulary acquisition in this context, with or without the use of dictionaries, is that not all previously unknown words in a text will receive the same treatment from learners; some unknown words will be focused on, others passed over, some will be looked up, some will not, some will be guessed correctly from context, while some will be guessed wrongly. For some words, the learner may make a mental note to learn the word, while other words may be dismissed as a waste of time.

A further characteristic of at least a proportion of our foreign language lexicon is that it is unstable: both in terms of our lexical knowledge and with regard to our confidence about this knowledge. This is indicated, but not discussed, in various studies by the sizeable numbers of items that are recorded as known in the pretests and not known in the post-tests (Krantz, 1990, for example). It is also reflected in the studies reported in Chapters Three to Six in the low retention rates as compared to the much higher comprehension rates recorded between one and three weeks earlier.

In sum, then, we can say that vocabulary development through encounters with unknown words in long texts tends to have the following characteristics: it often takes place in small increments, it is does not apply equally to all words, and much of it may be unstable. For words that are looked up in an MLD in this context, we may find many of the same characteristics: acquisition levels that vary widely from word to word, and lexical development that is often unstable. We may also, as with extensive reading without dictionaries, expect to find widely varying levels of comprehension. We now need to consider what kind of instrument may be suited to vocabulary development of this nature.

7.3 What kind of instrument do we need?

Considering the above aspects of L2 vocabulary learning, we can begin to picture the desired characteristics of an instrument which would be able to record the type of L2 vocabulary development that may occur in the context of extensive reading with or without dictionary use. Based on these characteristics, we will now specify our requirements of instruments to investigate this vocabulary development and suggest, practically, how these may be met.

To record small, partial changes in vocabulary growth, we would need an instrument that is sufficiently sensitive to small changes in word knowledge. This may involve a test in which participants record a range of states or extents of word knowledge, although this would depend on participants being able to evaluate and record their changing word knowledge in this way. Although the level of confidence required for language learners to publicly admit to partial knowledge may be an inhibitory factor, private self-assessment of vocabulary knowledge may be one way of making it easier for participants to record partial knowledge.

To reflect the incremental nature of the vocabulary growth, it would be valuable to have repeated encounters with the words and repeated tests to record changes in word knowledge. In connection with this, and to record vocabulary growth that is often unstable, we would need repeated testing to see what change there is between tests. We would also need to include large numbers of test items so that this instability may be reflected in the data without the data being distorted by the effect on word knowledge of repeated testing of the same items.

For three important reasons, a predictive aspect to the instrument would be valuable, enabling predictions of vocabulary development after, for example, five or ten encounters. One reason is that if there are only small incremental changes in vocabulary knowledge, we may expect little identifiable change through one or two encounters with a targeted word; a reliable prediction of the effect of subsequent encounters would give a fuller picture of the nature and rate of vocabulary development through one learning condition. A second reason is that prediction of subsequent vocabulary development would make it possible to compare the effect of two different learning conditions on the vocabulary development of individual learners. As the review of the literature shows, traditional studies have very often experienced difficulties with regard to establishing the equivalence of experimental groups when comparing the effect of different learning conditions (see, for example, Luppescu and Day (1993) or Krantz (1990)). A third reasons is that if the instrument is able to provide reliable projections of future vocabulary development, it will reduce the requirement to constantly test and retest, and so reduce the effect that this repeated testing may have on the participant's knowledge of the targeted items.

Finally, to gain a global picture of this incremental growth, within which we may find a large degree of variation, there should be a large number of test items. This would also give us the opportunity to investigate the effect that some of the different types of participant treatment for individual words (only guess, look up once, look up more than once…) may have on the vocabulary development of different types of words. In addition, the larger the number of test items, the smaller the effect of tests themselves on vocabulary knowledge.

To bring these requirements together, then, an instrument suited to recording vocabulary acquisition in this context should be able to test a large number of items, it should involve repeated exposure to and testing of the items, it should be in a format that would allow participants to record partial and changing knowledge of the items, and it should be able to provide a prediction of change following further encounters with the test items.

7.4 How may these requirements be realised?

Word lists for translation and multiple-choice tests, the instruments most widely used in previous studies, have most often been used for no more than 24 items in one test, and usually with only one or two testing sessions. These instruments each have their qualities, but they do not satisfy most of the requirements listed above. We will now consider how, practically, an instrument may be designed to suit the vocabulary acquisition environments under investigation.

If we want to record degrees of knowledge of words, Wesche and Paribakht's (1996) Vocabulary Knowledge Scale (VKS) is perhaps the first instrument to come to mind. Its aim is to identify in which of five states a learner's knowledge of a word may be, from recognition to accurate production. One major problem with the VKS, from our perspective, however, is that a thirty-minute test would only be sufficient for the testing of 15-20 items. If we imagine a "light" version of the VKS, involving no writing, then it does become a more practical possibility. Meara's V_States (v.03, 2000) is a test of this type. It is a computer-based test, and for each test item participants only have to click the box matching their estimation of their knowledge of the word, from a choice such as

I do not know this word	*I don't think I know this word*
I think I know this word	*I definitely know this word.*

This process is much quicker than the VKS, with each item requiring under 5 seconds, and allows for the inclusion of many more items; 300 items may be tested in a session lasting less than 30 minutes. As a quick and easy test, a further advantage is the very limited amount of time focused on each word, so reducing the effect that repeated testing would otherwise be likely to have on the testee's knowledge of the items targeted for the study.

V_States is a self-assessment task in which participants would rate their estimation of their own knowledge of the targeted items. Although such tasks have their own problems, which we will consider in more detail in the following chapters, they do have the advantage of more sensitively eliciting participants' partial knowledge of words than would a more test-like form of evaluation.

One way to achieve repeated encounters with words is to repeatedly read, or listen to, a text in which the targeted items appear. Although this means that targeted words are encountered by participants in the same context each time, rather than in the different contexts of a more natural reading environment, repeated reading does provide a predictable, controlled environment that is amenable to investigation. As for the test, if sufficiently large numbers of items are tested, with sufficiently little time spent on each item, we can expect the practice effect of repeated testing to be minimal. As a further safeguard, test items can be randomly reordered for each test.

One further, and very important, advantage of using a repeated, multi-state vocabulary test is that it furnishes data from which a transitional probability matrix may be created to make projections of future vocabulary development under one or more learning conditions. We will now go on to demonstrate how the data from this testing instrument can be used to create a matrix, and how the data from this matrix may be used to provide predictions of future vocabulary development for the set of lexical items under investigation. (For a more detailed explanation, see Meara and Rodriguez Sanchez, 2002.)

7.5 Creating a transitional probability matrix

We will begin with a brief explanation of how a transitional probability matrix works and what kind of data it produces. As an illustration of the procedure employed using matrices in the following studies, we can imagine that we have 100 words as test items and we ask a case study participant to rate each of these items as being in one of four states of knowledge ranging between *definitely do not know* (0) and *definitely know* (3). The result might at a given time (t1) be as follows:

States	0	1	2	3
tl	50	20	20	10

The figures shown here are for the number of items in each of the four states at one given time. Because they only show us numbers for one specific time, they do not tell us whether the participant's knowledge is changing over time, how it may be changing, or what the rate of change may be. If, after a certain period, the participant takes the test again – t2 – we will be able to identify movement of items between the two adjacent test times. The numbers of items that stay in the same state or move to another state will give us an indication, provided the learning condition remains constant, as to the numbers of items which are likely to remain in the same state or move to other states following subsequent learning sessions. This movement can be expressed in terms of proportions, as shown in the illustrative data in Table 7.1.

The column in Table 7.1 labelled 't1' shows the number of items in each state recorded at the first testing session, while the column labelled 't2' shows numbers of items in each state at the second session. The central four columns show the proportions of items in a particular state at t1 that at t2 stayed in that state or moved to another state.

Table 7.1
Deriving a transitional probability matrix

		Proportions moving to these states at t2				
State		0	1	2	3	
	t1					*t2*
0	50	.6	.2	.1	.1	37
1	20	.2	.5	.2	.1	25
2	20	.1	.2	.4	.3	19
3	10	.1	.1	.2	.6	19

If we take as an example the 50 items in state 0 at t1, we can see that at t2 60% of these (30 items) stayed in state 0, 20% (10 items) moved to state 1, 10% (5 items) moved to state 2, and 10% (5 items) moved to state 3. Viewed from the perspective of t2, the 37 items in state 0 at t2 is made up of 60% of the 50 items that were rated as state 0 at t1 (30 items), 20% of the 20 state 1 items at t1 (4 items), 10% of the 20 state 2 items at t1 (2 items) and 10% of the 10 state 3 items at t1 (1 item).

The proportions shown in Table 7.1 illustrate how numbers of items moving or staying in each state between t1 and t2 can be expressed can expressed as proportions. These proportions in the central columns of Table 7.1 can, then, be used a transitional probability matrix to predict how we may expect items to behave in subsequent sessions. Table 7.2, using this matrix with the same illustrative data, shows the projected figures for this for a further eight times.

Table 7.2
Projected movements for iterations 3 – 10

	Actual scores		Predicted scores							
State	t1	t2	*t3*	*t4*	*t5*	*t6*	*t7*	*t8*	*t9*	*t10*
0	50	37	*31*	*28*	*27*	*26*	*25*	*25*	*25*	*25*
1	20	25	*26*	*25*	*25*	*25*	*25*	*25*	*25*	*24*
2	20	19	*20*	*21*	*21*	*22*	*22*	*22*	*22*	*22*
3	10	19	*23*	*26*	*27*	*28*	*28*	*28*	*29*	*29*

What happens to the numbers of items in each state with successive iterations of the matrix? As Table 7.2 demonstrates, by far the greatest change in numbers in each state is between t1 and t2, with, in this illustration, changes of numbers per state falling or rising an average of 9 items. There is still a substantial amount of change between t2 and t3 – an average of 3 items per state. By the time we reach t6, however, there is a change from the previous time of less than one item in any state. By t10, there is virtually no change from t9 in the number of items in each state: under 0.1 item per state.

This does not mean that movement between states falls to the same degree over the same number of times. As, with a matrix, proportions of items in each state moving or staying remain constant and since numbers in each state remain fairly steady after about t4, the actual numbers moving between states will remain fairly constant too. This means that while after a few times the number of items in each state becomes almost static as a balance is reached between incoming and outgoing items, there may continue to be a large amount of hidden "under-the-surface" movement between states.

We can see something of the potential value of using a transitional probability matrix for predicting L2 vocabulary development in a stable learning environment, either in terms of acquisition or attrition. A number of studies have investigated the practical application of these theoretical assumptions (among them, Meara and Rodriguez Sanchez, 2000, Horst and Meara, 1999, Waring, 1999, and Milton, 2001). The studies have tended to focus on extensive reading, and although no other studies have investigated the use of transitional probability matrices in the context of dictionary use, this application is an obvious development. In addition, then, to investigating vocabulary acquisition in the context of extensive reading and listening, with and without the use of dictionaries, we will also investigate the value of using a transitional probability matrix in this context.

7.6 Conclusion

As the above discussion has indicated, the type of instrument for examining vocabulary acquisition that we have selected appears to meet our requirements in a large number of important ways, to an extent unequalled by any instruments employed in previous studies. However, because of the volume of reading required, the repeated testing, and the relatively long period over which this research would take place, the only practical way of proceeding with this research is with case studies of individual language learners. Larger scale studies may be possible in the future, but case studies are a useful vehicle as we investigate the nature and volume of vocabulary learning through dictionary use and extensive listening and reading. They are also a good means by which we may investigate the value of the transitional probability matrix as a predictive tool for vocabulary development.

Although the instrument V_States does appear to meet many of the proposed requirements for measuring L2 vocabulary acquisition in the two learning conditions under investigation, we need to anticipate and address, as required, various potential problems with its use. Specifically, participants' self-rating of vocabulary knowledge may result in differences between individuals in their conception of different states of word knowledge, in intentionally dishonest reporting, or in misidentification of words that are not known for words that are known and known words for unknown words.

We will now go on to these individual studies, explaining in more detail how they were conducted and revised, considering the real and projected data they produce, and reflecting on how this instrument informs us about vocabulary acquisition in the context under investigation.

Chapter Eight: An Exploratory Case Study

8.1 Introduction

This chapter describes the first in a series of four case studies investigating the vocabulary development of an intermediate level learner of English as a result of using a monolingual English learner dictionary during extensive and repeated reading of an English text. In this study, a Japanese adult learner of English read a complete book for older children seven times: three times without access to a dictionary and four times using an MLD. Before and after each reading, she evaluated her level of word knowledge for each of over 300 hundred words which appear in the text.

This study has two main purposes:

i) To investigate L2 vocabulary development in the two learning environments of extensive reading without the use of a dictionary and extensive reading while using an MLD;

ii) To evaluate the usefulness and suitability of the method, the instruments, the text, and the targeted items used in the study for measuring vocabulary development in the two L2 learning conditions under investigation.

As described in Chapter Seven, Horst and Meara (1999) report on the use of a model for measuring and predicting the partial and incremental L2 lexical growth that may occur through extensive reading. A broadly similar model is employed in this study, in which hundreds of test words are selected rather than just the 20 or 30 items tested in most studies. In this case study, one participant read one long text a number of times, ensuring multiple encounters with the same word, although admittedly in only one context.

Using a transitional probability matrix of the kind described in the previous chapter, the use of multiple readings of the same text by one participant followed by rating sessions, makes possible the comparison of projected vocabulary development in one learning condition with its actual development in a second condition.

8.2 The study

As this study was largely exploratory in nature, aiming to assess the applicability of the methods and instruments employed to the learning conditions under investigation, only two hypotheses were proposed relating to vocabulary growth resulting from the two learning conditions:

1. That the participant would demonstrate vocabulary growth resulting from extensive L2 reading, both with and without the use of a monolingual learner's dictionary.
2. That there would be a clear benefit in terms of vocabulary growth attributable to dictionary use over and above that attributable to extensive reading alone.

For this experiment, one intermediate level adult learner of English read a complete 150-page book once a week for a total of seven times. She took a vocabulary rating task prior to starting the reading sessions and took a further rating task after every reading. Because each reading of the whole text took the participant between eight and ten hours, no specific time was set for reading the book other than that it would be completed within a seven day period. Given the time required for reading and testing (almost 80 hours in total), a case study with a single willing, and financially compensated, participant was selected as the most suitable method for this project.

8.2.1 The participant

The participant asked to take part in this study was a 20-year-old 3rd year student majoring in English at a middle-ranking Japanese university. Almost all her experience of language learning had been in formal learning contexts: in class, doing set homework, or preparing for exams. She had not been abroad or spent any time in an exclusively English-speaking environment. She had very little previous experience of extensive reading in English: only one graded reader eighteen months previously. Her knowledge of English would be rated as intermediate, with a TOEFL score at the time in the region of 470.

8.2.2 The reading text

The text used for reading and as the source for 310 of the 330 test words was C.S. Lewis's *The Lion, the Witch and the Wardrobe* (1950). The book was chosen for the following reasons:

1. As a book for older children, with 30 pages of pictures, it would be relatively easy to read. At the same time, not having a specifically controlled vocabulary, there would be a place for dictionary use to aid comprehension.

2. As an allegorical story that can be read at different levels it was felt that in terms of interest it would support multiple readings.
3. A vocabulary recognition task in which a similar level student was asked to identify unknown words in the book suggested that over 95% of the text would be known. This is in line with that suggested as necessary for optimum comprehension (Laufer, 1992, Hu and Nation, 2001), and as such would provide a suitable environment for the guessing of meanings of unknown words in the text.
4. In terms of size, (about 40,000 tokens), the text was long enough to provide a sufficient number of task items: 310 words which occur only once in the text, and the majority of which would be unknown.

8.2.3 Targeted word selection

The whole text was scanned into a computer and a concordancing package which includes a word frequency count (*Wordsmith*, v.03, Scott, 1999) was used to identify words which occur only once in the text. There was a total of over 3,000 word forms, or types, occurring in the text, with 1,546 occurring only once. Over half of these, however, were excluded from consideration as task items as there were other members of the same lemma or word family in the text. Of the remaining 700 or so single-occurrence words, about 300 were judged to be probably known by the participant, leaving something in the region of 400 hundred words which could be used as task items. The final selection from these was made to present a variety of words to the participant. So, as far as possible where two similar words occurred, such as *chirping* and *chirruping*, only one was selected. The final set of 310 targeted words is listed in Appendix 8.1. An additional 20 low frequency words were drawn randomly from *The Longman Dictionary of Contemporary English, 3rd Edition* (Summers et al, 1995; henceforth LDOCE3), to serve as a control. These were the control items:

ain't, bitumen, burring, champing, chancel, contumely, extenuate, frugal, grenade, heady, intruding, loge, minion, mow, paisley, prep, rellos, selvage, spindle, tonnage

8.2.4 The dictionary

From the fourth reading onwards, the participant was encouraged to use a monolingual EFL dictionary. LDOCE3 was chosen because it was the best-selling MLD in Japan, because it was said to be closest in format to the bilingual dictionaries with which Japanese learners of English are more familiar, and because the case study participant already owned a copy of this dictionary and was, to some degree, familiar with it. The participant used her own copy of LDOCE3 during the study.

8.2.5 Vocabulary evaluation method

A computer programme called V_States (v.03, Meara, 2001) was used for this case study. The programme presents the targeted items, one by one, in random order and records the participant's responses. For each item that appears on the screen, the participant rates her knowledge of the word by clicking on one of the four buttons on the screen, labelled as follows:

0 – I don't know this word
1 – I'm not sure I know this word
2 – I think I know this word
3 – I definitely know this word

A computer log file recorded the participant's rating of each item at each rating session, as is shown below for a few of the targeted words. The full set of data for the participant is provided in Appendix 8.3. The responses for the first V_States session for each word are shown at the far left of each row, with the responses for the final session closest to the targeted word.

```
00000000:shins
00000000:dungeons
02233333:gloriously
01000001:grate
00000011:boughs
00010111:gorse
01122333:primroses
22333333:flushed
```

The participant took an average of about 4.5 seconds per item, and took under 30 minutes to complete one rating session for the total of 330 items. This time remained largely constant for all of the rating sessions.

After the experiment was completed, a final test was conducted in which the participant was required to give the meanings of the words she had rated as definitely known in the final V_States session.

8.2.6 Procedure

Before reading the text for the first time, the participant rated her knowledge of the 330 task items. In the following week, she read the text, without referring to a dictionary of any kind. The first reading took almost 10 hours to complete, over a period of seven days. The participant then rated her knowledge of the targeted items again. This continued for two further reading and rating sessions.

From the fourth reading onwards, the participant was allowed to use an MLD. At this point, the participant was given a very brief guide to using the dictionary (see Appendix 8.2), based on the guidelines suggested by Nation (1990: 136). She was advised to limit her dictionary use and to aim to keep within the 10 hours that it took her to read the text on the first reading. The participant was also told to affix a Post-it tab in the dictionary for each word she looked up during the reading. Although a copy of LDOCE3 had been prepared for the participant, she offered to use her own copy of this dictionary. However, as the participant was using her own dictionary, she felt free to keep a record of looked-up words by highlighting these words in the dictionary, as was her habit. We will discuss the consequences of this below.

8.2.7 Dictionary use

We will now report what use the participant made of the dictionary over the latter four "with dictionary" reading sessions. Table 8.1 provides a summary of recorded dictionary use by the participant during the study.

As Table 8.1 shows, the number of words recorded as having been looked up in one session ranged between 28 and 42 words, averaging just over 34 words per session. Proportions of looked up words which were targeted items averaged 24.1%, ranging between 19.0% and 30.1% of the number of words looked up in one session. In all, only 10.6% of the 310 targeted items from the text were looked up.

Table 8.1.

Participant's dictionary use in reading sessions 4 – 7

Reading session	Looked-up words	Targeted items
4	42	8
5	31	8
6	28	6
7	36	11
Total	137	33

These were the 33 targeted words from the text that were looked up:

aisle, badgers, beckoned, bluebottle, boughs, bunks, decent, decoy, dodging, dunces, flask, gasped, glittering, goose, hilt, knuckles, myth, overwhelming, pitter-patter, puddles, reckoned, sill, sizzling, sluice, snigger, splendid, splutter, stratagem, struggled, swiftest, swirling, tapping, wireless, ye.

A further notable feature of the participant's dictionary use is that no words were recorded as having been looked up more than once. This may have been a consequence of the participant's highlighting words looked up in her dictionary; when words were looked up a second time, the participant was aware of this and, out of a misplaced sense of shame, did not record any look-ups for words which had previously been looked up.

8.3 Results

Four sets of data were obtained through this study. One set of data consists of the results of the eight V_States rating sessions. A second set of data was derived by using matrices based on the results for the V_States session. A third set of data is the participant's rating of the 20 control items for the eight V_States sessions. The fourth set of data is from the Final Meaning Test, in which the participant was required to give the meaning of the targeted words which she had rated as definitely known in the final V_States rating session.

8.3.1 V_States rating session results

The results from the V_States rating sessions are shown in Table 8.2 below. The raw data from the participant's V_States log file are provided in Appendix 8.3, with totals for targeted items shown in Appendix 8.4.

The first rating session (t0) was completed before the participant read the text; three were completed without the benefit of dictionary use (tl to t3); and four were completed after readings where a dictionary was available (t4 to t7). We will also look at the participant's rating of the 20 control items over the 8 rating sessions. Finally, in conjunction with the V_States scores, we will also report the participant's answers in the Final Meaning Test for the words she had rated as definitely known (state 3) in the final V_States task.

Table 8.2

States of targeted words for each V_States rating task

Learning condition	Session	state 0	state 1	state 2	state 3
Pretest	t0	246	40	15	9
Reading,	t1	116	134	45	15
no dictionary	t2	169	69	46	26
	t3	142	81	57	30
Reading,	t4	112	73	80	34
dictionary used	t5	86	102	80	42
	t6	59	113	72	66
	t7	37	108	75	89

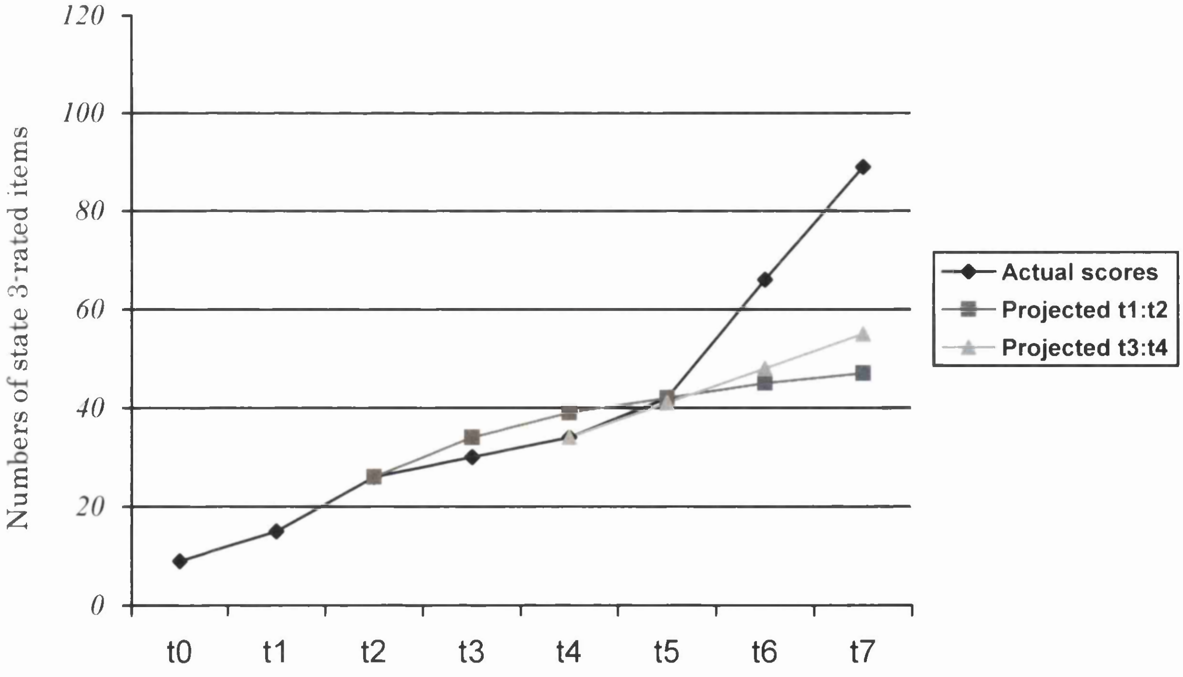
To summarise these results, Table 8.2 shows an overall steady fall in the number of state 0 items. However, there is a sudden drop of over 120 items from t0 to t1. This figure rises again by over 50 items at t2 before falling by small amounts at t3 and at each subsequent session. As for state 1-rated items, these tend to increase steadily over the eight sessions, with the exception of t1; at this session there was an abnormally large rise, mirroring the drop in state 0 items at this point. Numbers of state 2 items rise steadily over the first 5 tests, after which they remain largely constant, falling slightly in the final two sessions. Numbers of state 3 items rise steadily until t5, after which there are sharp increases at t6 and t7.

8.3.2 Matrix based projections

A matrix based on proportions of items in the four states at t1 and t2 was used to generate a projection of the numbers of words in each state in tasks up to t8 for the original learning condition. The data for this, together with raw data for the targeted items, are shown in Appendix 8.4. The use of the matrix makes it possible for the actual data for t4 to t8, reflecting the new learning condition, to be compared with this projection for the original condition. A further matrix, based on t3:t4, was used to create a projection for the second learning condition. The reason for this second projection was to assess the validity of this method by comparing actual and projected scores for the same condition. The actual and projected state 3 scores are shown in Figure 8.3.

Figure 8.3

Actual and projected numbers of state 3 items (total number of items = 310)



To compare the two learning conditions, we can look at state 3 scores from t3 onwards in Figure 8.3. These show actual scores reflecting the reading with dictionary use condition compared with projected scores based on a matrix from t1:t2 results for the first learning condition. As we can see, although at t4 the projected scores for reading without a dictionary are a little higher than actual scores with the new condition, by t5 the scores are almost identical. The projection for the reading with dictionary use condition, based on t3:t4, suggests that by t6 numbers of state 3-rated items in this condition would rise slightly above those projected for reading without a dictionary. In fact, actual scores rose sharply for t6 and t7 and diverged sharply from the projected figures for this condition. These results will be discussed in detail below.

Now we will consider comparisons between numbers of actual and projected state 3 items within each learning condition. For the reading without dictionary use condition, actual and projected scores are only available together for t3. Here, the two scores are relatively close, with 34 projected state 3 items as opposed to 30 items actually rated as state 3. For the reading with dictionary use condition, comparisons between actual and projected numbers of state 3-rated items are available for t5 to t7. For t5, actual and projected scores are almost identical: there are 42 projected state 3 items and 41 items actually rated as state 3. For t6 and t7, actual and projected scores are widely divergent: 66 actual state 3 items at t6 as compared to 48 projected state 3 items, and at t7 89 actual state 3-rated items as compared to 55 projected state 3 items. This suggests that there is some problem with the participant's rating of state 3 items, an issue that is discussed below in relation to the Final Meaning Test results.

8.3.3 Rating of control items

We will now look at the participant's rating of the 20 control items that were included in the V_States rating sessions together with the 310 targeted items from the reading text. The data for these items are shown in Table 8.4.

Table 8.4

States of control items for each V_States rating task

Learning condition	Session	state 0	state 1	state 2	state 3
Pre-test	t0	17	2	1	0
Reading, no dictionary	t1	6	10	3	1
	t2	9	8	3	0
	t3	10	6	4	0
Reading, MLD use	t4	7	7	4	2
	t5	6	8	5	1
	t6	3	8	6	3
	t7	4	7	5	4

As these control items were only encountered in isolation in the V_States sessions, without any contextual information regarding word meaning, we should expect no increase in the participant's reported knowledge of these items. However, we can see that participant rating for the control items tends to parallel her rating of the targeted items. This is especially apparent at points at which V_States rating for targeted items is atypical of trends otherwise observable over the eight V_States sessions: at t0, t1, and t6 and t7 for state 0; at t0 and t1 for state 1; and at t6 and t7 for state 3. There was, also a general decrease in state 0 rated control items over the eight sessions, but this is largely mirrored by a rise in state 1 items, with only slow and intermittent increases in the top two states. The control helps us to identify possible problems with the participant's rating of the items, which will be confirmed as we look at the Final Meaning Test results.

8.3.4 Final Meaning Test results

In addition to the eight computer-based V_States rating sessions, the participant took one final test, in which she was asked to give the meaning, in Japanese or English, for the 89 words she had rated as state 3 ("I definitely know this word") at t7. She only achieved an accuracy level of 52.81% for these items, even with a generous rating of the translation equivalents. However, when the state 3 responses from t6 rather than t7 were used, the proportion of acceptable answers was much better 65.62% (for 66 items). This confirms the suggestion that for this participant her least reliable ratings of word knowledge were in the final V_States session.

8.4 Discussion

We will begin by asking whether data from this study confirmed our hypotheses regarding L2 vocabulary development in the two learning environments under investigation. We will then focus on the three main sets of data relating to the two instruments used in this study: the V_States rating scores, the predictions of V_States scores, and the Final Meaning Test. Our main concern here is to evaluate the success of these instruments in fulfilling the aims of the study. Before concluding this chapter, we will also consider how the various problems encountered through this study may be addressed and overcome in subsequent studies.

8.4.1 Hypotheses addressed

The first of the two hypotheses was that the participant would demonstrate vocabulary growth resulting from extensive L2 reading, without the use of an MLD and while using an MLD. As the results show, both in the fall in combined numbers of state 0 and state 1 items and in the rise in state 3 items, the participant did demonstrate vocabulary growth resulting from extensive L2 reading, both with and without the use of an MLD. Within the first learning condition, the number of state 3 items rises from 9 in t0 to 30 in t3, while the combined number of items rated as state 0 or 1 fall from 286 to 233 over the same period. As for the second learning condition, from t3 to t5 the number of state 3 items rose from 30 to 42, and the number of states 0 and 1 items fell from 223 items to 188 items, falling further to 145 items by t7.

The second hypothesis was that there would be a clear benefit in terms of vocabulary growth attributable to dictionary use over and above that attributable to extensive reading alone. The actual scores for the reading with MLD condition, from t6 onwards, and projections for this second condition, were both higher than projections for the second condition of reading without a dictionary. The prediction based on vocabulary development in the first condition and the actual scores from t6 onwards are higher than the prediction for the first condition at these times based on the t1:t2 matrix. Problems with the participant's rating of state 3 items in the final two sessions, however, suggest that there is insufficient evidence to be able to maintain with confidence that for this learner there was a clear benefit, in terms of vocabulary development, gained from using an MLD dictionary during extensive reading in a foreign language. Clearly, more research is needed, and we will make suggestions regarding further research as we evaluate the various data obtained through the instruments used in this study.

8.4.2 V_States scores

The use of V_States over eight sessions, with 310 target words moving among the four states of vocabulary knowledge, provides a very impressive picture of vocabulary development in this case study. Although only a single subject study, this provides a record of a total of 2,480 ratings of targeted words over the seven week period, considerably more than that in the majority of multiple subject studies reviewed in Chapter Two.

As Table 8.2 shows, the numbers of items in each of the four states over the eight V_States rating sessions, and the changes in numbers from session to session, are generally what we might expect from a language learner reading a book a number of times. While there is considerable sustained movement to both upper and lower states throughout the eight rating sessions, overall movement of items is slowly and steadily away from state 0 and in the direction of state 3. The account starts with a considerable jump between the pretest t0 and t1; the number of state 0 items falling sharply and the number of state 1 items rising by almost the same amount. This, too, is not surprising and may, in part, be accounted for by the failure to recognize the isolated items in t0, the encounter with the items during the reading session between t0 and t1, and the recognition of these items in t1. This seems similar to the "Boulogne ferry effect" (Meara, 2005) in which substantial numbers of dormant vocabulary items are reactivated very quickly, in this case through reading a 40,000-word book.

A further factor that may help us gain a better understanding of the participant's rating of items in V_States is the issue of what knowing a word means and, specifically, the participant's understanding of the four word knowledge states in V_States. The description of state 1 – "I'm not sure I know this word" – may be interpreted as degree of familiarity with the word form as well as knowledge of the word's meaning. It does appear that, at times, the participant was equating knowing a word with recognition of, or familiarity with, the written word form. There is support for this hypothesis in the rating of items at t2 as the participant corrects her "over-rating" of items as state 1 at t1 so that this figure falls and the number of state 0 items rises again.

The participant seems especially to have a problem with the two lowest states, and in isolation they do not provide much useful information about the learner's vocabulary development. However, if we combine the scores for the two states 0 and 1, we can see a regular sustained decrease over the eight V_States sessions that is indicative of steady vocabulary growth in states 2 and 3. These are shown in Table 8.5 below. In Chapters Nine and Ten, we will investigate further the possibilities offered by the combination of items in neighbouring states, both as raw data and as a basis for transitional probability matrices.

Table 8.5

Participant's combined state scores

	t0	t1	t2	t3	t4	t5	t6	t7
States 0 and 1	286	250	238	223	192	188	172	145
States 2 and 3	24	60	72	87	118	122	138	165

One further aspect of the participant's scores that stand out as atypical of her rating of targeted words is the sharp rise in state 3-rated items in the last two V_States tests, t6 and t7. This is clearly visible in Table 8.2, in Figure 8.3, and also in the figures for the control items in Table 8.4. In the graph, it is highlighted by the extent to which actual numbers of state 3-rated items at t6 and t7 diverge from projections for state 3 items in these rating sessions.

There are two likely reasons for large rises in numbers of words that the participant rates as definitely known. One is the confusion, again, between increasing familiarity with the items and being sure about their meaning; repeated encounters through reading and evaluation may increase the participant's familiarity with the items but, alone, would not be sufficient to substantially increase word knowledge or confidence about that knowledge. The other reason may be the participant's desire to give a good final result for the researcher, who was also her teacher. This reason is proposed partly because the Final Meaning Test, in which she was required to give the meaning of the items rated as state 3 in t7, was a complete surprise for her, and partly because of her inability in this test to give the meaning of words which she had just rated as definitely known. A third possible reason, that the participant may have forgotten to use the dictionary during a reading session, may be dismissed by the evidence provided by the tabs placed in the dictionary to show looked-up words from the text.

8.4.3 Projections of V_States scores

In this study, the main focus has been on targeted words which were rated as state 3. This has also been the focus for projections obtained through transitional probability matrices derived from the participant's scores. The main projection obtained by this method is for comparing the two learning conditions of reading without and with the use of a dictionary. As Figure 8.3 shows, the projected figure was a little higher than the actual figure for state 3 items at t3 and t4, was virtually identical at t5, and was considerably lower at t6 and t7.

At first glance this divergence between actual and projected scores in the last two sessions suggests a problem with the predictive power of this model. However, as we have already noted, control item results and Final Meaning Test scores suggest that the actual state 3 scores at t6 and t7 are unreliable, rather than the projections. Rather, the projections help confirm this problem and indicate a further application of these projections; where the participant's responses are suspect in some way, this may be brought to light by comparison with the projected scores.

A further confirmation of the value of these projections may be found by comparing the projections using t3:t4, for the second condition of reading with the aid of a dictionary with projections based on t1:t2 for the first condition. These are very much how we might expect, with dictionary use while reading providing a small but distinct advantage over continued reading without a dictionary. Both projections differ widely from the actual state 3 scores at the final two rating sessions. As we look at the responses for the Final Meaning Test, we may gain a better understanding of other factors involved.

8.4.4 Final Meaning Test

In the Final Meaning Test, the participant was able to provide accurate meanings for only 52.8% of the 89 items rated state 3 in the final V_States test. The combination of a number of factors may account for this result. One factor is the same as that for the sudden jump in the state 3 ratings for the final V_States session: a desire by the participant to show a good result for the case study. The participant was chosen for her reliability and her willingness to participate, and perhaps a desire to "please the teacher" resulted in this over-rating of her knowledge of the test items in the final session. A further reason is the high number of words for which the meanings given indicate that they were misread in the Final Meaning Test, and presumably in the V_States sessions too. Almost a third of the wrong answers can be attributed to these misreadings, which are listed in Table 8.6 below.

As the list demonstrates, many of the test items were mistaken for more common words with similar renditions within the Japanese phonological system: 13 of the 89 FMT items, and over one third of the wrong answers in this test. One possible reason for the high proportion of such items is that while the 13 items represent almost one third of wrong state 3 responses at t7, they account for less than 17% of the total number of items rated state 3 at t7 and under 5% of all 310 V_States items. As they are mistakenly rated as known words, their influence is magnified. Factors accounting for this number of such items may be a combination of the isolated presentation of the targeted words in the V_States sessions and storing of English words in the learner's mental lexicon with Japanese pronunciation and orthography (Chikamatsu, 1996; Nakamura, 2001), with the result that words presented with their English spelling are not easily identified.

Table 8.6

Evidently misread state 3-rated items

Test item	Answer given	Meaning
bellowing	次の (*tsugino*)	following, next
craves	洞くつ (*doukutsu*)	cave(s)
device	分配する，わける (*bunpaisuru, wakeru*)	divide
flushed	ぱっと輝く (*pa-to kagayaku*)	suddenly shine (flash)
fluttering	平な (*tairana*)	flat
fond	池 (*ike*)	pond
forth	たたかい (*tatakai*)	fight (fought)
healed	かかとをならす (*kakato o narasu*)	to stamp with the heel
mortar	モーター (*mo-ta-*)	motor
puddle	こぐ (*kogu*)	paddle
revelry	レベルごとの (*level gotono*)	of each level
sluice	うすく切る (*usuku kiru*)	slice
vultures	*not real world*	virtual

It appears, then, that for this participant the pronunciation of English words is often transposed into sounds that are used in the pronunciation of Japanese, and the word's orthography is recorded in Japanese phonetic script. This suspicion is confirmed by the degree to which misread English test words are mistaken for other words which, if written in Japanese, would share the same orthographic form. Examples of this are *puddle* being mistaken for *paddle*, *flushed* for *flashed* and *fluttering* for *flat(tering)*. For each pair of words, the two vowel sounds [æ] and [ʌ] would be the same if represented in Japanese script:ア. Further evidence of this is found in the participant's look-up behaviour when, for example, she looked up the word *bank* after encountering the word *bunks* in the text. In other the cases, such as *fond* and *pond*, the initial sound of the pair of words use the same basic Japanese phonetic symbol: ホ (*f/ho*) and ポ (*po*) for *fond* and *pond*. This, too, suggests that many English words in the learner's mental lexicon may be stored, in terms of pronunciation, as Japanese words as if they were loanwords. In other cases, too, such as *revelry* being identified with *level* or *forth* with *fight* (or, rather, *fought*), mistakes can be attributed to the use of one sound in Japanese (レ) for both *le* and *re* or to a sound [θ] not existing in Japanese.

One further possible factor in the participant's inability to provide the meaning for so many of the words for which she gave a state 3 rating may be found in her understanding of what knowing a word means. At the beginning of the study she was told that in rating her knowledge of targeted words she should rate her knowledge of their meanings. Despite this, at times during the period of the study her understanding of knowing a word may have become closer to being how familiar the word is to her rather than how well she knows the word's meaning. This shifting of understanding can also be seen, as was discussed above, in the volatility of state 0 and 1 scores.

To summarise, then, it appears that the participant's low rate of success in the Final Meaning Test may be due to the following factors: her over-rating of vocabulary knowledge to please the researcher; her misreading of targeted items; and her unstable understanding of what knowing a word means. We will now go on to suggest means by which we may address these, and other, problems in future studies.

8.4.5 Problems and possible solutions

We will now summarise the problems identified through this study and propose changes that may improve the means by which the use of the instruments employed in this study may be adapted to better assess the extent and depth of vocabulary acquisition in the two conditions under examination. Five specific problems have been identified through this study. The first, easily solved, is that of the participant not recording second or subsequent look-ups of words. Lending a new dictionary to participants, rather than letting them use their own, will ensure that the participant keeps no record of look-ups from previous sessions. Additionally, participants could be given explicit assurance that repeated looking up of the same word is normal dictionary use behaviour and should be recorded by them in the study.

A second issue is with the large number of items that were misread and rated as state 3 by the participant. Although this problem may vary in severity depending on the individual, one way of reducing the effect of this is to replace words that have been, or are likely to be, misread. It is not difficult to replace words that were evidently misread if the same text is used again. Identifying test items that are likely to be misread is a little more difficult. There is some regularity in words that were misread: they were usually mistaken for higher frequency words, and for words that if pronounced with sounds used in Japanese would share a similar phonetic representation. If we exclude such words, we may largely overcome this problem. However, we should also recognize that by doing so, we may be simply avoiding a typical feature of Japanese learners of English at this level: that, typically, a certain proportion of words will be confused with others. Finally, we need to bear in mind that in this study we have been looking at data for a single learner. Only further research will reveal whether this participant's behaviour is typical of similar level learners with the same mother tongue, or whether she may experience idiosyncratic difficulties with storing and recognising English words.

A third problem is that of the participant over-rating her knowledge of test words in the final V_States test. There are various ways of addressing this issue: stressing frequently that rating a word as state 3 means being able to give its meaning; informing the participant that there will be a final test requiring the meaning of all state 3-rated words; having one more session of reading and rating word knowledge then discarding this data and basing the Final Meaning Test on words rated as state 3 in the penultimate V_States test; or suggesting initially that there would be eight readings and nine V_States sessions then stopping unexpectedly after the eighth session and using this data in the Final Meaning Test.

Fourthly, a related matter to that mentioned above is that the participant's understanding of the four states appears to have been unstable and tended to slip during the study, from knowledge of meaning of the words to familiarity with the word forms. Solutions to this are to stress that participants should rate knowledge of word meanings rather than familiarity with word forms, to repeat this statement prior to each V_States test, and to change the description of the states onscreen from, for example, *I definitely know this word* to *I definitely know this word's meaning.*

The fifth issue is that relatively few test items were looked up by the participant: less than 40 in total over the four reading with dictionary sessions. In one respect this is not as great a problem as may first appear; dictionary use cannot be seen as merely helping comprehension of the looked-up words alone, but also of the text in general and, more specifically, of words in the environment of looked-up words. On the other hand, if we want to observe the direct effect of dictionary use on individual words, the more test items that are looked up, the more reliable our data. It is not, however, an easy matter to increase the number of test items that are looked up. One reason is that with so large a text, it is hard to predict which lower frequency words intermediate learners may know. Another reason is that only words which occur once in the text were included as test items; learners are perhaps more likely to look up words which occur more than once in a text. Further, motivation for dictionary use may not always be as straightforward as is suggested by the implicit hypothesis that learners will look up unknown words. Rather, a learner's motives may often be to look up words that they feel they already know partially: either seeking confirmation of their beliefs or taking the route of vocabulary building through focusing on words that will require less effort to learn. In subsequent chapters we will see some ways in which this particular issue is addressed; using a shorter text, with fewer different words, and selecting as test items targeted words that are used more than once in the text.

8.5 Conclusion

The theoretical value of using the self-reporting test in the form of the V_States test has already been considered in Chapter Seven and reported in other contexts, but it is only in this study that its practical application could be evaluated in the context of vocabulary growth resulting from the two learning conditions under investigation. Overall, this study has shown V_States to be a highly effective tool: easy to use, quick to administer, and producing a breadth and depth of valuable data that would otherwise be unavailable. As for the use of the test results to create transitional probability matrices from which predictions of future development could be obtained, this too has been generally successful, both in correctly predicting future vocabulary development and, in this exploratory study, in confirming the identification of problems in other areas of the study.

We should note that although the vocabulary growth through this study does appear quite impressive, it is less so when we reflect that it is the result of the participant devoting around eighty hours of her time to reading, dictionary use and testing. On the other hand, the benefits of extensive reading and monolingual dictionary use are usually recognised as being broader than simply increasing knowledge of the meaning of individual words.

As we proceed to Chapter Nine, we will gain a better understanding of how successful the solutions proposed above may be in improving the effectiveness of the methods employed for researching vocabulary development in the two learning conditions under investigation. It will also give us a clearer picture of what behaviour, in terms of reading, dictionary use, and test-taking, has been a reflection of the idiosyncratic behaviour of this learner and what may apply to other learners of English with similar L2 backgrounds and levels of ability.

Chapter Nine: Sharpening the Tools

9.1 Introduction

The case study reported in Chapter Eight was successful in a variety of ways. The text chosen for the participant to read was of a length and level of difficulty suited to the participant and to the experiment; the main instrument used, V_States (v.3; Meara, 2000), was ideal in many respects for the purpose of rating word knowledge of a large number of items in a reasonable period of time; and the participant was reliable and conscientious throughout the 80 hours of reading and word knowledge rating over a seven week period. The study produced large amounts of valuable data and confirmed that vocabulary growth took place as a result of extensive L2 reading. It also confirmed additional vocabulary growth when the L2 reading was aided by the use of a monolingual learner dictionary. In addition, the study demonstrated the value of using a method and instrument that produces data that can be used to provide predictions of vocabulary growth for the two learning conditions under inspection.

As an exploratory case study, it was also successful in a quite different manner. It successfully revealed three substantial problems with the method, materials, and instruments as they were employed in the study described in Chapter Eight. These were concerned with the participant's accurate recognition of the targeted items, with her understanding of the meaning of "knowing a word", and with the reliability of her rating of word knowledge in the final V_States sessions. In this chapter, we will begin by describing these problems and proposing means by which each of the problems may be overcome in subsequent studies. This will be followed by a description of a further study in which the methods, materials, and instruments employed were adjusted in line with these proposals.

This study will also differ from that in Chapter Eight in terms of the degree of attention paid to the effect of dictionary use on vocabulary knowledge. Here, in addition to looking at general changes in terms of vocabulary growth for the complete set of targeted words, we will draw on the rich reserves of data produced through the studies to look at the effect of dictionary use while reading on the retention of word knowledge for individual words and groups of words. Specifically, we will compare vocabulary development for targeted words that were looked up with that for targeted words that were not, and compare vocabulary development for looked-up nouns, verbs, and adjectives.

9.1.2 Identifying and addressing problems

There are three main issues arising from the study reported in Chapter Eight that need to be addressed: concerning the choice of targeted words, concerning the participant's understanding of what knowing what a word means, and concerning the excessive rating of targeted items as state 3. The main problem with the targeted words used in the study described in Chapter Eight was that a sizeable proportion of unknown words appear to have been misread or otherwise confused with similarly spelled English words that were known by the participant. Examples of such words include *flushed* mistaken for *flashed*, *fond* for *pond*, *puddle* for *paddle* and *sluice* for *slice*. As a simple solution to this, 10 of these words were deleted from the list of targeted words to be investigated in the study, reducing the number of targeted words from 310 to 300. Although a more thorough “cleansing” of the targeted word list of such items might be possible by also excluding words likely to be confused for other words, this was not attempted as the misreading of some words is a natural part of the reading and vocabulary acquisition experience of many Japanese learners of English.

The problem with the participant’s understanding of the meaning of “knowing a word” is that it appears to have been both unclear and unsteady, wavering between the meanings of recognizing a word and knowing a word’s meaning. Both the lack of clarity and the instability of these states for the participant need to be addressed. To ensure that participants in subsequent studies would be clear about the meaning of the four states used in V_States, the wording on the screen for each state was changed to include the words “...what this word means”. State 1, then, changes from “I’m not sure I know this word” to “I’m not sure I know what this word means”, with state 3 changing from “I definitely know this word” to “I definitely know what this word means”. To make sure that understanding of this meaning does not slip from one V_States session to the next, participants should also be reminded what the four states mean before each rating session.

The third problem encountered in the study was that of the participant rating an exceptionally high proportion of the targeted words as state 3, “definitely known”, in the final V_States sessions then being unable to provide the meaning for these words in the Final Meaning Test. This may be associated with the problem described above of confusion about what knowing a word means, and this issue has already been addressed. However, other factors also seem to be involved. Specifically, it appeared that the participant wanted to provide a good result for the study and was not aware that knowledge of the state 3 rated items would be tested following the final V_States session. The same attitude or desire to give a good result may also be found in participants in subsequent studies, with the same danger that too many items would be over-rated as state 3. To avoid this, the participant should be told from the beginning that the study would end with a test in which (s)he would be required to provide the meaning for all items rated as state 3 in the final V_States session. As a further measure, additional reading and V_States sessions were added to the study, penultimate rating session data could be used so as to avoid the effect of the rise in the final V_States session.

A further related point is that state 3 may not be the best place to look for vocabulary development. Rather, if we believe that vocabulary development is incremental in the two conditions under investigation, and if most targeted items are initially unfamiliar to the participant, we might be better advised to look for evidence of L2 vocabulary acquisition in falls in numbers of 0- and 1-rated items, and in combined numbers of items in adjacent states. In this study we will, then, focus mainly on falls in numbers of unknown and hardly known items rather than solely on items that the participant rates as definitely known.

9.2 The study

The case study described below followed largely the same procedures and used the same materials and instruments as those in the study described in Chapter Eight. One intermediate level learner of English repeatedly read a long English text, in this study a total of eight times. Following each reading he rated his knowledge of a large number of lexical items drawn from the text. For the first three reading sessions, no dictionary use was permitted, but from the fourth session onwards for a total of five sessions, he was allowed to use an MLD while reading.

For the study, the following two general hypotheses were proposed:

1. That the participant would demonstrate L2 vocabulary growth as a result of repeated extensive reading.

2. That there would be a clear benefit in terms of L2 vocabulary growth attributable to dictionary use while reading beyond that attained through reading without the aid of a dictionary.

Two more specific hypotheses were proposed regarding sets of words within the set of targeted items. The first of these concerns the effect of dictionary use implied in hypothesis 2 above. The second, following on from research described mainly in Chapter Six, focuses on vocabulary acquisition and part of speech.

3. That targeted items that were looked up would show greater rises in terms of vocabulary growth than targeted items which were not looked up.

4. That there would be greater observable retention of looked-up nouns than looked up verbs.

There were a further three expected outcomes resulting from changes made to this study following problems identified in Chapter Eight. They are:

a) That there would not be very large changes in numbers of any state from one V_States rating session to the next, except perhaps following the first reading of the text.

b) That there would not be exceptionally high rises for state 3 items in the final rating sessions, except following the first reading of the text.

c) That the results for the Final Meaning Test would show a higher proportion of correct answers than that obtained in the study described in Chapter Eight.

9.2.1 The participant

As in the previous study, the Japanese learner of English who took part in this study was a 20-year-old 3rd year student majoring in English at a middle-ranking Japanese university. He had little experience of language use outside formal learning contexts: in class, doing set homework, or preparing for exams. He had been to Britain for 12 weeks six months earlier but apart from this had had little opportunity to use English outside class. His only experience of extensive reading in English was of one graded reader almost two years previously. His knowledge of English would be rated as intermediate, and at the time of the study he had a TOEFL score close to 450.

9.2.2 The reading text

As in the previous study, the text used for reading and as the source for the targeted words was C.S. Lewis's *The Lion, the Witch and the Wardrobe* (1950). As the book proved to be of a suitable level and acceptable length for the participant in the previous study, and she expressed her liking of the story, this confirmed the likely value of using the text again in further studies. Detailed reasons for the initial selection of this text may be found in Chapter Eight.

9.2.3 Targeted word selection

The set of 300 targeted words from the text was largely the same as that used in Chapter Eight (see Appendix 8.1). Each targeted word only occurred once in the text and the words were chosen as items which the participant was unlikely to know, or at least unlikely to know well. Ten words were excluded from the set of 310 targeted words used in Chapter Eight because the Final Meaning Test in that study showed that they had been confused with other words:

bellowing, *device*, *flushed*, *fond*, *forth*, *healed*, *puddle*, *revelry*, *sluice*, *vultures*

The same 20 low frequency words selected from LDOCE3 served as a control in this study as in the previous one. This brought the total number of items in this study to 320.

9.2.4 The dictionary

From the fourth reading onwards, the participant was encouraged to use a monolingual EFL dictionary while reading. As in the previous study, LDOCE3 was chosen for this purpose. The participant did not own a copy of this dictionary but he had some familiarity with using another MLD. He was lent a copy of LDOCE3 to use during the study, with the understanding that he would return it on completion of the study. The reason for stipulating that the dictionary would be returned was to ensure that he kept no record of words looked up in the dictionary, such as highlighting or underlining, as this may have affected his look-up behaviour and retention of looked-up words.

9.2.5 Vocabulary evaluation method

The computer programme V_States (v.03, Meara, 2000) was also used for this case study. The programme presents the targeted items, one by one, in random order and records the participant's responses. For each item that appears on the screen, the participant would rate his knowledge of the word by clicking on one of four buttons on the screen. The labels for this study were changed from those in the program as used in Chapter Eight to stress that the participant should record knowledge of the meaning of the word, as opposed to, for example, degree of familiarity with the word form. The newly labelled states were as follows:

0 – I don't know what this word means

1 – I'm not sure I know what this word means

2 – I think I know what this word means

3 – I definitely know what this word means

As with the previous participant, the rating task took an average of about 4.5 seconds per item, totalling around 25 minutes per testing session for the total of 320 items. Following the final V_States rating session, a final test was conducted to check whether the participant was able to give the meanings of the words he had rated as state 3, definitely known, in the two final V_States sessions.

9.2.6 Procedure

Before reading the text, the participant rated his knowledge of the 320 targeted items. In the following week, he read the text, without referring to a dictionary of any kind. The first reading took him 7 hours to complete, over a period of seven days. The participant then rated his knowledge of the task items again. This continued for two further reading and rating sessions, for which reading times were 7.5 hours and 8 hours.

From the fourth reading onwards, for five sessions, the participant was given a monolingual EFL dictionary to use while reading. He was also given a very brief guide to using the dictionary (see Appendix 8.2). He was advised to limit his dictionary use by aiming to keep within the 7 to 8 hours that it took him to read the text on the first three readings. The participant was also instructed to affix a Post-it tab in the dictionary for each word he looked up during the reading. At the end of the study, following the ninth V_States rating session, the participant was given a Final Meaning Test in which he was asked to give the meaning for all words which he had rated as definitely known in the last two V_States sessions.

9.2.7 Dictionary use

In this section we will report only two aspects of participant dictionary use in this study: the frequency of dictionary use for the five latter reading sessions, and the part of speech of the targeted items that were looked up. In the Results section we will report various other dictionary use data that is more directly concerned with vocabulary development.

As Table 9.1 shows, words in the text were looked up a total of 188 times, with 87 of these look-ups for targeted words.

Table 9.1.
Participant's dictionary use in reading sessions 4 – 8

Reading session	Total look-ups	Not targeted items	Targeted items
4	47	25	22
5	41	27	14
6	41	21	20
7	38	17	21
8	21	11	10
Total	188	101	87

A small number of words were looked up twice over the five reading sessions. The 188 look-ups represents a total of 175 words that were looked up, with 162 of these looked up once and the remaining 13 looked up twice. As for the 87 look-ups for targeted words, they represent a total of 84 words: 81 words that were looked up once and three words that were looked up twice. This leaves 216 targeted words which were not looked up. The figures for numbers of look-ups for targeted words according to word class are as follows: Nouns: 42 Verbs: 31 Adjectives: 15 Adverbs: 1

We will consider further data relating to different aspects of the participant's look-up behaviour as we go on to look at the data obtained through the study that relates more directly to vocabulary development.

9.3 Results

We will now look at various types of data relating to the effect of L2 reading with and without a dictionary on the participant's vocabulary development. We will begin with two sets of data that will help establish the credibility of V_States as the main instrument used in this study: the participant's Final Meaning Test results and his V_States rating for the control items. We will then go on to the results of the V_States rating sessions for the 300 targeted words for the pretest, during the reading only condition, and while reading and using a dictionary. Following this, we will report the use of V_States data to create transitional probability matrices by means of which predictions of vocabulary development in the first condition may be compared with the actual vocabulary development in the second condition. We will then focus on V_States data for the 84 words that were looked up: the states of these words prior to being looked up and their states after being looked up. We will also compare this data with that for targeted words that were not looked up. Finally, we will look specifically at data relating to vocabulary acquisition and word class that may help address questions raised in earlier chapters.

9.3.1 Final Meaning Test

In the Final Meaning Test, the participant was able to give acceptable translation equivalents for 79% of the items that he rated as definitely known in the final V_States session. There was some confusion with morphologically similar words, for example mistaking *rattle* for *kettle*, some lack of semantic clarity, such as confusing *dungeon* with *cave*, and some apparent confusion with Japanese words too, giving *"ugomeku"* (wriggle) instead of *"umeku"* as the meaning of *moan*. Overall, though, this test confirmed that the participant did know the meanings of words which he claimed to know. This enables us to rely on the V_States data for the targeted items with much more confidence than was possible for the data gained through the study reported in Chapter Eight. On the other hand, we should note that we have no indication of the number or proportion of items not rated as definitely known but for which the participant could have provided the correct meaning. In other words, there is no indication of the extent to which this participant may have under-rated his knowledge of targeted words.

9.3.2 Control items

We are further encouraged when we look at the ratings for the 20 control items which were rated alongside the 300 targeted items from the text. Sixteen of the items remain 0-rated throughout the nine V_States sessions, with the other four items showing no overall gain from the first session to the last. Inevitably, nine encounters with these items in the rating sessions, however brief, may increase familiarity with them. However, unless the items are onomatopoeic to some degree, or are encountered outside this study, these contextless encounters should not increase knowledge of the meanings of the items. The participant's rating of these control items confirms that he was recording his knowledge of the meaning of these items rather than degree of familiarity with the word forms. This leads us to assume that he also recorded his knowledge of word meanings for the 300 targeted items.

9.3.3 V_States sessions

We will now go on to the results of the nine V_States sessions: the one session prior to reading, and the eight further sessions following each of the eight readings of the text. These are shown in Table 9.2. As we can see, for 0-rated items, there is a steady fall from t1 to t7, after which the numbers show signs of levelling out. 1-rated items are generally steady from t0 to t6, with items substantially higher for the last two sessions. The numbers of state 2 items rises steadily from t0 to t6, after which they fall for t7 and t8. Somewhat surprisingly, and requiring further investigation, numbers of state 3 items remain largely unchanged from session to session, and in fact fall by over 20% from the first session to the last.

Table 9.2
States of targeted words for each V_States rating task

Learning condition	Session	state 0	state 1	state 2	state 3
Pre-test	t0	196	31	24	49
Reading, no dictionary	t1	192	14	42	52
	t2	174	34	57	35
	t3	166	31	72	31
Reading, dictionary used	t4	142	33	85	40
	t5	124	39	106	31
	t6	107	32	120	41
	t7	82	62	113	43
	t8	86	62	114	38

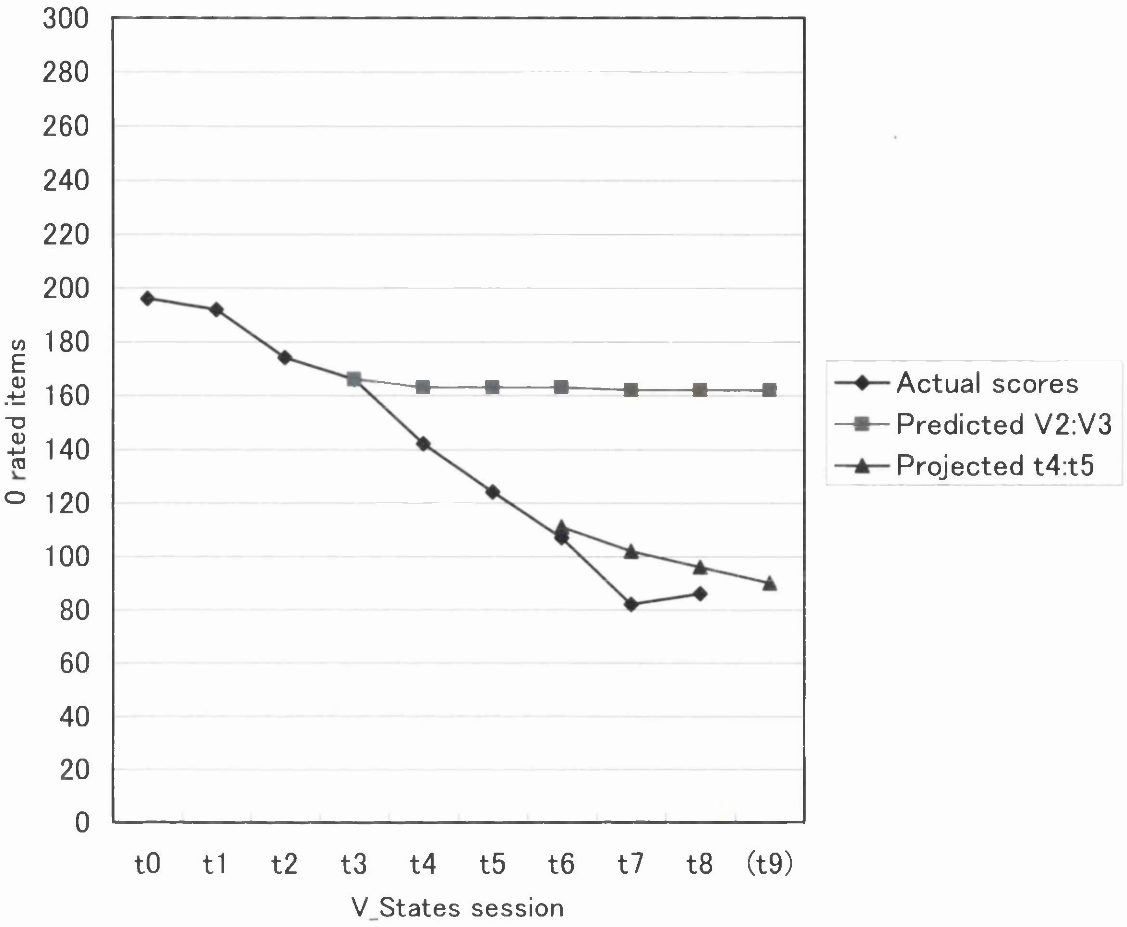
The largely accurate Final Meaning Test results, the almost unchanging V_States rating for the control items, and the lack of major changes in numbers of state 3 items all alert us to two issues regarding the participant's L2 vocabulary development. The first is that this participant is much more accurate, if not conservative, in his estimation of his word knowledge than was the participant in the previous study. Secondly, we are reminded that we should expect incremental vocabulary growth to take place through the learning conditions under inspection; consequently, we may expect to find evidence of this growth in states of vocabulary knowledge other than "definitely known". This is confirmed by the substantial, and steady, fall in numbers of state 0 items and the rise in numbers of state 1 and 2 items over the nine sessions.

9.3.4 Projections of vocabulary development

We will now investigate the use of V_States data to create transitional probability matrices by means of which projections of vocabulary development in one condition may be compared with the actual vocabulary development in a second condition. Although we might expect vocabulary growth or development to be represented as increase in numbers of items that are known, we can see that there are problems in this regard for state 3 items. We could look at combined figures for items rated showing some extent of vocabulary knowledge, i.e., states 1, 2 and 3 combined. It is simpler, however, to show the obverse side of this same data through numbers of items in one single state, and for that reason we will focus on falls in unknown, state 0, items over the nine V_States sessions. Actual and projected number of items in state 0 are shown in Figure 9.3 below, with figures for all states given in Appendix 9.1.

Figure 9.3.

Actual and projected numbers of state 0 items



The projected figures for continued reading without dictionary use, based on t2:t3, show a rapid slowdown in the fall in numbers of state 0 items; fewer and fewer words would become known in subsequent sessions, to the point that between sessions t4 and t5 there may be no drop in numbers of 0-rated items. This contrasts with the continued steady fall in 0-rated items with the new learning condition of reading with a dictionary as more items become at least partly known. The projection based on t4:t5, by which we may compare real and projected figures for the same condition, was made to confirm the accuracy of the predictive power of these V_States generated matrices. Although the real figures' movements from session to session are inevitably less regular than the projected figures, we can see that by t8 the projected and real figures for this condition are very close and are set to become closer. This confirms, again, the value of V_States in this regard.

The tables and graph above provide a description of changes to the vocabulary states for the 300 targeted items from the text. Within this set of items, however, there are two distinct sets of words: words that were looked up and words that were not looked up. We will now look at vocabulary development recorded through V_States for these sets of words.

9.3.5 The effect of look-up on word state

Table 9.4 shows the effect of looking up words while reading for the 84 targeted words that were looked up.

Table 9.4
The effect of look-up on word state

Prior	Number of	State following look-up				Total state
state	items	0	1	2	3	gain
0	67	22	17	24	4	77
1	16	3	0	11	2	12
2	4	0	0	2	2	2
3	1	0	0	0	1	0
Total	87	25	17	37	8	91

Three of the words were looked up twice, so in total there were 87 look-ups for these items. The column labelled "Prior state" shows how the word was rated in the V_States session immediately prior to the reading during which the item was looked up, the four columns labelled "State following look-up" show the state record in the V_States session following the reading in which the word was looked up, and the column labelled "Gain" shows the total number of states by which items in the same prior state rose in the V_States sessions following their look-up.

We will begin by looking at the "Prior state" column as this gives us a useful insight into the state of words that were looked up. As we can see, the vast majority of items that the participant looked up were previously 0-rated: not, as far as he could judge, even partially known to him. This is followed by numbers of words that he had rated 1, as "I'm not sure I know what this word means", while the two higher states 2 and 3 only accounted for about 6% of the total look-ups for targeted words.

If we go on to see what effect looking up a word had on its retention, we can see that for initially 0-rated items, there was an average rise per item of a little over one state. For 1-rated words, the average gain is a little under one state per item. As for words which were in one of the top two states when looked up, the average gain is of under half a state per item: a reflection of the limited room for increasing knowledge of these words.

9.3.6 Looked up words vs. not looked up words

Although the above data is of interest in itself, it becomes more meaningful when we compare data for looked-up words with that for words which were not looked up. In order to compare like with like, I will focus only on words which the participant consistently rated as 0 – unknown – for the first four sessions. For this reason, the figures for looked up items differ from those in Table 9.4. Table 9.5 shows the number of these items in each state together with the state at which these were recorded in the final V_States rating session. Figures are given for words that were looked up, for words that were not looked up, and for the control items.

Table 9.5
Comparison of development for looked-up, not looked-up, and control items

	0-rated for	state at final V_States session				Total
	t0-t3	0	1	2	3	gain
Looked-up words	69	18	18	31	2	86
Not looked up words	84	60	10	13	1	37
Control items	18	17	1	0	0	1

Of the 153 targeted words that were 0-rated for the first four V_States sessions, 69 were subsequently looked up and 84 were not. As we can see for the looked-up items, about a quarter of these remained in state 0, with a quarter rising to state 1 by the final V_States session, and almost half of the items rising to state 2. Very few rose to state 3. As for the words that were not looked up, over two thirds were also 0-rated in the final V_States session. The remaining 24 items were mainly divided between states 1 and 2. Only one of these items rose from state 0 to state 3. To summarise, the most likely outcome of looking up a previously unknown targeted word was for that word to rise from state 0 to state 2. Unknown items that were not looked up typically remained 0-rated.

It is also worth noting that there is a difference between final scores for the targeted words from the text that were not looked up and the control items, which were neither encountered in the text nor looked up. Of the 18 control items that were 0-rated for the first four V_States sessions, 17 remained 0-rated until the end and one was 1-rated. This figure is much lower than that for targeted words that were not looked up. We may assume that the gain for targeted words that were not looked up is attributable to meanings that were guessed from context and retained, or that comprehension of looked-up words aided the comprehension of words in the text that were not looked up.

9.3.7 Comparison of development for nouns, verbs, and adjectives

Finally, we will look specifically at data for looked up targeted words of different word classes. We will again focus on the final states for words that were 0-rated for the first four V_States sessions since these account for the vast majority of subsequently looked-up items from each of the three word classes of verbs, adjectives and nouns. They are shown in Table 9.6.

Table 9.6
Comparison of development for looked-up nouns, verbs, and adjectives

	0-rated for	Final V_States state				Gain
	t0-t3	0	1	2	3	
Nouns	31	7	10	13	1	39
Verbs	22	7	7	7	1	24
Adjectives	14	3	2	9	0	20

Although the numbers of items per word class are rather small, it is clear that there is no major difference between the retention rates recorded for nouns and for verbs. In both cases, there is an average gain per item of a little over one state. Adjectives do seem to be retained rather better, with an average state change of almost 1.5 states per item.

We have looked at various types of data that have been produced using the method of repeated reading of a text with or without the aid of a dictionary and repeated ratings of word knowledge of a large number of words from the text. We will now go on to consider what the data mean and whether through this study we have been able to satisfactorily address the issues raised in the hypotheses proposed earlier in this chapter.

9.4 Discussion

We will begin by returning to the hypotheses and other expected outcomes proposed at the beginning of this chapter. For each of these, we will consider the extent to which the study confirmed the hypothesis and propose explanations for the results obtained. We will also examine and seek to understand the substantial differences between results obtained in this study and those reported in Chapter Eight. Finally, as we review the achievements of this study, we will summarise the findings that have been made and identify issues that still need to be addressed.

9.4.1 Hypotheses addressed

Of the four hypotheses proposed with regard to this study, two were general hypotheses regarding anticipated vocabulary development in the two main learning conditions under investigation. One is that the participant would demonstrate L2 vocabulary growth attributable to repeated extensive reading. The study's findings supported this hypothesis; especially for the first of the three reading sessions for which dictionary use was not allowed, considerable vocabulary growth was recorded, although most of the word knowledge gains were partial and for items previously rated as unknown. While there is clear evidence of vocabulary growth, the data provide no indication as to the nature of the recorded vocabulary growth: whether, for example, words that were encountered in isolation in the first V_States session and rated as unknown were subsequently recognised in the context of the reading text; whether the reading text provided a context in which previously unknown words' meanings could be at least partially understood; or whether the items which showed gains were mistakenly believed to have been understood but were actually misunderstood. We may expect the total vocabulary growth recorded here to be attributable to a combination of these three factors, but it is not possible from the data to determine the influence of each.

The second general hypothesis was that there would be a clear benefit in terms of L2 vocabulary growth attributable to dictionary use while reading beyond that attained through reading without the aid of a dictionary. This, too, was confirmed, as is shown in Figure 9.3. Just as the benefit of repeated extensive reading without dictionary use was beginning to wane, the introduction of permitted dictionary use had the effect of accelerating vocabulary acquisition once more, to the extent that half of the items that are 0-rated at this point rise above this level by the time of the final V_States session. Unlike the low level of return in terms of vocabulary development of repeated reading alone, the rate of vocabulary growth through reading with dictionary use was largely sustained over the five reading with dictionary use sessions. It should be pointed out, however, that the gains made from the fourth reading session onwards cannot be attributed to dictionary use alone, since the participant did not use the dictionary in isolation. Rather, we need to recognise that the participant would rely on the combined resources of the written context for the word in the text and the information contained within the dictionary entry.

A further set of hypotheses was concerned with groups of words within the total set of 300 targeted items. The first of these was that looked-up targeted words would show greater vocabulary growth than targeted words which were not looked up. As we can see from the data presented in Table 9.4, most looked-up words were previously rated as unknown, and Table 9.5 shows that knowledge of previously unknown targeted words that were looked up typically rose by one or two states, with the majority rising by two states. This compares with an average growth rate of less than half this for initially unknown targeted words that were not looked up. It appears that although a reader can make a guess at the meaning of unknown words from the context, in the case of the participant in this study, guessing alone was usually insufficient for the word to be learned and the meaning even partially retained. The participant seemed, usually, to need confirmation of word meaning through dictionary use before he would feel that he knew the word better than he did prior to encountering it in this study. There appears to be a reciprocal relationship between extent or state of word knowledge and confidence about the accuracy of that word knowledge. This means that a learner may invest more effort in committing to memory a word for which they are sure of the meaning (or, for example, collocations). In this regard, especially for language learners who have little experience of extensive L2 reading or of monolingual learner dictionary use, we may expect a hierarchy among L2 vocabulary learning resources both in terms of learner confidence and in actual rates of acquisition. Guessing word meaning from context may be the resource that inspires least confidence among such learners, followed by MLDs, with bilingual dictionaries inspiring most confidence in having understood the meaning of the L2 word in question.

Also in Table 9.5 we can see the low levels of vocabulary growth for unknown words that were not looked up, and that may be attributable in part to dictionary use. For this learner at least, this suggests that while the effect of multiple reading of a long text may reactivate forgotten word knowledge, beyond the first couple of readings it appears to do little in increasing knowledge of previously unknown words. In other words, there was little evidence of an anticipated ripple effect by which words in the textual vicinity of looked-up words in the text would also become better known. This may not be surprising when we consider that for many unknown words a single written context may give few clues as to the word's meaning, and that this will be especially true for a learner with a limited vocabulary for whom the meaning of the context of some unknown words is also largely unknown.

One further hypothesis about groups of words within the total set of targeted words concerns parts of speech. Our hypothesis was that there would be greater observable retention of looked-up nouns than looked-up verbs. In fact, as we can see from Table 9.6, retention rates for looked-up nouns and verbs were very similar, with adjectives showing slightly higher retention rates than either of these. This result may reflect the complex combination of factors involved in comparing retention of different parts of speech in this learning context rather than any reflecting intrinsic word difficulty for nouns, verbs, or adjectives. Factors may include typical intrinsic word difficulty for the part of speech, but also involved will be typical accessibility of word meaning from the written context, typical comprehensibility of the dictionary entry for the part of speech, and typical memorability of the word meaning from these two contexts. This last factor may include clear collocations for the targeted words which would create or strengthen links for those words within the mental lexicon. This may have contributed, for example, to the advantage shown for adjectives in this study, since adjectives impart and derive meaning largely by association with limited sets of nouns. Perhaps most important, however, is the recognition that the set of looked-up items for each part of speech may not be a representative sample for that part of speech. Within each word class some words will be easier to guess from context, some will be easier to understand from definitions or example sentences, and some will be more memorable. The participant's look-up behaviour may produce a bias in favour of one part of speech over the others. Although this means that we cannot draw definite conclusions about the effect of intrinsic or extrinsic factors on the learnability of words in different word classes, we can say that this participant's look-up choices, combined with the factors listed above, resulted in proportionately greater learning of adjectives than nouns or verbs.

9.4.2 Expected outcomes

There were three further expected outcomes proposed for this study that were related directly to problems identified in the study reported in Chapter Eight and changes made to address these problems. The first of these was that there would not be very large changes in numbers of any state from one V_States rating session to the next, except following the first reading of the text. As Table 9.2 shows, this is true for state 0 and state 3 items, with rises and falls between sessions generally around 10-20% and on no occasion exceeding 35%. This is in marked contrast to the large rises and falls between sessions reported in Chapter Eight: a fall of over 50% in state 0 items, and a rise of over 50% for state 3 items. In this study, there are more sudden rises and falls for the two middle states than for states 0 and 3, with numbers of items falling by half in one case and almost doubling in two other instances. Two of these cases do occur, as predicted, between the pre-reading session and the session following the first reading, with the fall in state 1 items matched by a rise of a similar amount in state 2. The two central states, expressed as *"I'm not sure I know what this word means"* and *"I think I know what this word means"*, are the most vague of the states and are also the states for which the participant's rating is most volatile. This suggests that there may be a case for collapsing the results for the top two states and for the bottom two states. Although this may appear to reduce the sensitivity of the instrument, it may well increase its reliability.

The second of the three expected outcomes concerning this study is that there would not be exceptionally high rises for state 3 items in the final rating sessions. In fact, the number of state 3 items stayed remarkably steady from the third V_States session (t2) onwards, with the figure rising by only two items in the penultimate session and even falling slightly at the final session. This compares with rises of 24 points for each of the final two sessions in the study reported in Chapter Eight. This may provide confirmation of the effectiveness of changing the on-screen wording for the four states in the latter study, together with repeated announcements that the participant's knowledge of state

3-rated items would be tested at the end of the study. It may also be a further indication of the difference between the learners who took part in the two studies. We will consider this issue below.

The third expected outcome, that the results for this study's Final Meaning Test would show a higher proportion of correct answers than that obtained in the study described in Chapter Eight, was also confirmed, with close to 80% of the 38 items rated as state 3 correctly identified in the Final Meaning Test. This compares with just over 50% for the 89 state 3-rated items in the study in Chapter Eight. Again, this may reflect the effect of the repeated announcements during the study reported in this chapter that the final V_States sessions would be followed by a test of word meaning knowledge for items that were rated state 3. In addition, as the two participants knew each other, it is likely that the participant in this study would have been warned about the Final Meaning Test and so remained conservative in his estimation of his knowledge of the targeted items. A third factor that should be considered is simply that the participant in this study was, generally, a more conservative rater of his knowledge of the items than was the participant in Chapter Eight.

9.4.3 Comparing the two case studies

As we compare the results of the two studies, we will see how many ways there are in which the two participants differ: in their dictionary use; in their rating of the control items, in their Final Meaning Test scores, and in their rating of the targeted items. We will refer to the participant for the study reported in Chapter Eight as "C" and that in this chapter as "S".

We will begin by considering the participants' dictionary use. There was little difference between the average number of lookups for the two participants: an average of just over 34 lookups per reading session for C and of approaching 38 lookups per session for S. Overall, however, S did look up considerably more words than C (188 as compared to C's 137) but this was mainly a consequence of S having one more reading session with dictionary use than C. Where there was a marked difference between the two participants was in the proportions of looked up words that were targeted words. Only 24%, or 34, of C's look-ups were for targeted items, while for S a much higher proportion of looked up words were targeted items: 46%, or 87 look-ups. With S looking up a considerably larger number of targeted items, we would expect him to show greater gains than C in word knowledge recorded using V_States through the study. This is not the case, and the scores for control items and Final Meaning Test results may help us understand why this is.

For the 20 control items, there was a large difference between the two participants. For C, in the first V_States session 17 of the 20 items were 0-rated but most ended up rated 1, 2, or 3 by the final session, with only 4 items still 0-rated. For S, 16 of the 20 items remained 0-rated throughout the nine V_States sessions, with the remaining 4 items also, on average, showing no movement. This suggests that S's rating standard was unchanged through the study, while for C there are two possible interpretations of her progressively higher rating of the control items. One is that she was rating familiarity with the items; through the repeated V_States sessions it may be that these items became increasingly familiar. In contrast, it is hard to imagine how knowledge of meaning might have increased as a result of meeting these words in isolation. A second interpretation may be that C's rating of all the items was increasingly lenient, session by session, and that both the targeted items and the control items benefited from this.

In the Final Meaning Test of state 3-rated items from the final V_States session, C gave acceptable meanings for only 53% of the items as compared to 79% for S, with this figure for S rising to 85% for items rated as state 3 in the penultimate V_States session. In a large number of cases in which C was unable to give correct meanings for these items, the cause seems to have been mistaking unknown words for known words with similar spelling or, more specifically, for words which would have similar phonological representation in Japanese. This suggests that C may have a systematic problem as far as her recognition of English words is concerned. The data strongly suggest that she may be storing these words in a way which relies heavily on Japanese phonology. (See Nakamura, 2001 for a discussion of the implications of this kind of behaviour.) For S, the small number of items for which he gave wrong answers may be attributed to a variety of causes.

For the targeted items, there are also large differences between the two participants in the V_States rating sessions. These may be seen in Table 9.7. Although C starts with higher numbers of state 0 items and lower numbers of state 3 items than S, this situation is reversed as the studies progress, until by t7, the eighth rating session, C has less than half the 0-rated items of S, and more than twice as many 3-rated items. For C, there are greater rises and falls in items in particular states than for S, with numbers of items in a particular state sometimes halving or doubling from one session to the next. For both participants, numbers of 0-rated items fall through the study, although for C this is from

245 items at t0 to only 37 items at t7 while the figures for S are from 196 to 82. Numbers of state 1 items appear volatile for both participants, with both more than doubling by the end of the study. State 2 items rise through the study for both participants although by a greater amount for S. As for numbers of 3-rated items, for C these rise from 9 items to 90 items in her final V_States session, while for S there is little overall change, with the number of items falling from 49 items to 38 items.

Table 9.7
Comparison of V_States scores of targeted words for C and S

Session	state 0		state 1		state 2		state 3	
	C	S	**C**	S	**C**	S	**C**	S
t0	**245**	196	**41**	31	**15**	24	**9**	49
t1	**116**	192	**134**	14	**45**	42	**15**	52
t2	**169**	174	**69**	34	**46**	57	**26**	35
t3	**142**	166	**81**	31	**57**	72	**30**	31
t4	**112**	142	**80**	33	**84**	85	**34**	40
t5	**86**	124	**102**	39	**80**	106	**42**	31
t6	**59**	107	**113**	32	**72**	120	**66**	41
t7	**37**	82	**108**	62	**75**	113	**90**	43
t8	**X**	86	**X**	62	**X**	114	**X**	38

Judging by the V_States scores, C appears to gain much more word meaning knowledge of the targeted items through reading and dictionary use than S, despite the fact that she looked up far fewer targeted items than S. However, as her results for the control items and for the Final Meaning Test show, C consistently overrated her knowledge of a large number of the targeted items, so we cannot take her V_States scores at face value. We do, however, need to recognise that she does seem to have made large gains, at least for state 3 items; in the first V_States session she only rated 9 items as definitely known but in the Final Meaning Test was able to give acceptable meanings for 46 state 3-rated items. As we have noted, S's rating of word meaning knowledge is much more conservative, and his greater dictionary use is reflected in the substantial rise in state 2 items.

In terms of methods and materials, the two studies differed in terms of the numbers of targeted items (310 for C, 300 for S), in the number of easily confusable items (10 fewer for S), in ignorance (C) or awareness (S) of the impending Final Meaning Test, and in the number of reading sessions (7 for C, 8 for S) and V_States sessions (8 for C, 9 for S). The wording of the meaning states for the V_States program was changed too, and verbal explanations were added, so that S was constantly reminded that the word knowledge he was rating was word *meaning* knowledge. These changes would undoubtedly be factors contributing to making S more conservative than C in his rating of the targeted items. It seems unlikely, however, that these changes alone would account for the considerable differences between the two participants in their V_States rating of both targeted items and control items, and in their Final Meaning Test scores. Nor would these changes be likely to result in S's greater dictionary use, or in his greater focus on the targeted items. Rather, we have to recognise that this data suggests that there are important differences between these two participants in the process by which they learn words, and in how these words may be retained as part of the learner's mental lexicon.

9.4.4 Implication for further research

The comparison of these two studies has brought to light three issues that merit further attention. The first is that even participants who are similar in a large number of ways, such as native language, L2 proficiency levels, and foreign language learning experience, may behave very differently with regard to dictionary use and in how they learn foreign words. Given the changes made in how the studies reported in Chapters Eight and Nine were conducted, it was not possible to judge the extent to which differences in results were attributable to these changes or to more fundamental differences between the two participants. There may be value in investigating this further by conducting parallel case studies with two participants following identical experimental procedures.

The second issue is that of text difficulty. The text used in the two case studies reported here was selected as a text that would be largely comprehensible without the use of a dictionary. This has been worthwhile as it provides an environment in which we can investigate the acquisition of words that are guessed from context with that for words that, in addition to being encountered in the text, are looked up in a monolingual learner dictionary. There is, however, justification for using more difficult texts, for which near-total comprehension will not be achieved without extensive dictionary use.

Intermediate level language learners may often encounter such texts both inside and beyond the classroom and it is with these texts rather than, for example, graded readers or books intended for children or teenagers, that they are likely to make most use of their dictionaries.

The third issue concerns the matter of repeated reading of a text. Although we may encounter instances of repeated reading of long texts outside experimental conditions such as those in these studies, they are rare and are not typical of most language learners' foreign language reading experience. Although we must recognise that, in terms of reading and dictionary use environments, we cannot expect to produce experimental learning conditions identical to those that form part of the participants' normal language learning experience, we may be able to come closer to this by focusing on texts that language learners may typically read and re-read a number of times.

9.5 Conclusion

As we have noted, the study described in this chapter has been surprisingly successful in fulfilling the purposes for which it was conceived and conducted. The participant showed himself to be capable of reliably assessing his word meaning knowledge, with confirmation of this provided by Final Meaning Test scores, control item scores, and V_States scores for looked-up targeted words. With a reliable participant such as this, we have been able to evaluate both the effectiveness of the instruments employed for the study and the effect of the different experimental learning conditions on the participant's acquisition of previously unknown L2 word meanings. V_States, the main instrument employed, fulfilled its purposes of providing a sensitive measure of vocabulary knowledge, and of providing a longitudinal record of L2 vocabulary development for a large number of targeted items. Even more importantly, it produced transitional probability matrices by which accurate predictions could be made of continuing vocabulary development for the targeted words in the different learning conditions under investigation. It was by this means that it was possible to compare the effect of two, or even three, learning conditions on one participant's vocabulary knowledge development.

Following the study described in Chapter Eight, all the changes made to how the study was conducted appear to have been justified, with this possible exception of increasing the number of readings and V_States sessions in this study. As the participant in this study proved to be so reliable, if not too conservative, in his assessment of word meaning knowledge, and showed no sudden rises for items in states 2 or 3, this additional insurance measure proved to have been unnecessary.

Investigation of the results of this study, and comparison with those reported in Chapter Eight, have suggested two main ways in which further research in this field may be conducted. The first is that two parallel studies with comparable participants may confirm issues that these two studies have brought to our attention but for which no definite conclusions may yet be drawn. The second is that there may be various advantages in using shorter and harder texts: texts which typically call for intensive dictionary use. These are issues which we will address in Chapter Ten.

Chapter Ten: Shorter, Harder Texts; Listening and Reading

10.1 Introduction

The case study in Chapter Nine, with a participant who was a generally reliable, if conservative, judge of his own changing L2 word knowledge, was successful in investigating the value of repeated reading of a text with and without the use of a monolingual learner dictionary. It also confirmed the value of V_States (Meara, 2000) as an instrument for recording and comparing changing word knowledge for large numbers of items in the two learning contexts under investigation.

In some respects, the L2 text used in the studies reported in Chapters Eight and Nine also served its purposes well. The text was *The Lion, the Witch and the Wardrobe* (Lewis, 1950), a 150-page, 40,000 token, fantasy story for older children. Only a small proportion (2-5%) of the text was estimated to have been unknown to the participants prior to the study. This, and the fact that the story was written for children, meant that the text could be read and largely understood by intermediate level EFL learners without too much difficulty. At the same time, as it was a long text, it still was able to yield over 300 items which would be suited for use as test items: words in the text that occur only once and most of which would be unknown to the participants. Also, importantly, it was a text that could sustain participants' interest through multiple readings, since this was a requirement of the experimental procedure employed in these studies.

Despite the above qualities, there were three main ways in which this text, and text type, was not ideally suited for the purposes to which it was put in these studies:

1. While the main focus of the studies is the comparison of L2 vocabulary learning through reading with and without the aid of a dictionary, intermediate level readers of this L2 text should be able to gain an understanding of the majority of the text without the use of a dictionary. In other words, this is a text for which it could be argued that extensive dictionary use is largely unnecessary.

2. Although the text used may have been better able to sustain the interest of the readers for multiple readings than many other books, the repeated reading of whole books (seven or eight times) is not part of the usual reading experience of the majority of young language learners, whether reading in their L1 or in the L2.

3. From an administrative perspective, it is difficult to find participants who are willing and able to assist in a study for which they are required to give up as much as 60 or 70 hours of their time, and willing to read a whole book six or seven times.

For these reasons, a preferred type of text might be one that is shorter but with a greater proportion of unknown words. It would also be a type of text for which repeated reading is widespread both among people for whom the text is in their mother tongue and among language learners. In the light of these considerations, a set of popular song lyrics, together with the songs themselves, was selected as the text to be used in the two case studies reported in this chapter. Reasons for this specific choice are as follows:

a) The activity of listening to popular songs repeatedly is familiar to young people in the culture of the participants, both for songs in their L1 and for songs in English. Coupled with this, L2 listening is often accompanied by careful reading of the printed lyrics of the songs. Especially for favourite groups or singers, this might involve using a translation of the lyrics or intensive use of a dictionary to gain a fuller understanding of the songs.

b) Learners often choose to, or are obliged to, read texts that are considerably beyond the level at which near-total comprehension would be possible without the use of a dictionary. The lyrics of popular songs in the L2 are an example of this type of text. With texts such as these, dictionary use would naturally accompany reading the text.

From the perspective of case study design, there were further reasons for using this type and size of L2 text:

c) Each session of listening to the songs, reading the lyrics and trying to understand the songs by inferring from context or using a dictionary, together with the vocabulary rating sessions, would take the participants no more than six hours, totalling around 40 hours in all for the seven sessions. This is considerably less than the 8-10 hours required per session for the story reading in previous studies, and makes it possible for each session in this study, including a rest day, to be completed in six days.

d) With a smaller body of language than the previous text, as the songs contained a greater proportion of unknown words, it should still be possible to identify 300 suitable words as targeted items, with most of these words unknown or only partly known to the participants.

We will now go on to describe and explain the administration of the studies, including the reasons for conducting two parallel studies rather than one individual study. We will also explain about the specific texts chosen for these studies.

10.2 The studies

Although the case studies reported in this chapter use materials that are different from those used in the studies reported in Chapters Eight and Nine, the instruments employed are the same and the procedures only differ to the extent that they are adapted to suit the materials used in these studies, and that they were conducted with parallel studies. In each of the two studies, one intermediate to advanced level learner of English listened to and read the lyrics of 47 English songs over a period of four days. Following each listening and reading session, the participant rated the state of her knowledge of the meaning of 300 words from the songs. This procedure was repeated a total of seven times: three listening and reading sessions without the use of a dictionary and four listening and reading sessions in which participants had the help of an MLD.

The following hypotheses were proposed regarding these studies:

1. That L2 vocabulary growth would be evident in both learning conditions.

2. That there would be evidence of L2 vocabulary growth in the listening and reading with MLD use condition beyond that achieved through listening and reading without the use of a dictionary.

3. That during the with MLD sessions there would be gains in reported word knowledge meaning, both for targeted words that were looked up and for targeted words that were not looked up.

We will begin by explaining the reasons for conducting the parallel studies. Following this, we will describe the two participants, explain the procedures by which the studies were conducted, give more details about the text used in the studies, and give an account of selection of the targeted items. We will then go on to report participants' dictionary use and results from the studies.

10.2.1 Parallel studies

In a case study with an individual participant it may be difficult to determine the extent to which any outcomes are attributable to factors intrinsic to the study itself, such as the text, targeted words, or procedures employed, and how much may be attributable to the particular participant's behaviour or abilities. This is apparent from the two studies in Chapters Eight and Nine, in which the two participants varied widely in their dictionary use, in their changing reported knowledge of the targeted items, and in the accuracy of their self-assessment of knowledge of these items. Those studies were not, however, identical in terms of items targeted, in instructions to the participants, or in participants' awareness of how the study was to be administered. All these factors may have affected the participants' reading, their dictionary use, the nature of their L2 vocabulary development, and their rating of this vocabulary development. As a result of this, all we can say with confidence is that observed differences in dictionary use, in rating of knowledge of the targeted words, and in Final Meaning Test scores for these studies appear to be beyond what we might expect as a result of differences in how the two studies were conducted. With parallel studies, conducted with the aim of being identical for the two participants, we would be able to say with much more confidence that any observable differences may be attributed to differences between the two participants. By conducting parallel studies, we may also gain a clearer understanding of the relationship between L2 proficiency, dictionary use, and vocabulary development.

10.2.2 The participants

The two participants for these studies, as for the previous case studies, were in the third of four years as students majoring in English at the same middle-ranking Japanese university. As before, the participants in these studies were asked to help because their conscientious attitude as students suggested that they would also be reliable participants for these long and time-consuming studies. In common with the participant in Chapter Nine, these participants had spent 12 weeks studying in the UK almost one year earlier. Those in the present studies, who we will refer to as E and N, differed from those in the two previous studies in that they were both actively involved in using English outside the classroom and joined various extra-curricular English-related activities. One of them, E, also had an exceptionally good general command of English, reflected in a TOEIC score of over 800 at the time. Her TOEFL score would have been in the region of 550. She was also a keen reader of books in English. The English proficiency level of N was comparable to that of the participants in the two previous studies. Her latest TOEFL score at this time was in the region of 450.

10.2.3 Procedures

The focus of these studies is to compare the effect of the following two learning conditions on the English vocabulary of an individual foreign language learner:

Sessions 1 – 3:	Listening to and reading the lyrics of English songs without dictionary use.
Sessions 4 – 7:	Listening to and reading the lyrics of English songs with MLD use.

As with the studies reported in Chapters Eight and Nine, an iterative approach was adopted to do this, with each participant going through repeated sessions of listening and reading followed by a rating session for knowledge of large numbers of lexical items from the song lyrics. As in the study in Chapter Nine, the participants had to rate their knowledge of the meaning of each of the targeted words, and the control items, as being in one of the following states:

0 – I don't know what this word means

1 – I'm not sure I know what this word means

2 – I think I know what this word means

3 – I definitely know what this word means

The computer programme V_States (v.03, Meara, 2001) was again used to implement this repeated vocabulary rating and to create transitional probability matrices with which to make forecasts of subsequent vocabulary development within the same learning condition. This would render possible the comparison of these projections for the first learning condition with actual results for the second learning condition.

For a total of seven sessions, the two case study participants were required to listen to four albums of popular songs sung in English, 47 songs in total, while reading the printed lyrics of the songs. Each complete session comprised four days of listening and reading (one album per day), one rest day in which there was no contact with the listening and reading materials, and one day in which participants rated their knowledge of 301 words from the text using V_States. In total, the participants were each required to devote a total of about 40 hours, over a period of 42 days, to listening, reading, inferring from context, dictionary use, and rating of knowledge of the targeted words.

During the first three sessions, the participants were allowed time, up to about thirty minutes per day and album, to try to infer the meaning of unknown words or parts of songs from the surrounding context. For the subsequent four sessions, the same amount of time was allowed for using an MLD to look up words from the text. This would be done either after listening to individual songs or after listening to a complete album.

At the end of the study, a final test was conducted in which the participants were required to give the meanings of the words they had rated as definitely known in their last V_States session. This Final Meaning Test for each participant was administered thirty minutes after this final V_States session. While this final test was a complete surprise for the participant for the first of the series of case studies, for participants in subsequent studies, including the two participants in the studies reported in this chapter, this Final Meaning Test was expected. They were told about it as part of the explanation of the procedures of the study, and undoubtedly heard about it in more detail from the participants in previous studies.

10.2.4. Materials

Four albums of popular English songs, on CD and with printed lyrics, were the listening and reading materials selected for these case studies. The specific albums were as follows:

Delirious: *Cutting Edge* (1997, Furious? Records) – two albums

Joni Mitchell: *The Hissing of Summer Lawns* (1977, Asylum Records)

Elvis Costello: *Painted from Memory* (1998, Mercury Records)

The particular albums were chosen for their extensive use of vocabulary that was unlikely to be known by the participants, for the clarity of their sung lyrics, and for their relative obscurity, at least from the perspective of 20-year-old Japanese students. *Cutting Edge* is an example of contemporary Christian music; that much of the vocabulary of the songs is specifically related to Christianity, which for most Japanese is perceived as a Western and largely unfamiliar religion, suggested that it would provide a reasonable number of unknown items. *The Hissing of Summer Lawns* is a collection of songs mostly focusing on life and art in North America during the 1970s. Its subject matter and oblique, referential style suggested that many of the words and references would be unknown to the participants. *Painted from Memory* is a more recent collection of songs written about 'mature' topics such as separation and the disintegration of relationships; it was hoped that the vocabulary used to describe these themes would be largely unfamiliar to the participants.

A further factor for choosing the above collections of songs was that, unlike many more popular contemporary groups or singers, those chosen were unlikely to be known to the participants. This proved to be so: none of the performers had been heard of, and only one song had been encountered before: one song by Delirious that both had heard in a church attended when studying English in England one year previously.

10.2.5 Targeted items

A total of 301 targeted words were selected from the song lyrics for these studies (as listed in Appendix 10.1). The main criterion for selecting the targeted items from the lyrics of the songs was that most of them would be unknown to the participants. With one participant, this was definitely the case, with 75% of targeted items being rated as either not known or not sure. For the other participant, the figure for these two states was closer to 50%. Certainly, as the figures for dictionary use presented below confirm, the targeted items were very largely the words that the participants chose to look up. A range of factors may be involved in determining which words would be likely to be known. These include whether words appearing in the participants' school textbooks and word frequency. The banded word list JACET 8000, compiled to take account of both these factors, provides a useful indicator of words likely to be known to the participants. Just over a third of the targeted items were listed among the first 3,000 words, with a further third found in the 4,000 to 7,000 word bands, and the final third either in the 8,000 word band or not listed among the 8,000 words at all.

It is not a straightforward matter to relate likely participant knowledge of the targeted words with their ranking in a word list. As a general guide, however, we might expect the participants to know the majority of words in the first 3,000 words of JACET 8000, to know some of the words in the next 4,000, and to be unlikely to know many of the words which are either in the 8,000 word band or not listed at all. In terms of depth of vocabulary knowledge, we might expect the participants to have fuller knowledge of the more frequent items, and more partial or hazy knowledge of less frequent items. Similarly, we might expect participants to be generally more confident about their knowledge of more frequent words and less confident regarding knowledge of lower frequency words.

A further indicator of numbers of items in the chosen texts likely to be unknown to the participants was obtained by asking two students with similar backgrounds to the participants in these studies to highlight unknown words in the texts. Of the total of 1641 word types in the texts, one student marked 256 word types (16%) as unknown and one marked 152 word types (9%) as unknown. These figures, together with ranking for the targeted words in JACET 8000, suggest that although it would be unlikely that all 301 targeted items would be unknown to such participants, there would still be a sufficient number of items which the participants would not know and would be likely to look up.

An additional condition for targeted item selection, as some balance in the number of items encountered each day was desired, was that the number of targeted items per album should be no fewer than 60 and no more than 100. There were no other restrictions. Given the relatively small size of this text, and the customary repetition of lines or phrases in songs, it was unrealistic to make the stipulation applied in the previous studies that only words occurring only once should be targeted items. The vast majority of the items do in fact only occur once, about a dozen occur twice and just a handful more than twice. Where items do occur more than once, these are almost all in lines or phrases that are repeated in the songs. In the two previous studies, a further stipulation applied was that the items should be single-sense words. Again, because of the relative small pool of words suitable as targeted items, this requirement was not applied in these two studies, with the result that some polysemous words were included in the set of targeted items targeted from the text.

A further 30 items were added to the test to serve as a control. These were selected from low frequency words found in the Longman Dictionary of Contemporary English, 3rd Edition (LDOCE3; Summers, 1995). Wherever possible, one word was chosen from those alphabetically located in the dictionary between the entries for every tenth and eleventh targeted item. The control items were added with the same criteria as for the other test items: that they would be probably not be known to the participants, would be low frequency words, and with the additional consideration that, morphologically, they should not stand out from the test items selected from the song lyrics.

10.2.6 Dictionary use

From the fourth reading and listening session onwards, the participants were allowed to use a monolingual learner dictionary to look up words from the lyrics of the songs. As in previous case studies, the *Longman Dictionary of Contemporary English, 3rd Edition*

(Summers et al., 1995) was chosen for this purpose. They were also given basically the same brief guide for dictionary use as the participants in the two previous case studies (See Appendix 8.2).

Because the activity in these studies involved listening to the songs as well as reading the lyrics, dictionary use would have been difficult if attempted while the songs were playing. The participants were advised to either stop after each song and look up unknown words or wait until the album had finished and go through all the lyrics for that album.

When the participants looked up a word, they would place a Post-it tab at that point on the page of the dictionary to indicate which word they had looked up. They used different colour tabs to differentiate between words in the inner and outer columns of the page. During the 20-30 minutes of the V_States rating sessions, the participant would return the dictionary so that the tabs could be removed from the dictionary and the looked up word written on each tab. The unmarked dictionary would then be returned to the participants, ready for use in the following listening and reading session.

We will now go on to report the following information about the words looked up by each of the participants:

i) How many words were looked up, when they were looked up, and how often they were looked up,

ii) How many of these looked-up words were targeted items used in the V_States sessions,

iii) The extent of participants' reported prior knowledge of looked-up targeted words.

First, Table 10.1, below, shows the use made by the participants of their MLD during the four sessions in which dictionary use was allowed.

Table 10.1
Participants' dictionary use

Participant	N	E
No. of look-ups in session 4	48	61
No. of look-ups in session 5	51 (34 new)	43 (19 new)
No. of look-ups in session 6	42 (36 new)	52 (16 new)
No. of look-ups in session 7	40 (19 new)	32 (9 new)
Total no. of look-ups	181	188
Total no. of words looked up	134	109
Average no. look-ups for looked up words	1.35	1.72
Targeted items looked up	100 (33.0% of total)	92 (30.4% of total)
Looked up words which are targeted items	74.5%	84.5%

As we can see, the two participants did not differ greatly in the overall number of times they consulted their dictionary, although E's dictionary use was more erratic from session to session, ranging between 32 and 61 look-ups per session as compared with between 40 and 51 for N. Where the participants did differ markedly is in the numbers of words that they looked up more than once; this is reflected in Table 10.1 both in the totals for numbers of words looked up and in the average number of look-ups for looked up words.

Overall, E looked up almost 20% fewer words than N, but looked up many more words two or more times. This difference is amplified in the large difference between the participants in the numbers of words they looked up for the first time in sessions 5 – 7, shown as new in Table 10.1. In each of these sessions N looked up around twice as many words for the first time as E. In the Results section we will investigate what effect the repeated looking up of the same words, and the lower numbers of new items looked up, may have had on E's vocabulary development. The two participants looked up largely similar numbers and proportions of targeted items; for both N and E, at least three-quarters of looked up words were targeted items. This provides us with an ample number of looked up words whose development can be tracked through V_States sessions along with that for targeted words that were not looked up.

We will now address the question of how well the participants knew the words they chose to look up: whether they typically looked up words which they felt they already knew at least partially, or whether they tended to look up words for which they already had some idea about the meaning. V_States session t3 records participants' assessment of their own knowledge of the targeted words prior to the four listening and reading sessions in which dictionary use would be allowed. This provides data for the targeted words that each participant would subsequently look up, as shown in Table 10.2.

Table 10.2

State at t3 of words looked up in sessions 4 to 7

	Number of items in each state at t3	
	N	E
0	62	68
1	28	7
2	7	13
3	4	9

Between two thirds (for N) and three quarters (for E) of words that would be looked up were previously rated as 0, not known. N also looked up quite a large number of items which she had rated as 1 (*I don't think I know the meaning of this word*). This still leaves quite a large number of items that the participants looked up despite their being fairly confident about knowing them at t3: a little over 10% of looked-up targeted items for N and over 20% for E.

There may be at least two related reasons for participants' looking up words which they had already rated as known. One possible reason was that participants only knew the word as a loan word in Japanese and were not sure whether the meaning would be the same in English. Another reason is that while participants knew one sense of the word, they felt that the word must be polysemous since the apparent sense used in the songs was unknown to them. A further reason may be attributed to the volatility of much vocabulary in the L2 mental lexicon, as may for example be inferred from the data in Krantz' study (1990), in which initially known items were later not recognized. In these studies, some words that participants felt they knew at t3 may not have been recognized as known words two or more sessions later.

10.3 Results

We will now consider the results obtained through these studies. These studies depend on the central use of V_States (v.03, Meara, 2000), an instrument that relies on the stable and accurate assessment by participants of a large number of targeted words. Our first task is to confirm, following on from the reliable results reported in Chapter Eight, that the data obtained by this means would be trustworthy. To do this, we will look at data from the two measures adopted to address this issue: control item rating and Final Meaning Test scores. The participants' ratings of the 30 control items will tell us mostly about the accuracy with which the participants' were able to recognize unknown targeted words as unknown, while the participants' Final Meaning Test scores will give us a firm indication of the accuracy with which the participants were able to rate words which they felt they knew.

10.3.1 Control items

As we observed with the study reported in Chapter Nine, a participant's rating of the control items in the V_States sessions provides a good indicator of the reliability of his or her rating, under the same conditions, of the targeted items from the text. While we might expect some incremental increase in familiarity with the word forms of the control items through repeated exposure to them over the eight V_States sessions, we would not usually expect any increase in knowledge of the word meanings of these items. If we do observe such an increase in their V_States rating of the control items, it may suggest that the participant is not only rating her estimated knowledge of the meaning of the items but also rating her increasing familiarity with their word forms. The two participants' rating of the control items in the eight V_States session is shown in Appendix 10.2.

Control items for N

N rated 19 of the 30 control items as 0 (definitely unknown) throughout the eight V_States rating sessions. A further four control items were rated as at least partly known for five or more of the eight rating sessions; for these items there was considerable variation from session to session. The remaining seven items were 0-rated for all but one or two of the sessions, with four of these items rated 1 or 2 in the final session. These generally steady figures for the control items suggest that the participant will have provided generally reliable rating of the 301 targeted items through the eight V_States sessions. In addition, N's rating of these control items in the final V_States session suggests that she would not have been over-conservative in her rating either of the control items or of 301 targeted items.

Control items for E

For E, 24 of the 30 control items remained 0-rated for all eight V_States rating sessions. The data for these items suggest that E was usually able to identify words that she did not know, over three quarters of the total, and that this ability remained stable for the course of the study. Five of the remaining control items showed considerable variation;

they started as 0- or 1-rated, changed states three or more times over the eight sessions, with all reaching state 3 for at least one session, and in the final session were rated between 0 and 3. The volatility in the rating of these items may be accounted for by the participant's wavering confidence, unaided by context or dictionary consultation, about whether she has correctly recognised the item in question. The results overall also suggest that E did not confuse unchanging word meaning knowledge for the control items with the increasing familiarity with word forms that may result from a total of eight encounters with the items during the V_States rating sessions.

10.3.2 Final Meaning Test

A further measure taken to ensure the reliability of participants' responses was the Final Meaning Test. In this test, participants were requested to give the meaning of all the targeted items which they had rated as state 3 (definitely known) in either of the last two V_States rating sessions. At the end of the study both participants were asked to give the meaning of the words they had rated as definitely known in the final V_States test 30 minutes earlier (and, incidentally, in the penultimate test six days earlier). They were able to demonstrate their knowledge of the meaning of the words by giving the meaning in English or Japanese; in fact, the answers were overwhelmingly in Japanese, with only 10 answers out of 177 given partly or totally in English by E, and only 5 out 117 by N.

In the Final Meaning Test, both participants were able to give a correct or near-correct meaning for a high proportion of the words they had rated as definitely known in the final V_States test: just under 90% for E and almost 80% for N. This high score provides confirmation that, at least for items rated state 3, the participants' rating of their confidence about knowing the test items was largely accurate. The relatively few errors tended to be either due to the misreading of words such as *racking* for *lacking* or *altar* for *alter*, or demonstrated a lack of clarity about the meaning of the word; *ivy*, for example, was described as a kind of flower and *grant* as a lot of money.

The results of the Final Meaning Test together with those for the control items provide confirmation that we should expect these participants' assessment of the 301 targeted items in the eight V_States sessions to be generally accurate. This is true specifically for 0-rated items, as confirmed by rating of control items, and for 3-rated items, tested in the Final Meaning Test. There was no test of accuracy for the two central states, both representing partial knowledge of targeted item word meanings, but we may expect similar variation between the two participants in the studies reported in this chapter to that found for the participants in Chapters Eight and Nine.

10.3.3 V_States results

Before we look at the data for actual V_States scores, which we will do separately for the two participants, it is worth pointing out two possible pitfalls in interpreting the data obtained. One is where, for example, there is, from one session to the next, a fall of 40 items in state 0, no change in states 1 and 2, and a rise of 40 items in state 3. We should not simply interpret this as meaning that 40 items have moved from state 0 to state 3; many items may have moved from state 0 to states 1 or 2, with others moving from states 1 or 2 to state 3. A further danger is assuming that no change in numbers of items in states 0 – 3 from one session to the next means that there has been no movement of items between states. While this may be the case, it may also be that movement between states has occurred but that it is balanced. For example, if 10 items move from state 0 to state 1, 10 from state 1 to state 2, and 10 from state 2 to state 0, then despite this movement of items between states, numbers of items in each state will remain static.

V_States results for N

We will begin by looking at the scores for N over the eight V_States sessions. N's results from the V_States sessions are shown in Table 10.3 and Figure 10.4.

Table 10.3
V_States scores for N (Total = 301 items)

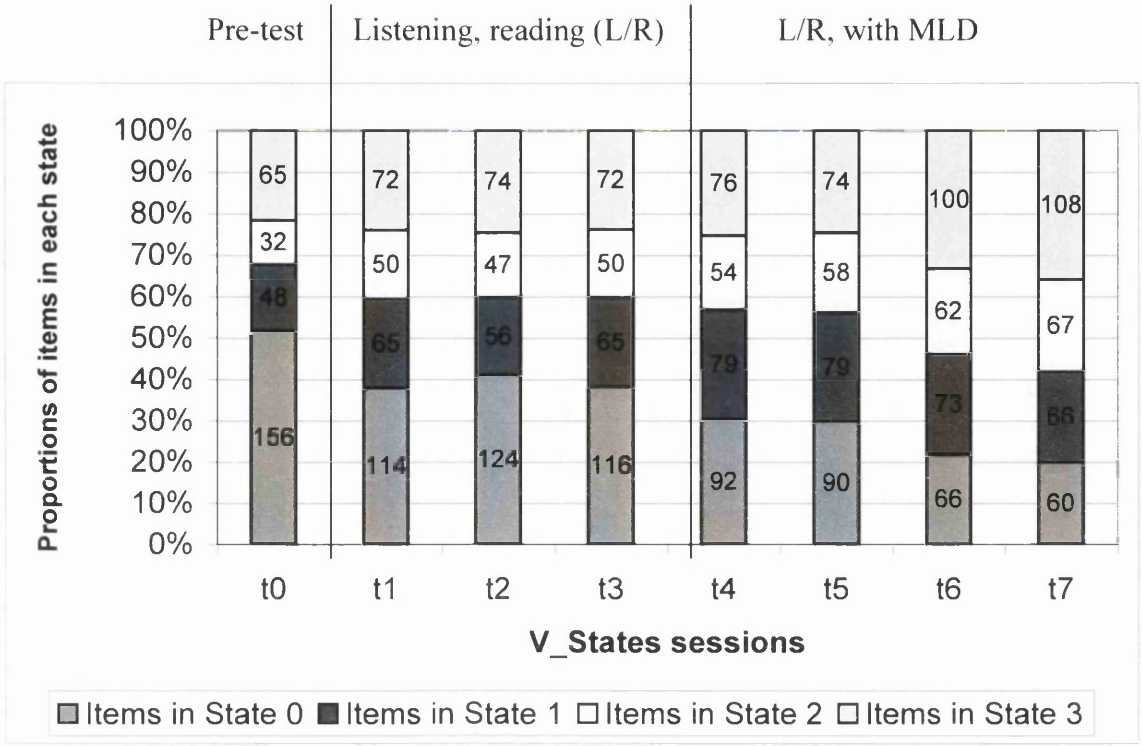
		state			
		0	1	2	3
pre-listening/reading	t0	156	48	32	65
	t1	114	65	50	72
listening/reading, no MLD	t2	124	56	47	74
	t3	116	65	52	68
	t4	92	79	54	76
listening/reading with MLD	t5	90	79	58	74
	t6	66	73	62	100
	t7	60	66	67	108

For N, in the first V_States session, t0, the meanings of over half the items are rated as unknown (state 0), 16% as not sure (state 1), just over 10% as probably known (state 2) and a little over 20% as definitely known (state 3). There is considerable variation over the next seven sessions, but by t7 states 0, 1 and 2 each account for about 20% of items, with items rated as state 3 rising to almost 40% of the total. It is worth noting, however, that even after eight V_States sessions, four of which took place within the with MLD condition, equilibrium in numbers of items in given states has yet to be achieved. The relatively small difference in numbers of states 0 and 3 items in the final two sessions does, however, suggest that this participant is approaching a state of equilibrium. The number of items in state 1 varies between 16% and over 26%, with the two greatest rises observed in the sessions following the introduction of the two new learning conditions. State 2 items vary between the t0 score of under 11% and the t7 score of just over 22%, with the only major rise following the first listening and reading without a dictionary.

The number of items in state 3, averaging 24% of the total, varies from session to session but does not rise or fall overall for the first six sessions. Although this may give the impression that there is no movement in and out of state 3 during this period, this is not true. While just over a half of the items (39) do remain stable state 3 items throughout, the remaining portion is made up by a total of 78 items which move in or out of state 3 during these five sessions.

Figure 10.4 shows us clearly what outward effect the introduction of the new learning conditions has on N's vocabulary knowledge for the targeted items. We can see a drop, for example, of over 40 items rated as not known in the change from the pre-test to listening and reading without a dictionary (t0 – t1), with rises in the number of items rated as not sure and probably known of 17 and 18 items respectively.

Figure 10.4.
Changes in proportions of states for N



Where MLD use is introduced (t4), there is again a drop in the not known (state 0) category, of over 20 items, with a rise of 14 items in the not sure category, although the number of state 3 items remains almost static. Also noteworthy about N's results is that not all substantial falls and rises are observed directly after changes in learning conditions. For example, one of the largest changes is between t5 and t6, both V_States sessions reflecting listening and reading with MLD use, in which the number of items in the not known category falls by 24 while the number of items rated as definitely known rises by 26 items.

It is also worth looking at combined numbers of items in adjacent states, since despite being from similar linguistic and educational backgrounds, and receiving identical rating instructions, any two participants may still differ markedly in their use of the four states in rating the extent of their knowledge of word meaning. In addition, as noted in Chapter Nine, combined states scores may provide a less volatile and, arguably, more reliable indication of vocabulary development than scores given for single states. Combined states numbers for N are shown in Table 10.5.

Table 10.5
Combined scores for adjacent states for N (Total = 301 items)

		states		
		0+1	1+2	2+3
pre- listening/reading	t0	204	80	97
	t1	179	115	122
listening/reading, no MLD	t2	180	103	121
	t3	181	117	120
	t4	171	133	130
listening/reading with MLD	t5	169	137	132
	t6	139	135	162
	t7	126	133	175

At t0, prior to any exposure to the targeted words from the text, N records about two thirds of the targeted words as 0 or 1 and approaching one third as 2 or 3. For the three sessions t1 to t3, around 60% of words are 0- or 1-rated and 40% 2- or 3-rated.There is considerable change over the next four sessions, and by t7 proportions of items in combined states are reversed, with 0 and 1-rated items accounting for just over 40% of the total and 2- and 3-rated items accounting for almost 60%. The two central states 1 and 2 account for a little over 25% of items at t0, rising to over 30% in the three listening and reading sessions t1 to t3, and to around 45% for t4 to t7.

V_States results for E

We will begin by looking at E's results over the eight V_States sessions as shown in Table 10.6 and Figure 10.7.

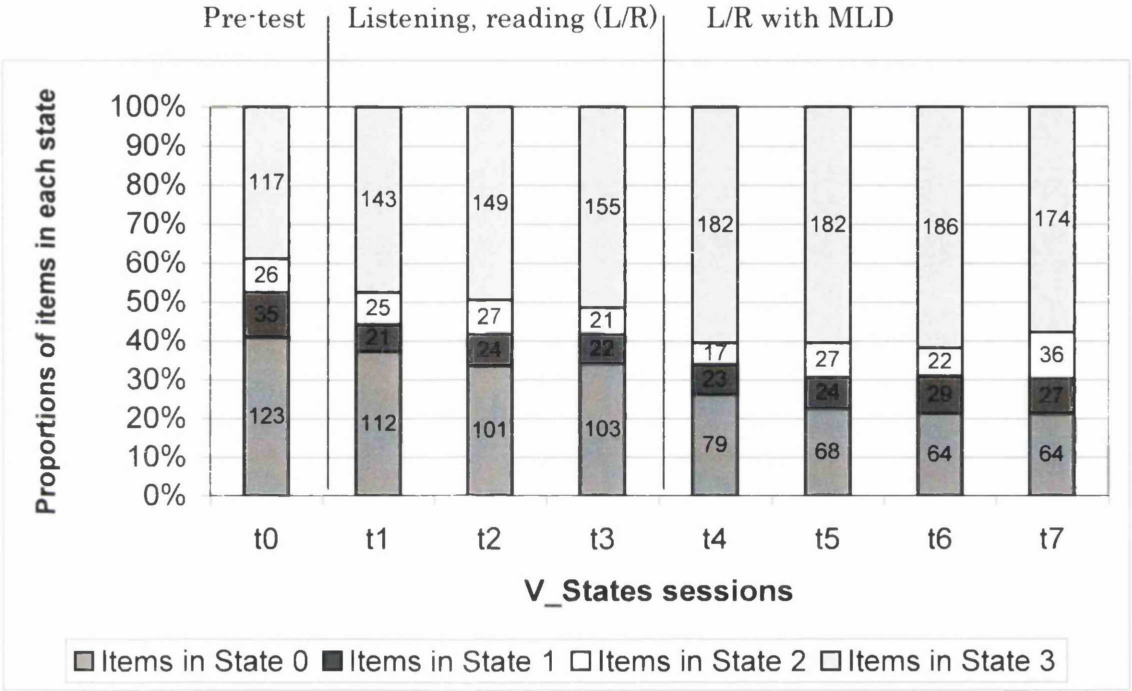
Table 10.6
V_States scores for E (Total = 301 items)

		state			
		0	1	2	3
pre- listening/reading	t0	123	35	26	117
	t1	112	21	25	143
listening/reading, no MLD	t2	101	24	27	149
	t3	103	22	21	155
	t4	79	23	17	182
listening/reading with MLD	t5	68	24	27	182
	t6	64	29	22	186
	t7	64	27	36	174

As we can see, in the first test around 40% of items were rated as not known and 40% as definitely known, with about 10% for each of the two middle states. By the eighth session, the proportion of items rated as unknown has fallen to just over 20%, while that for items rated as definitely known has risen by a similar amount to almost 60%, with the two middle states remaining fairly stable throughout, each ranging between around 5% and 12% over the eight V_States sessions.

Figure 10.7

Changes in proportions of states for E



There are substantial changes in at least some of the states at both of the stages at which a new learning condition for E is introduced. In the change from the pre-test to listening and reading without a dictionary (t0 – t1), the not known and not sure states both fall by over 10 items each, while the number of words rated as definitely known increases by 26 items. With the introduction of MLD use while listening and reading (t4), there is again a drop in the not known state, of over 20 items, and a further rise of 27 items in the definitely known state. One further change of note is the fairly substantial fall in numbers of items rated as definitely known items in session 7.

We will now go on to look at numbers of targeted words in combined adjacent states for E. These are shown in Table 10.8.

Table 10.8
Combined scores for adjacent states for E (Total items = 301)

		states		
		0+1	1+2	2+3
pre- listening/reading	t0	158	61	143
	t1	133	46	168
listening/reading, no MLD	t2	125	51	176
	t3	125	43	176
	t4	102	40	199
listening/reading with MLD	t5	92	51	209
	t6	93	51	208
	t7	91	63	210

At t0, the number of items E rated 0 or 1 represents just over 50% of the total number of targeted items. There is a steady movement of numbers of items from states 0-1 to 2-3 from t0 to t5, by which stage 0-1 rated items account for about 30% of the total and 2-3 rated items for around 70%. Between t5 and t7 there is very little further change in these proportions of numbers in the combined states. As for numbers of items in the two central states, E rates between 13% and 20% of the targeted words as state 1 or 2 over the eight V_States sessions. There is some variation in the number of items in these states from one V_States session to the next, but no overall rise or fall in numbers over the eight sessions.

10.3.4 V_States results projections and comparisons

As we have seen, V_States provides very rich data concerning vocabulary knowledge for large numbers of items both synchronically and diachronically. However, perhaps the most valuable property of the V_States programme is that matrices of scores from adjacent V_States sessions can produce projections of vocabulary development in a given learning condition. These projected results from one learning condition can then be compared with the actual results obtained in a subsequent learning condition for the same participant, allowing the comparison of the effect of two learning conditions on a single participant and for a single set of targeted items.

In both studies reported in this chapter, V_States sessions t1 to t3 reflect the effect on the words selected as targeted items under the condition of reading and listening to the songs without the use of a dictionary while sessions t4 to t7 reflect the condition of reading and listening while using an MLD. For this reason, we will look at projections obtained through the t2:t3 matrix, representing the reading and listening with no dictionary use condition. The t2:t3 matrix will be used rather than the t1:t2 matrix because, using two mid-condition scores, it may give a more representative and settled picture of the continuing effect of the listening and reading with no dictionary condition. Data for the two participants derived from this matrix are provided in Appendix 10.3.

We will begin by focusing on the results for state 3, items rated as definitely known. We will then go on to look at figures for states 2 and 3 combined, as the results from the study in the previous chapter suggest that combined states data may be less volatile than data based on changes of numbers of items in a single state. As before, we will look at the data for the two participants separately, only comparing results, where relevant, in the Discussion section.

Matrix data for N

We will begin with actual and projected state 3 scores for N: for numbers of words rated as definitely known. These are shown in Table 10.9 and Figure 10.10 below.

Table 10.9

Actual and projected state 3 items for N (Total = 301 items)

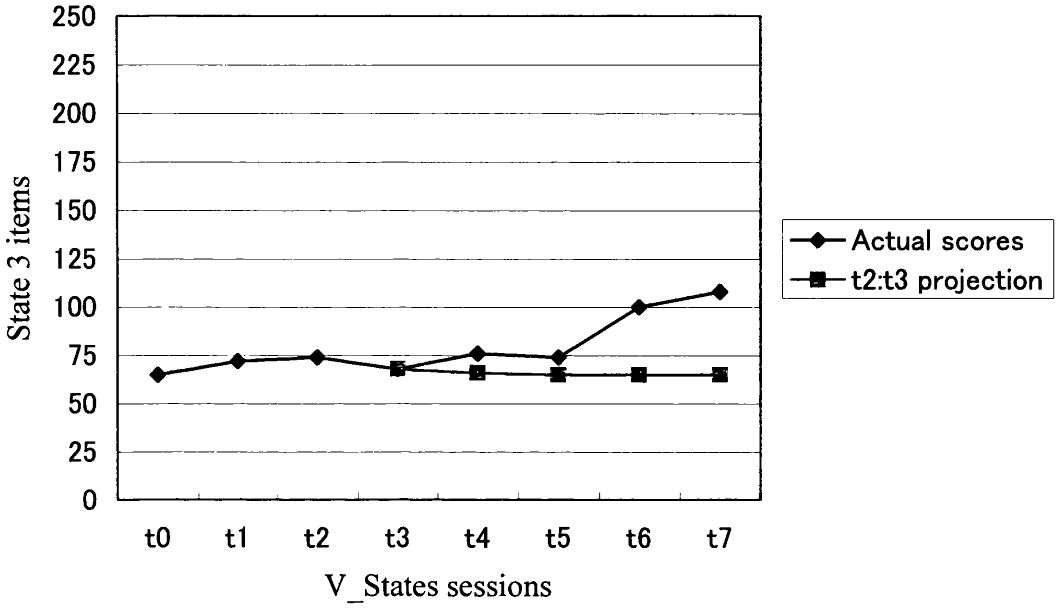
Figures shown in bold are projections based on the t2:t3 matrix.

	Pre-test							
Session	t0	t1	t2	t3	t4	t5	t6	t7
No MLD	65	72	74	68	***66***	***65***	***65***	***65***
With MLD					76	74	100	108

The t2:t3 matrix of scores for state 3 items in the listening and reading without a dictionary condition provides us with projections of numbers in this state for subsequent V_States sessions with this learning condition. We can then compare these with the actual state 3-rated items obtained for the sessions reflecting listening and reading with MLD use.

Figure 10.10

Actual and projected state 3 items for N (Total = 301 items)



Apart from small rises and falls, the number of state 3 items recorded by N over the first six sessions remained fairly static, only rising from the previous session by a sizeable number at t6 and t7. For the first two sessions reflecting MLD use, then, dictionary use has little effect on numbers of state 3 items other than to reverse a fall in numbers at t3. This is shown by the closeness of the actual numbers of state 3 items at t4 and t5 to projections for the listening and reading sessions without dictionary use, and by the almost parallel lines at this stage on the graph of Figure 10.10. From t6, however, numbers for these two conditions diverge as the numbers of state 3-rated items increase. This is what we might expect, with the effect of MLD use becoming apparent, although the delayed effect is unexpected and requires further investigation. We may gain a better understanding of what is behind these figures as we look at numbers of items, and projections of these numbers, for states 2 and 3 combined.

Table 10.11 and Figure 10.12 show combined numbers of items in states 2 and 3 for N for the eight V_States sessions, together with figures for these two states produced by the matrix of numbers for t2 and t3.

Table 10.11

Actual and projected combined states 2 and 3 items for N (Total = 301 items)

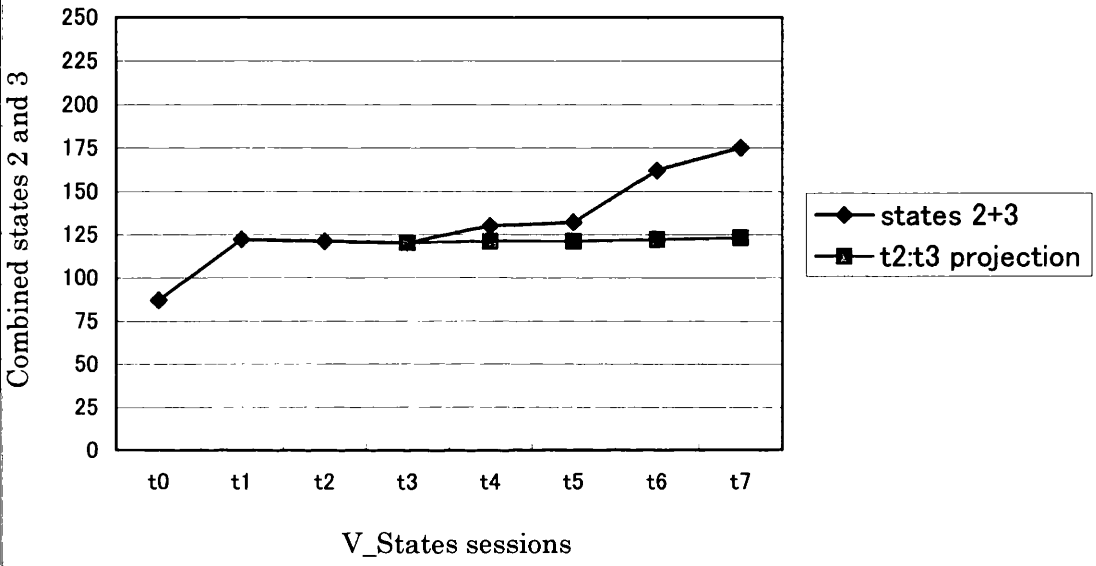
Figures shown in bold are projections based on the t2:t3 matrix.

	Pre-test							
Session	t0	t1	t2	t3	t4	t5	t6	t7
No MLD	87	122	121	120	***121***	***121***	***122***	***123***
With MLD					130	132	162	175

The data for the two states combined, compared with those for state 3 alone, provide two important insights regarding N's vocabulary development for the targeted items. The first, as predicted, is that the combined figures show considerably less volatility than those for state 3 alone; numbers of items rise at certain points and remain stable at others but do not rise and fall from session to session. Within the listening and reading condition, for example, there is almost no change in combined numbers of state 2 and 3 items. This is also true for the first two sessions in which dictionary use was permitted. The second insight, for N, is into the immediate identifiable effect of dictionary use on knowledge of the targeted items. While at t4 and t5 the figures for state 3 alone show no immediate effect of dictionary use, the combined state 2 and 3 data do show an immediate and sustained rise in numbers. This suggests that, contrary to expectations, the combining of data for adjacent states may not only reduce volatility in the numbers of items from session to session but may also increase the capability of the instrument to reveal small changes in vocabulary development.

Figure 10.12

Actual and projected combined states 2 and 3 items for N (Total = 301 items)



Matrix data for E

We will begin by looking at actual and projected numbers of items for E for state 3 items. These are shown in Table 10.13 and Figure 10.14 below. The data derived from the t2:t3 matrix provide projections of numbers of state 3 items in subsequent V_States sessions for the learning condition of listening and reading without dictionary use. These projections for this first condition can then be compared with actual figures for the second condition of listening and reading aided by MLD use.

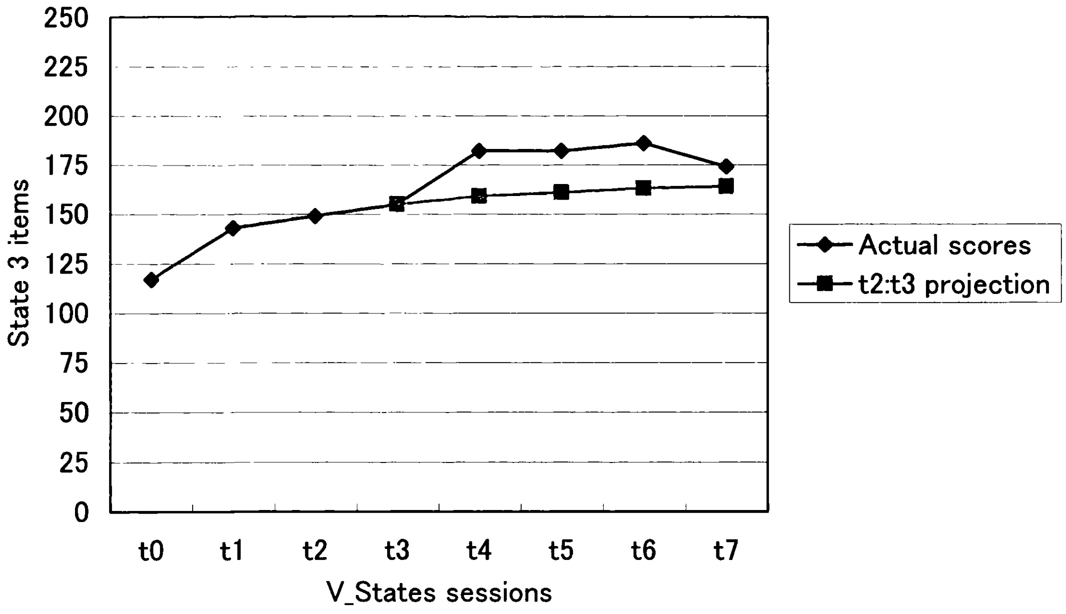
Table 10.13

Actual and projected state 3 items for E (Total = 301 items)

Figures shown in bold are projections based on the t2:t3 matrix.

	pre-test							
Session	t0	t1	t2	t3	t4	t5	t6	t7
No MLD	117	143	149	155	***159***	***161***	***163***	***164***
With MLD					182	182	186	174

Figure 10.14
Actual and projected state 3 scores for E (Total = 301 items)



As is clear from both the table and the graph, the effect of dictionary use on knowledge of the targeted items is for E to produce an immediate and sustained rise in numbers of state 3-rated items, markedly higher than the steady but very slow rise predicted for the first learning condition. Although the actual number of state 3 items remains almost constant between t4 and t6, and actually falls at t7, it still remains strikingly higher throughout than projections for the condition of listening and reading without MLD use.

We will now go on to look at actual and projected numbers of items for E for states 2 and 3 combined. These are shown in Table 10.15 and Figure 10.16.

Table 10.15

Actual and projected combined states 2 and 3 items for E (Total = 301 items)

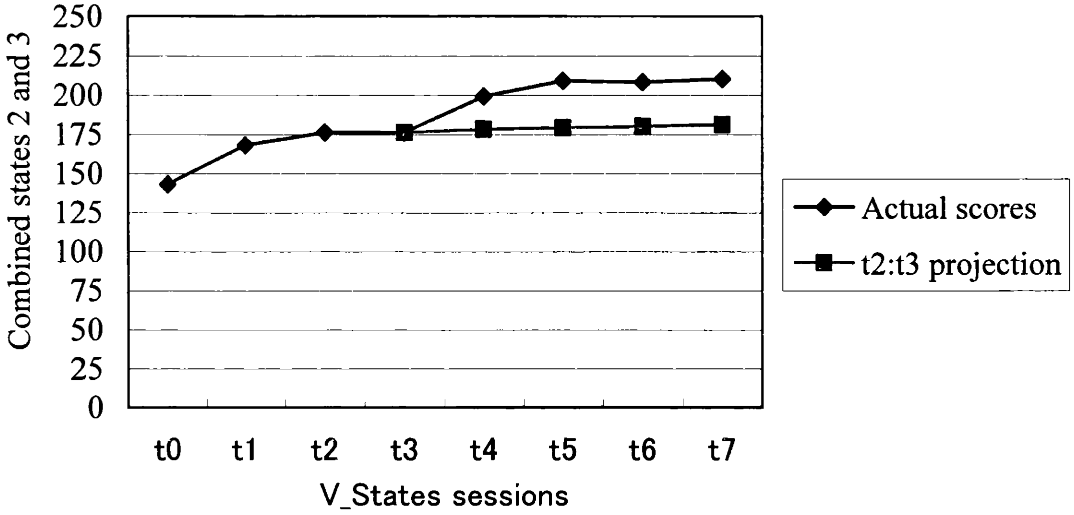
Figures shown in bold are projections based on the t2:t3 matrix.

	Pre-test							
Session	t0	t1	t2	t3	t4	t5	t6	t7
No MLD	143	168	176	176	***178***	***179***	***180***	***181***
With MLD					199	209	208	210

Two aspects of the data become apparent as we look at the figures for the combined states: the points at which there are rises in numbers of items in the two states and the points at which there are plateaux where there is little or no change in numbers from one V_States session to the next. The two largest rises follow the introduction of the two learning conditions: in the first case a rise of 25 items and in the second a rise of 23 items. In the session following each of these rises there are further smaller rises, each of about 10 items. Apart from this, in the remaining listening and reading session and in the two remaining sessions with dictionary use, figures remain almost static.

Figure 10.16

Actual and projected combined states 2 and 3 items for E (Total = 301 items)



As the table and graph show, projected figures for the continuation of the first learning condition for the four sessions of t4 - t7 are 20 items short of the actual figure at t4 and close to 30 items short for the remaining three sessions. With the combined state 2 and 3 figures, as there is no drop in numbers in the final V_States session, we do not see any closing of the gap between the two conditions. Rather, we see that the figures for the two conditions are almost parallel for the last three sessions.

As we have seen, V_States can provide a large amount of valuable data, both for actual numbers of targeted items in one or more states and for predictions and comparisons of these numbers with data for a second learning condition. It provides a means by which this data may be obtained automatically.. Our focus here has largely been on projections of state 3 data but, as the data for combined states confirm, V_States may also be used to provide projections for items in other states, in both single and combined states.

10.3.5 The effect of MLD use on targeted words that are or are not looked up

V_States has provided an extensive longitudinal overview of the effect of MLD use while reading on the participants' knowledge of the targeted items. This overview does, however, obscure the fact that within this one learning condition the targeted words receive one of two very different treatments; some targeted words are looked up and some are not. When we want to compare data for subsets of words within the total subset of targeted words, manual intervention and word by word sorting of responses does become necessary. We will now look at the manually extracted data for one set of these targeted words: words which are 0-rated at the V_States session immediately prior to the participants' beginning to use an MLD while listening and reading. Tables 10.17 and 10.18 below show the final rating for these targeted words by the two participants.

Table 10.17

Final scores for N for 0-rated items at V_States session 3 (Total = 116 words)

	Total	Final state	Number (%)	
Looked-up state 0 words	61	0	31	(50.8)
		1	18	(29.5)
		2	5	(8.2)
		3	7	(11.5)
Not looked-up state 0 words	55	0	27	(49.1)
		1	17	(30.9)
		2	7	(12.7)
		3	4	(7.3)

For N, figures for the final states of items 0-rated at V_States session 3 are almost identical for items that were looked up and items that were not. In both cases, about 50% of items remained 0-rated, about 30% rose one state to state 1, and the remaining 20% rose to state 2 or 3.

The picture for E is very different. While almost 40% of words E looked up remained 0-rated, well over 90% of words which were not looked up remained 0-rated. Approaching 40% of the looked-up items rose to states 2 or 3 by the final V_States session, as compared to under 6% for targeted words which were not looked up.

Table 10.18

Final states for E of words rated as state 0 at V_States session 3

	Total	Final state	Number (%)	
Looked-up state 0 words	68	0	27	(39.7)
		1	15	(22.0)
		2	8	(11.8)
		3	18	(26.5)
Not looked-up state 0 words	55	0	33	(94.3)
		1	0	(0.0)
		2	1	(2.9)
		3	1	(2.9)

10.4 Discussion

In our discussion, we will focus on three main aspects of the two studies reported in this chapter. We will:

1) Address the hypotheses proposed regarding vocabulary acquisition in the two conditions under investigation.
2) Evaluate the effectiveness of the materials used in these studies.
3) Consider the instruments employed in these studies to investigate the issues under investigation.

10.4.1 Addressing the hypotheses

There were three main hypotheses proposed regarding the data from the studies described in this chapter. They are:

i) That L2 vocabulary growth would be evident in both learning conditions;

ii) That there would be evidence of vocabulary growth from the second condition exceeding that expected from the first condition;

iii) That during the With_MLD sessions there would be gains both for targeted words that were looked up and for words that were not looked up.

10.4.1.1 Evidence of L2 vocabulary growth in both conditions

The first hypothesis was that L2 vocabulary growth would be evident in both learning conditions. The data from both studies confirmed that there was clear evidence of vocabulary growth both from listening and reading without a dictionary and from listening and reading with the use of an MLD.

For N, the benefit of reading and listening alone was most clearly evident in the first of the V_States sessions for this learning condition, in which there was a rise of 35 items for combined states 2 and 3 items. In the two subsequent sessions this gain was sustained but there were no further rises either for state 3 items alone or for combined state 2 and 3 items.

E produces an initial rise of 26 state 3 items in the first V_States session for the listening and reading without MLD condition. There are small further rises in the two subsequent sessions for state 3 items, although this continuing rise is not evident for combined numbers of state 2 and 3 items. The admittedly limited data from these studies suggest that the greatest benefit to vocabulary development of reading an L2 text may be in the first reading of the text. Subsequent readings may serve to consolidate previously acquired vocabulary knowledge but does not appear to add substantially to numbers of items known.

The two participants both showed evidence of vocabulary growth in the With_MLD condition, although they differed in when and how this vocabulary growth was recorded. For N, there was no sustained rise in numbers of state 3 items for the first two V_States sessions for this condition. Only from the third of these sessions did we see a large and sustained rise in state 3 items. The picture regarding combined state 2 and 3 items is rather different; we do see a small sustained rise in combined numbers of items for the first two sessions, followed by a larger rise in the third, and a further rise in the fourth of these sessions. E showed a large rise in numbers of state 3 items from the first of the V_States sessions for the With_MLD condition. This rise was, largely, sustained but there were no further rises from this figure in the three remaining sessions, with numbers in fact falling slightly in the final session.

The impression we gain from numbers of combined state 2 and 3 items is, again, different from that obtained from that for state 3 items alone. For the combined states, the initial rise in numbers at the first session for this condition is followed by a smaller rise in the second session, following which figures remain constant for the final two sessions. For E, then, MLD use had a pronounced immediate effect, as compared to a delayed effect for N. This difference may be attributed both to the differing English proficiency levels of the two participants and to the related issue of the differing extent and type of dictionary use by the participants over the four sessions.

10.4.1.2 Evidence of greater L2 vocabulary growth through MLD use

The second hypothesis was that actual gains in word knowledge during the With_MLD sessions would exceed projected gains for further sessions of listening and reading without dictionary use. For participants in both studies, increases in numbers of state 3 items in V_States sessions for the With_MLD condition were considerably higher overall than projections for the condition without dictionary use. For N, however, there was initially little perceptible benefit of MLD use; gains in state 3 of only 10 items in the first two V_States sessions for this condition above those projected for the without dictionary condition. Only in the last two V_States sessions does a more marked benefit become evident; in both of these sessions there is an advantage of at least 40 items for the With_MLD condition over projections for these sessions with the previous condition. This is true both for state 3 items alone and for state 2 and 3 items combined.

For E there is an immediate benefit attributable to MLD use while reading and listening over that projected for the condition without dictionary use. This advantage over projections of more than 20 state 3 items is maintained for the first three sessions in this condition, but drops to a difference of only 10 state 3 items in the final V_States session. When we look at actual and projected numbers of combined state 2 and 3 items, the picture is more complex. At the first V_States session for the with dictionary condition, actual combined numbers of state 2 and 3 items are just over 20 items higher for the with dictionary condition than numbers projected for the condition without dictionary use. At the next session this rises to a difference of 30 items, which is largely sustained for the two remaining V_States sessions, with no fall in the final V_States session.

We can see that the hypothesis is confirmed by data from both participants but that the timing and extent of the benefit of MLD use differs for the two participants, again suggesting that the extent of this benefit will be affected by the dictionary user's L2 proficiency level and by how he or she uses the MLD while reading. Specifically, E's greater English proficiency may account both for her more extensive dictionary use in the first With_MLD session and for the much reduced number of words looked up for the first time in subsequent sessions. Her greater initial use of the MLD may be due to two factors: the greater ease and efficiency with which an experienced L2 reader with a large vocabulary can use an MLD, coupled with a desire to look up all unknown words in the text. The smaller number of words subsequently looked up for the first time is a consequence of the large number looked up before; she had reached a kind of ceiling, with few unknown, and unlooked up, words outstanding from the text.

10.4.1.3 L2 vocabulary growth of words that were and were not looked up

The third hypothesis was that during the with MLD sessions there would be gains both for targeted words that were looked up and for words that were not looked up. Here, we focused on the final V_States rating for targeted words that were 0-rated in the session prior to the commencement of MLD use while listening and reading. The data for both participants showed a clear effect for looked-up words that were previously rated as unknown. However, this effect is far from universal for these looked-up words; for both participants, between around 40% and 50% of words remained rated as unknown after having been looked up. This may mean that looked-up words that were 0-rated in subsequent *V_States* sessions were not recognized on the computer screen, it may mean that the participant forgot the meaning of the word after looking it up, it may mean that the participant was unable to grasp the meaning of the word from the dictionary entry, or it may mean that the participant was unable to reconcile the meaning in the dictionary with the apparent meaning in the text and so was left without a clear understanding of the word's meaning.

Regarding targeted words that were not looked up, data for the two participants differed widely. For N, about 50% of items remained 0-rated, about 30% rose one state to state 1, and the remaining 20% rose to state 2 or 3. This is almost identical to the figures for words that were looked up and suggests that N's looking up of previously 0-rated targeted words affected word knowledge not only of the looked-up words but also of other previously unknown words in the text. The picture is very different for E regarding words which were not looked up; the proportion of words which were not looked up and remained 0-rated was 94.3%. For E, then, the effect on vocabulary acquisition of looking up unknown words appears to be largely limited to those words that are looked up. There is little spill-over effect of MLD use to unknown words which were not looked up, and these typically remain rated as unknown.

As we consider why there was little spill-over effect for unknown words that were not looked up, we need to consider how these learners' mental lexicons and the L2 environments may differ from those described by Meara (2005). Meara's conditions for the reactivating of a dormant L2 vocabulary are, naturally enough, the existence of a dormant L2 vocabulary together with an L2 environment that would serve to reactivate it. It may be that neither of these conditions was met to any real extent in the studies under investigation. For these learners of English, the study of English was one part of their everyday life; there was no "before" and "after" of no contact with the L2 to contrast with immersion in the L2. This would mean that they have relatively little dormant L2 vocabulary. Also, for Japanese learners, there is much less L2 vocabulary that can be retrieved or activated through L1 or L3 cognates than for speakers of many European languages. In addition, although some L2 reading may be equated to an L2 immersion experience, the experience of reading opaque L2 song lyrics, as in these studies, may not be a particularly favourable environment for the reactivating of knowledge of forgotten L2 words.

10.4.2 Evaluating the materials

We will now consider whether the reading and listening materials used in these studies fulfilled the various purposes for which they were selected.

From an administrative perspective, the listening and reading materials used in these studies fulfilled the functions expected of them. The use of song lyrics reduced by around one third the time required for reading as compared to the book used in the studies reported in Chapters Eight and Nine. Thanks largely to the text type and length, the study was, overall, an activity that was both familiar and enjoyable for the participants in these studies, and one that did not demand too much time of them.

As a source for suitable targeted words for the studies, the song lyrics cannot be described as totally successful in all respects. First, we need to consider whether the texts provided a sufficient number of suitable items for the studies. In other words, we need to ask whether there were sufficient numbers of words that were at least largely unknown to the participants at each of the seven reading and listening sessions. For N, the answer appears to have been yes throughout the study. At the first V_States session, prior to contact with the listening and reading materials, over 200 items were rated as unknown or probably unknown. After the last session of reading and listening without dictionary use, there were still 181 items that were 0- or 1-rated. This left a large number of items that could be looked up in the four With_MLD sessions. A large number were in fact looked up: a total of 134 words, 100 of which were targeted items, with 90% of these previously 0-rated.

For E, the materials used in this study may not have provided sufficient numbers of targeted words. At the V_States session prior to the first encounter with the materials, E rated 158 of the targeted items as unknown or probably unknown. This number fell to 125 items by the final V_States session without MLD use. Although this may appear to have been a sufficiently large pool of items from which the effects of dictionary look-ups could be recorded, E's look-up behaviour suggests that this may not have been the case. She looked up words a total of 188 times, with over 150 of these for targeted items, more than the totals for 0- and 1-rated items at this stage. This helps explain why so many of her look-ups were for words previously looked up, and why she tended to look up fewer words in later sessions. This is clearly apparent if we contrast E's look-up data for the first two With_MLD sessions with that for the last two sessions. In the first two With_MLD sessions with MLD use, E had a total of 104 look-ups, with 80 of these being for words that had not been previously looked up in the study. In the last two sessions, E used the dictionary a total of 84 times, with only 25 of these being for words not looked up in previous sessions. This fall in number of words looked up for the first time is reflected in the much reduced rate of increase of 2- and 3-rated items in later sessions. E appears to have reached a ceiling as far as the effect of MLD use of vocabulary development is concerned, and this may well be a result of the text containing an insufficient number of unknown words that she would be likely to look up.

A second question is whether the targeted words were selected with sufficient accuracy in terms of containing the words in the text that the participants were most likely to look up. The banding of the targeted words according to the JACET 8000 word list, with around one third of targeted words in the first three 1,000 word bands, suggested that the set of targeted words may have contained too many high frequency words. However, the high proportions of looked up words that were targeted words (almost 75% for N and near to 85% for E) suggests that this itself was not a major problem. Rather, as noted above, the problem was that the whole text contained too few items likely to be looked up by an advanced level learner of English such as E.

A third issue regarding the words from the text that were selected as targeted items concerns their comprehensibility in the context in the song lyrics. We should not always expect single contexts of unknown words to provide us with sufficient information for us to guess the words' meaning. If that were the case, language learners would have no need of dictionaries. On the other hand, information from the context may be confirmed through information in dictionary entries, and the correct sense within a dictionary entry identified by reference to the context of the word being looked up. However, the more opaque and less clear the language use of the text, the less the context of an unknown word will confirm a learner's comprehension of the MLD entry for the word. A few examples from the song lyrics used in this study, with targeted words shown in italics, may illustrate this problem:

She could see the valley barbecues from her window *sill…*

Maybe you will see my face reflected there on the *pane*
In the window of our poor *forlorn* and broken home.

Magnolias hopeful in her *auburn* hair.

Now wear your *scars* like medals/Defender of the faith

The perils of *benefactors*/The blessings of parasites

Rather than confirming comprehension, atypical or otherwise misleading contexts may lead MLD users to doubt their correct comprehension of the definition of an unknown word. Examples are the use of the words *forlorn* or *hopeful* with inanimate objects, or the implication that *benefactors* are something undesirable. Encounters with words in these contexts would, in turn, result in dictionary use for such items being less effective not only with regard to comprehension but also as concerns retention of looked-up words. This does appear to be the case for a fair number of items in this study which are used in the song lyrics in ways that are far from typical for the word. With such items, even repeatedly looking up the items, as E did, may not help in the comprehension of the items as they are used in the text.

Our consideration of the above contexts does help us to understand something of the challenged faced by language learners when trying to understand words from their contexts. However, for a language learner the contextual richness of a lexical environment depends not only on the text itself but also on the amount of information in the context of the unknown word that is accessible to the learner (Mondria and Wit-de Boer, 1991). In terms of the negative impact of unhelpful contexts, we can also see that in some cases the effect of the contexts on comprehension, and learning, of a looked-up word will be inversely proportional to the L2 proficiency of the learner; the less a learner understands of the unhelpful context of a looked-up word, the less it will conflict with his or her comprehension of the MLD entry for the word.

10.4.3 Considering the instrument

Although only one main instrument, V_States, was used in these studies, it performed three different functions:

i) To give a cross-sectional picture of states of word knowledge of large numbers of targeted items at a particular time;

ii) To provide a longitudinal record of changing states of word knowledge in one or more learning condition;

iii) To create a means of comparing the effect on L2 word knowledge of two different learning conditions.

V_States performed as hoped in all three functions and its reliability, with reliable participants, was confirmed by the results for the 30 controls items and by the results of the Final Meaning Test, in which participants were able to give correct meanings for a high proportion of the targeted items which they had rated as definitely known. There were problems, or at least unexpected outcomes, observed in these studies such as the sudden slowing down for E of increases in numbers of items rated as definitely known, or the lack for N of any apparent immediate effect of dictionary use. These problems were identified through V_States, another indication of its reliability and usefulness. However, the problems cannot be attributed to the use of V_States. Rather, as suggested above, they may probably be attributed to the lack of sufficient numbers of unknown words in the text or to the participant's dictionary use.

The outcome observed in previous studies of unexpected and counter-intuitive falls in state 3 items also occurs once or twice in these studies. One response proposed regarding this question was to combine the data for the two lower states (0 and 1) and that for the two higher states (2 and 3). As expected, this measure did serve to provide more stable data, in which numbers of known items may rise or stabilise within a particular learning condition but do not usually fall. Another expected effect of combining data for adjacent states was that the sensitivity of the instrument would be blunted. Rather than showing movement of targeted items among four states we would be reduced to the less sophisticated, and arguably less true to life, model of two states of word knowledge: *know* and *do not know*. Pleasingly, and contrary to expectations, the combination of results from two states tended to produce data that was more sensitive than those for a single state. For N, for example, we are able to observe the effect of dictionary use on knowledge of the targeted items in data for combined states 2 and 3 prior to it becoming evident for state 3 alone. For E, in her final V_States session, the fall in numbers of state 3 items is rendered in terms of combined states 2 and 3 as a continued plateau.

We also used the data for combined states for the matrices to compare projected data for one condition with actual data for a second condition, The effect of doing this, shown most clearly in Figures 10.12 and 10.16, is to produce data that appears much better than that based on data for a single state. Especially for matrix data in which we are looking at trends of vocabulary development, volatility in numbers from session to session may be seen as the opposite of reliability. It appears, then, that rather than providing a blunter, less sophisticated instrument, the use of these combined data may serve to clarify and to sharpen our understanding of a participant's vocabulary development in one or more learning condition.

Having established that combined states data appear to work better than that for a single state, we need to consider what these states may mean to the participants and, perhaps, to challenge our conception of states of vocabulary knowledge. It is tempting to equate the four states of word knowledge recorded in V_States with four points on a single scale, with state 0 at one end meaning "no knowledge", and state 3 at the other end, meaning "full knowledge" of the word. Two aspects of the participants' V_States data appear to reflect an understanding of the meaning of the four states that is more complex than this. The differences among participants in the extent to which they made use of the central states is one indicator of this; for N, these states accounted for between 25% and 40% of items in the eight V_States session while for E only between 13% and 20% of items are rated as state 1 or 2. This difference in the use of states 1 and 2 could be said to indicate differences in the mental lexicon of the two learners: that for E there are fewer half-known L2 words than for N; for E, the vast majority of English words are either definitely known or not known at all. We should, however, also consider that these differences may not necessarily be of extent of word knowledge alone.

In a self-assessment instrument such as V_States, confidence about the accuracy of the rating is an intrinsic element of the instrument. This can be seen in the descriptions of the four states:

0 – I don't know what this word means

1 – I'm not sure I know what this word means

2 – I think I know what this word means

3 – I definitely know what this word means

The words *not sure*, *think* and *definitely* in the descriptions of the states offer confirmation, both for us and for the participants, that the instrument is not measuring word knowledge alone but confidence about that word knowledge as well. If we reflect on how these two measures might work together, we may reformulate our understanding of what these states may mean in terms of two questions posed for any targeted word:

i) Do you know the word?

ii) Are you sure of that evaluation?

The four states of V_States would then represent the following sets of answers for these two questions:

	state 0	state1	state 2	state 3
i) Do you know the word?	no	no	yes	yes
ii) Are you sure?	yes	no	no	yes

This way of interpreting the V_States data states helps us to understand how the combining of states may produce a more reliable and more sensitive measure of vocabulary change. It also helps explain the falls in state 3 items in the final V_States sessions, both for E and for the participant S reported in Chapter Nine. We would not expect a fall in word knowledge at this stage but we should not be surprised to see a drop in confidence regarding state 3 items when participants know that their knowledge of state 3 items will be tested immediately after the V_States session. In fact, rather than simply causing a drop in an abstract type of confidence, the impending test changes the second question to "Are you confident that you will be able to give the meaning of this word in a test which will take place in a few minutes?” It is not surprising that participants feel able to say yes to this question for a smaller number of items.

These final comments about what the four states may mean to the participants, and how the data may be interpreted, do not negate the value of V_States data. They do, though, remind us of the need to treat these data, and other experimental data for L2 vocabulary development, with care.

10. 5 Conclusion

In this chapter we investigated L2 vocabulary development through reading and dictionary use with a different text type from that used in Chapters Eight and Nine. The English proficiency of one of the participants was also of a different level from those in previous studies. Despite these differences, the studies were largely successful in measuring and comparing the learning conditions under investigation. The studies also confirmed the value of following the word knowledge state of hundreds of targeted words over a total of six learning sessions, as opposed to the couple of dozen targeted words for one learning session, as in many other studies.

The main instrument, V_States, was again invaluable as a tool which made possible the comparison of projections for one learning condition with actual results for a second condition. Specifically, the justified use of results from combined word knowledge states rendered the data from this instrument both more stable and more sensitive than that from a single state. By this means it was possible to overcome the volatility between sessions which may reduce the value of data derived through matrices. The resulting V_States data clearly demonstrated the extent of the benefit of encountering unknown or partially known words in the context of English songs, and the additional benefit of using an MLD while doing this.

Finally, the data from these studies helped point out potential problems relating to text type and level and learner proficiency levels. These were not problems with the studies in themselves but provided clear indications as to when and how often language learners might benefit from rereading a text with or without the use of a dictionary.

This chapter brings to a close our reports of experimental investigations into the effect of MLD use on L2 vocabulary development. It has not, and can not, answer all the questions raised through these investigations. We will address some of the outstanding issues central to this topic in Chapter Eleven.

Chapter Eleven: Review and Discussion

The purpose of this chapter is to highlight and discuss major issues relating to vocabulary learning and MLD use revealed through the studies reported in this thesis. We will begin by drawing together findings from the studies that raise questions regarding three areas relating to the use of MLDs: rates of successful dictionary use by language learners; individual differences in language learners' ability to use learner dictionaries effectively; and language learners' varying success in dealing with entries for individual L2 words in monolingual dictionaries. These will lead us to discussions of issues grouped around three concerns that are central to research into the use by language learners of learner dictionaries: successful MLD use and learner choices regarding dictionary use; the nature of L2 vocabulary acquisition that takes place in the context of MLD use; and, finally, methods for investigating L2 vocabulary acquisition through MLD use.

11.1 Review of findings

The primary focus of attention in the research reported in the previous chapters of this thesis has been the comparison of two learning conditions for unknown L2 words: use of monolingual learner dictionaries and learning through inference of meaning from written contexts of the words. We will now turn to three further aspects of the studies to which much less attention has been paid: rates of successful dictionary use in terms of production, comprehension, and retention; variation among individual participants in the studies; and variation among results for individual words.

11.1.1 Productive MLD use

In the first two studies, those reported in Chapters Three and Four, one task required of participants was the production of acceptable sentences using the target words. In both studies, participants with access to MLD dictionary entries for the target words were only able to produce sentences demonstrating acceptable use of the target words in about 60% of the cases. Production of language in the target language is one use for which monolingual dictionaries has been proposed (e.g., Hanks, 1987: 117) yet these intermediate level learners, admittedly largely unpractised in using an MLD, were unable to produce acceptable sentences for an average of two out of five target words. This raises questions both about the L2 proficiency level at which language learners should start using an MLD and about the extent to which language learners need training in the use of learner dictionaries.

The low success rates also raise questions about this method of requiring learners to produce sentences for target words. One issue is that the requirements of the task may not be clear either to researchers or participants. Should a sentence demonstrate knowledge of the meaning of the word? Should it be typical of sentences using that word? Or should it simply show a possible context for the target word? These questions are also reflected in differences in raters' evaluations of these sentences; inter-rater reliability levels remained low, at around .7 for both studies, despite raters' receiving detailed guidance about the rating of sentence acceptability. The persistence of this problem is, perhaps, understandable since it is hard to specify parameters for the two standards of possibility and typicality, and since the two standards of typicality and demonstration of meaning may often be in opposition to each other.

A further problem is that it is not clear what data produced by sentence production may be evidence of. Participants' ability to produce acceptable sentences for target words is not automatically evidence of their comprehension of the target words through the dictionary entries or written contexts. The figure for production of acceptable sentences by MLD-using participants in the study reported in Chapter Four was comparable with that achieved by participants with access only to example sentences in which the target words were used, yet figures reflecting comprehension of the target words differed markedly for the two groups. In most cases where MLD users produced acceptable sentences they also gave accurate translation equivalents, and for the target words overall they gave 25% more accurate translation equivalents than acceptable sentences. However, participants with access to example sentences very often produced acceptable sentences for target words for which they were unable to give the meaning. Overall, participants with access to these materials alone typically gave 70% more acceptable sentences than accurate translation equivalents for the target words.

11.1.2 Comprehension of MLD entries

This brings us to comprehension rates for target words. In the first four studies, reported in Chapters Three to Six, participants' comprehension of target words was demonstrated by production of accurate L1 translation equivalents for the words. We have already looked in detail at comparative rates of comprehension for target words from MLD entries and from various contexts in which the target words were used, and we have observed the much superior rates of comprehension for users of MLDs over those for participants attempting to infer meaning from context. However, there has been little consideration of levels of successful comprehension using MLD entries, despite comprehension being crucial to the use, or non-use, of MLDs by language learners outside experimental conditions. In all four studies comprehension rates were low for the groups of participants with access either to full MLD entries for the target words or to MLD definitions for the target words. In these studies, these participants were able to give accurate translation equivalents for an average of between 50% and 60% of target words, with partially correct translation equivalents, in some studies, bringing this figure up to almost 85%. If these figures are in any way representative of intermediate level language learners' comprehension rates of words consulted in an MLD outside experimental conditions, they suggest that such learners will often fail to gain an accurate understanding of looked-up words from MLDs and in some cases will fail to understand a word's meaning at all.

In the context of the studies reported in this thesis, the main conclusion regarding these data was that MLD use is much more likely to lead to accurate comprehension of unknown words than is guessing their meaning from written texts in which the words appear. Beyond experimental contexts, language learners' conclusions regarding successful comprehension rates may be rather different. They may question the value both of guessing from context and of using an MLD, preferring the more assured comprehensibility of L1 glosses, bilingual dictionary entries, or teachers' explanations.

11.1.3 Retention of word knowledge

Rates for language learners' retention of lexical information gained through MLD use vary according to the type of lexical information under investigation. In the studies reported in Chapters Three, Eight, Nine, and Ten, the participants were tested for, or asked to report, their retention of word meanings learned through MLD use, while in the studies reported in Chapters Four, Five, and Six, participants were tested for their ability to recognise the contexts, including MLD entries, in which they had encountered the targeted words.

In the study reported in Chapter Three, MLD-using participants were typically able to give meanings for only 20% of target words one week after a learning activity using the dictionary entries for the words. This low figure is partly a result of the low comprehension rates for the words; comprehension rates were low, and accurate accurate retention of meaning is only possible when it has been preceded by accurate comprehension.

The data regarding lexical retention from the case studies reported in Chapters Nine and Ten are more complex in various ways than data from Chapter Three. In one respect the data are more precise and detailed in that partial changes in knowledge of word meanings are recorded, but in another respect they are less clear since they are based on participants' self-evaluation of knowledge of words which will include unrecognised miscomprehension of targeted words. Two features of the data from these studies stand out. One point is that for some participants MLD use rarely results in looked-up words being rated as definitely known; words previously rated as definitely unknown are typically rated as probably unknown or probably known after MLD use. The other striking feature of the data is the high proportions of looked-up items for which word knowledge retention is recorded; between 40% and 70% of looked-up items show some benefit of MLD use. These figures suggest that the general dichotomy of *know/not know* applied to lexical retention may be more a product of the instruments used to investigate retention than a true reflection of the nature of L2 vocabulary retention through MLD use.

In the studies reported in Chapters Four to Six, our focus was not on retention of word meaning but on retention of knowledge of contexts in which the target words were presented. The reason for focusing on recognition of contexts rather than production of retained word meaning was that we believed that greater levels of retention take place than those reported in studies which have focused on retention of word meaning. In the study reported in Chapter Four, participants were able to match target words with gapped definitions in almost 50% of cases one week after the learning session, while this figure was close to 55% for participants in the study reported in Chapter Five. The study reported in Chapter Six investigated retention over three weeks, as opposed to one week in the two previous studies, and using an instrument in which participants chose from ten possible answers rather than five. Under these harder circumstances, figures dropped drastically to a 10% success rate.

11.1.4 Variation among participants

In our review of data for production, comprehension, and retention of lexical knowledge, most of the data have been presented in terms of mean scores for experimental groups of participants. However, these averages fail to reflect the fairly large individual differences among dictionary-using participants. For example, in the study reported in Chapter Four, over 50% of the participants in this group were able to give acceptable translation equivalents for all but one to three of the 24 target items. Such data suggest that these participants usually understood sufficient information from the dictionary definitions to work out or guess at the meanings of the target words. On the other hand, around a quarter of participants gave acceptable equivalents for half or fewer of the items. These participants were either often unable to understand the information in the definitions or were unable to use the information that they could understand to make accurate guesses as to Japanese translation equivalents for the words.

Data for sentence production using the target words show similar diversity among participants, but to a lesser extent. Over a third of Dictionary group participants in this study demonstrated acceptable use of over two-thirds of target words in English sentences, while over a third were only able to produce acceptable sentences for half or fewer of the target words.

11.1.5 Variation among items

The studies reported in Chapters Three to Six all focused in some way on the issue of word difficulty and part of speech. In the studies reported in Chapters Three to Five, the sets of target words were composed of words of only one word class, either adjectives or verbs, while in the study reported in Chapter Six there were two sets of target words: one set of verbs and one set of adjectives. We will review two aspects of the study reported in Chapter Six: target item selection and data regarding comprehension of the individual target items.

The main criteria for selection of target words were word frequency and their not being known, or at least not recognised, by the participants. Comparable word frequency was ensured by the use of three measures: the word frequency bands in Collins COBUILD

English Dictionary (1995); the JACET 8000 word list; and numbers of occurrences in the 50-million word COBUILD Direct corpus. Word frequency is one important factor regarding likely participant knowledge of the selected items. In addition to frequency, the JACET 8000 word list includes consideration of word inclusion in high school text books used in Japan: a factor also likely to determine participants' prior knowledge of items. Finally, a pre-test was given to identify, and exclude, words recognised by the participants.

In order to compare comprehension and retention of English verbs and adjectives it was necessary to use comparable sets of target items for the two parts of speech. In terms of word frequency and participants' prior knowledge of items, we may confidently conclude that the two sets of items were comparable. However, regardless of target words' uniformity in terms of frequency, comprehension rates for individual words vary widely from word to word, both for participants using dictionary definitions and for those with example sentences for the target words, as shown in Table 11.1.

Table 11.1

Participant comprehension rates for target words in study reported in Chapter Six

	0% - 20%	21% - 40%	41% - 60%	61% - 80%	81%-100%
Dictionary definitions					
Adjectives	4	4	6	4	2
Verbs	1	6	5	6	2
Example sentences					
Adjectives	10	6	4	0	0
Verbs	10	4	5	1	0

Some target words (*blatant*, *morbid*, *poignant*) were understood by no, or very few, participants with access to dictionary definitions, while other words were understood by all or almost all the participants in this group (*obese*, *elope*, *amputate*). Such results suggest that factors other than word frequency or prior familiarity will be important in determining how likely MLD users may be to understand, and learn, looked-up words. This will be among the issues that we will address as we go on to consider effective dictionary use, the nature of vocabulary learning via MLD use, and means of investigating vocabulary learning that takes place in this context.

The data drawn together from the studies reported in Chapters Three to Ten focused on MLD-using language learners' low comprehension rates for L2 words, wide variation among participants abilities to perform various dictionary-related tasks, and wide variation among targeted words. Each of these issues is especially pertinent to one or more of the fundamental aspects of research into vocabulary acquisition through MLD use which we will now go on to discuss: the relationship between MLDs and their users; the nature of vocabulary acquisition through MLD use; and methods for investigating vocabulary acquisition in this context.

11.2 Monolingual learner dictionaries and their users

Our concern here is the relationship between language learners and MLDs. We will consider the contexts in which these dictionaries are used, and the purposes for which they are consulted. We will also focus on the process of use, from the trigger that leads the language learner to reach for the dictionary to the three steps of locating the information sought, comprehension of that information, and application of the information. This will lead us to consider comprehension success rates and consequent questions about the suitability of this type of dictionary for L2 learners of different proficiency levels. From this general discussion of MLD difficulty levels we will reflect on different word types in the dictionary and their varying difficulty levels for the dictionary user. We will then consider the contexts or circumstances in which there is no relationship between the language learner and the monolingual learner dictionary; i.e. why, very often, language learners do not use an MLD at all but rely exclusively on bilingual dictionaries. Finally, based on this discussion, we will suggest means by which the monolingual learner dictionary and the language learner may be made more suitable partners for each other.

11.2.1 Contexts and purposes for MLD use

Put in the most general terms, the MLD may be seen as a repository of information about words, with this information either provided in the target language or, as in the case with grammar or pronunciation, in partially codified form. The basic purposes of MLD use may be summarised as being either to gain knowledge or to confirm knowledge about particular words: to find out or to check. A dictionary user's need to confirm lexical knowledge depends on three conditions: the prior possession of some kind of lexical knowledge, a lack of sufficient confidence about the accuracy or completeness of that knowledge, and a current or anticipated need to apply the knowledge. We will now consider three sources of lexical information for a word prior to dictionary consultation: previous encounters with the word; the environment of the word in a particular text; and the written form of the word.

11.2.1.1 Previous encounters

There are various causes for our incomplete knowledge about a previously encountered word. Our knowledge regarding an L2 word may also have never been more than partial, or its learning may have been restricted to a particular social domain or linguistic context. For example, a student of English may have learnt the word *homie* from the lyrics of a rap song and know that it means *friend* but be unsure whether it refers to a particular type of friend. The learner may also be unsure whether the use of the word is restricted to rap song lyrics, to Black English vernacular, to young people's language, to American English, or to spoken English generally. Or a foreigner visiting Hiroshima may learn that "*Nambo?*" means "How much (is it)?" or even "How old (are you)?" but not be sure whether this expression would be understood in Tokyo or Osaka, or whether its use would make the speaker an object of ridicule.

Knowledge of words gained through previous encounters with L2 words may also fade with time, as may our confidence about that knowledge. We may, for example, remember that "*bouillabaisse*" is the word for a specific French dish, or that it is a kind of soup, but may not remember what kind of soup it is or where the dish originates.

In each of these instances, the potential dictionary user does have partial knowledge about a word but also has a need for more accurate information, and a need for greater confidence to apply that knowledge correctly.

11.2.1.2 Context

Unless foreign language learners are embarking on an intensive vocabulary learning course and using word lists, they generally encounter new L2 words in a written or spoken context, such as the passage used in the research described in Chapter Three, the children's story in Chapters Eight and Nine, or the song lyrics in Chapter Ten. After encountering a word in a particular context, we may gain various types of information regarding the word about which we are more or less confident. The location of the word within a sentence may tell us about the part of speech of the word, while the subject matter of the text, or the word's lexical environment may give us information about likely meanings of the word. These sources of information may also accord with, or conflict with, previously acquired or assumed information about the word. Where they accord, we may expect confidence regarding that knowledge to increase, whereas where there is apparent conflict, confidence about the knowledge would fall. A couple of examples from the texts used in Chapter Ten may serve to illustrate a possible circumstance.

1. *Magnolias hopeful in her auburn hair.*
2. *You wear your scars like medals…*

In 1, from a song by Joni Mitchell, there are two targeted words: *magnolias* and *auburn*. The words on either side of *auburn*, *her* and *hair*, would help confirm a participant's belief that *auburn* is a colour word used to describe hair. For *magnolias*, the context is less helpful. With the exception of the word *hopeful*, the rest of the line may help confirm someone's understanding of *magnolia* as a type of flower, but the adjective *hopeful* applied to *magnolias* may hinder rather than confirm participants' previous understanding of the word.

Line 2, from a song by Delirious, also illustrates the sometimes misleading nature of context. Within this narrow context, readers' previous beliefs regarding the meaning of *scars* may be challenged by the word *wear* which is usually only used for clothes, glasses, jewellery or other accessories.

While contexts such as those for *magnolias* or *scars* may not add to or confirm a learner's knowledge of a word, the influence of unhelpful contexts such as these may be tempered both by participants' understanding of the subject matter of the whole song and by their perception of song lyrics as typically containing poetic or non-literal uses of words.

11.2.1.3 Word form

The isolated written form of a word may be a rich source of different kinds of assumptions about the word. The spelling may alert us to the likely part of speech of the word. For example, a learner of English may guess, rightly or wrongly, that a particular unknown word ending *–ive* is an adjective or that a word ending in *–tion* is an abstract noun. For words encountered in a written or spoken context, the surrounding words may serve to confirm or challenge these hypotheses. Or the learner may reach for the dictionary, recognising that there is insufficient evidence to justify reliance on these assumptions. As for pronunciation, a learner of English seeing the word *tow* for the first time may assume that its pronunciation would rhyme with that of *blow* or *show* or with *now* and *how*. The learner may also realise that, without further information, it is impossible to make a confident guess as to which pronunciation is correct and turn to a dictionary.

One further point is that the first encounter with a particular word form may not be the first encounter with that word in some other form: as part of a word family of which another word is known; as a word that has been heard before but never seen in writing; or in the form of a known cognate or loan word in another language. Again, language learners may recognise that knowledge ascribed to these secondary or indirect sources is not always reliable and seek to confirm assumed meanings by reference to a dictionary.

11.2.2 User purposes for MLD use

Incomplete prior knowledge, uncertain lexical context, and unknown or half-recognised word forms, then, are three basic circumstances in which a language learner may reach for a dictionary, with either the purpose of finding out something specific about an unknown word or seeking to confirm specific knowledge about a word.

Uncertain or incomplete information about an L2 word is not usually, in itself, sufficient reason to use a dictionary. If it were, dictionary use would be a ceaseless activity since the complexity of the most common words, such as the articles in English, means that we constantly encounter L2 words for which our knowledge is far from complete. For language learners to reach for a dictionary, they usually require the recognition of a lack of knowledge, the awareness of a need to know, and a belief that the dictionary can supply the knowledge sought in an accessible way.

The most straightforward circumstance for dictionary use is, then, the encounter with an unknown known or partially known L2 word in a social or linguistic context in which this lack of knowledge limits the learner's understanding of that context. From a productive language perspective, this may be equated with the need for a word with a particular meaning or register in a written or spoken context that the learner is creating.

In addition to clear context-triggered needs that lead to dictionary use, there are three other factors that may lead language learners to consult a dictionary or to pay attention to particular information in an entry. The first factor is what has been termed a "dictionary reflex". For some language learners, the simple presence of an unknown word in a text is sufficient to send them to the dictionary, regardless of whether the word could be understood by studying the context of the word, and regardless of whether knowledge of the unknown word is necessary for comprehension of the text.

The second factor is language learner interest: a desire beyond the need of the moment to know about a particular encountered word, or to know lexical information other than that needed. For example, where only knowledge of a word's meaning is needed, the dictionary user may be interested in the register, the pronunciation, or the collocates of the word.

The third factor, part of a language learner's desire to learn the foreign language, is the motivation to retain looked up lexical information for future unspecified use. This may also be an element, whether conscious or unconscious, both in the two factors listed above and in the primarily need-driven consultation of a dictionary entry.

Similar learning-motivated considerations may account for a language learner' decision to turn to an MLD rather than to another kind of dictionary. In many circumstances or environments in which a language learner chooses to use an MLD, he or she could have chosen to use a bilingual dictionary. This is increasingly true as electronic dictionaries gain in popularity, since many models contain both monolingual and bilingual learner dictionaries; availability is increasingly not a factor determining which dictionary type to consult. It is, then, worth considering learners' reasons for consulting an MLD. One important reason may be the belief that MLDs contain more detailed, accurate, and up-to-date information than that typically available in bilingual dictionaries. The importance of this factor will vary a large amount according to the availability and quality of bilingual dictionaries in the learner's L1 and of monolingual dictionaries in the target language, as well as to the attitude or behaviour to MLDs observed in language teachers or more advanced language learners.

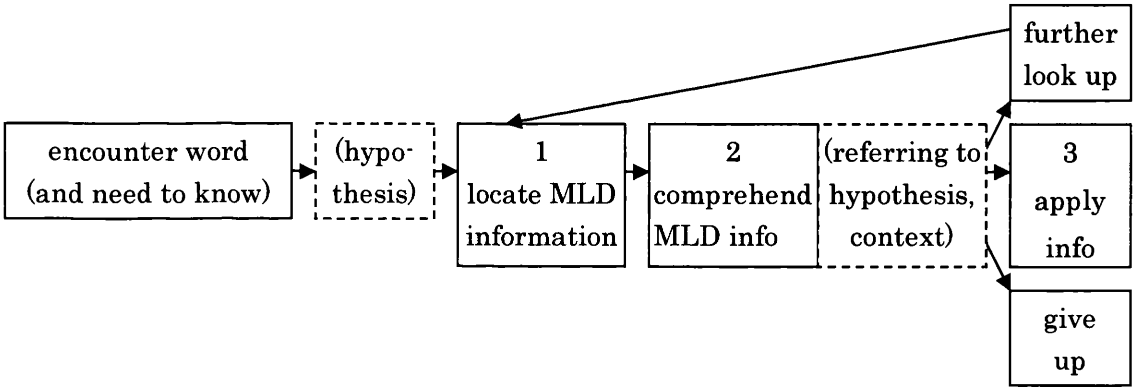
A second reason for using an MLD rather than a bilingual dictionary may be that it is seen itself as a medium for learning the target language. Especially in environments in which language learners have little contact with the foreign language they are learning, MLD use will provide almost unlimited opportunities for communication of meaning via the target language. A third reason for choosing to use an MLD may be the perception of language learning in terms of the aphorism "No pain, no gain": the belief that the greater effort involved in using an MLD, as compared with bilingual dictionary use, will be rewarded with greater learning of the looked up lexical information.

11.2.3 Process of use

Following the circumstances described above which lead a language learner to reach for a dictionary, the actual process of looking up information usually follows three basic steps: location of the information sought, comprehension of that information, and application of the information in some way. The first step may also be preceded by hypothesis forming, whether conscious or unconscious, about the information sought: for example, about the likely sense of the word, the probable pronunciation, or likely collocates. The flow chart in Figure 11.2 illustrates this process.

Figure 11.2

Process of dictionary use



We will now go on to look at the three main stages in dictionary use (labelled 1, 2, and 3 in Figure 11.1) of the location of required information, comprehension of that information, and application of the information. As we do this, we may gain a fuller understanding of the problems faced by MLD users as reflected in the low success rates in the data from the studies in terms of comprehension, production, and retention of looked-up words.

11.2.3.1 Location of information

The location of specific required information in a dictionary may proceed by a series of stages: locating the entry for the word, locating the appropriate sense of the word (if the word is polysemous), and locating the specific information sought within the entry for the sense. The method of finding the entry for the word will, of course, differ according to whether the language learner is using a printed dictionary or an electronic dictionary but in both cases, with the possible exception of phrases or idioms, this step is relatively straightforward provided that the spelling of the word is known, and provided that the dictionary user is sufficiently familiar with the alphabetic system used for the organisation of the dictionary. With polysemous words, finding the relevant sense may not be so simple, may take a considerable amount of time, and in some cases may not be successful. With a paper dictionary, the appropriate sense is located by scanning the entry, in some dictionaries with the assistance of "signpost" words (Tono, 2001: 174-187). With an electronic dictionary, the user either chooses the appropriate sense from a "menu screen" for the entry or scrolls down the entry until the appropriate sense is identified. Again, if the MLD contained within an electronic dictionary contains "signpost" words, these may also aid the dictionary user. Locating the specific information sought within the entry should not be difficult once the dictionary user has become familiar with the layout of the entries: where, and in what form, are recorded the meaning, the pronunciation, grammatical information, or information regarding collocates.

11.2.3.2 Comprehension of information

Once the dictionary user has located the information sought in the dictionary, the next major step is comprehension of the information. In a monolingual learner dictionary, most of the information in an entry is provided in the target language, with grammar and pronunciation usually presented in a partially codified form, and information about register or language variety often given in some abbreviated form. In addition, pictures are provided for a small number of headwords in most MLDs. Each type of information, and each form in which it is presented, has its own challenges as far as comprehension is concerned. Here, we will focus mainly on word meaning and the conveyance of this meaning via definitions and, to a lesser degree, in example sentences.

Comprehension of the written text of an MLD entry depends, on the user's side, upon the user's proficiency in the target language, in reading ability in the language, and in familiarity with the location and presentation style of different information within the dictionary entries. On the dictionary side, users' difficulties in understanding definitions and example sentences may depend on the length of the definitions and example sentences, on their syntactic complexity, and on the difficulty of the vocabulary used. A further consideration in this matter is that greater accuracy and completeness of information provided through definitions and example sentences may often only be achieved at a cost of increased definition or sentence length, greater syntactic complexity and, possibly, harder vocabulary.

11.2.3.3 Application of information

Comprehension of the information sought in a dictionary is not usually an end in itself. Neither is it an isolated stage in the process, since its focus depends on the context in which the word was encountered, on the purpose for looking up the word, and on the application of the information that has been understood. Regardless of the specific application of the information sought in the dictionary, one step that links comprehension of information from the dictionary with its application is the fitting of the information to the context or purpose for which it was sought. For language learners looking up an unknown word in an L2 text, the application of the information is to have a better understanding of the text, for whatever reason: in order to understand and comply with what the text says; to understand the attitude or intentions of the writer of the text; or with an educational purpose such as answering comprehension questions or writing an essay.

Where the word is encountered in a word list, or presented as one of a set of words as in a simple bilingual dictionary, an MLD user's application of the knowledge gained through consulting a dictionary may be to select the correct or most suitable word from the list, to learn the information for a test, or to use the word in writing or speaking.

In language learning contexts, as we have discussed, a dictionary user's proposed application of information found in the dictionary may be twofold: to complete a task such as reading a text or writing an essay and to learn the looked up lexical information. This dual purpose, with the second perhaps unstated and at least partly unconscious, suggests that, in an educational or testing context, researchers' perception of L2 vocabulary acquisition as a dichotomy, either incidental or intentional, may be both misleading and inaccurate.

11.3 Comprehension, guessing, confidence, and retention

A number of unexpected aspects of the data from the various studies reported in this thesis raise questions about the nature of L2 vocabulary acquisition through MLD use. From the studies reported in Chapters Three to Six, low comprehension rates for target words, wide variation in comprehension rates among participants, and wide variation in comprehension rates for target words suggest that comprehension of MLD entries is not a straightforward matter for intermediate level L2 learners, and that the difficulty of this task varies widely both from person to person and from word to word. Data from the longitudinal studies reported in Chapters Eight to Ten showed unexpected falls in numbers of targeted words rated as definitely known by the participants, as well as, for some participants, an apparent reluctance to rate looked-up words as definitely known. We will now consider what light these various data may shed on our understanding of how L2 learners' vocabulary may be affected by MLD use.

We have suggested that from the perspective of language learners, MLD entries may be viewed as short reading comprehension passages. This perspective may help us as we go on to consider the central importance of guessing and confidence in MLD users' comprehension and retention of lexical information from MLD entries. Our main concern throughout this thesis has been, and remains, the retention of lexical information rather than its comprehension. However, we will begin here by considering comprehension of lexical information because, in most cases, comprehension is a prerequisite for retention, and the level or nature of that comprehension will inevitably affect the retention of lexical information.

Comprehension is not usually the black-and-white, all-or-nothing, issue that the word may suggest. Just as word knowledge may be seen in relative terms, in term of completeness or depth of knowledge, so may an MLD user's comprehension of the information in an entry. For example, a dictionary user may understand only one type of the information in an MLD entry, such as meaning or grammar, or only some part of one type of information. From another perspective, an MLD user may understand information from an MLD entry sufficiently to make sense of a reading passage, or sufficiently to use the looked-up word in writing.

11.3.1 Guessing

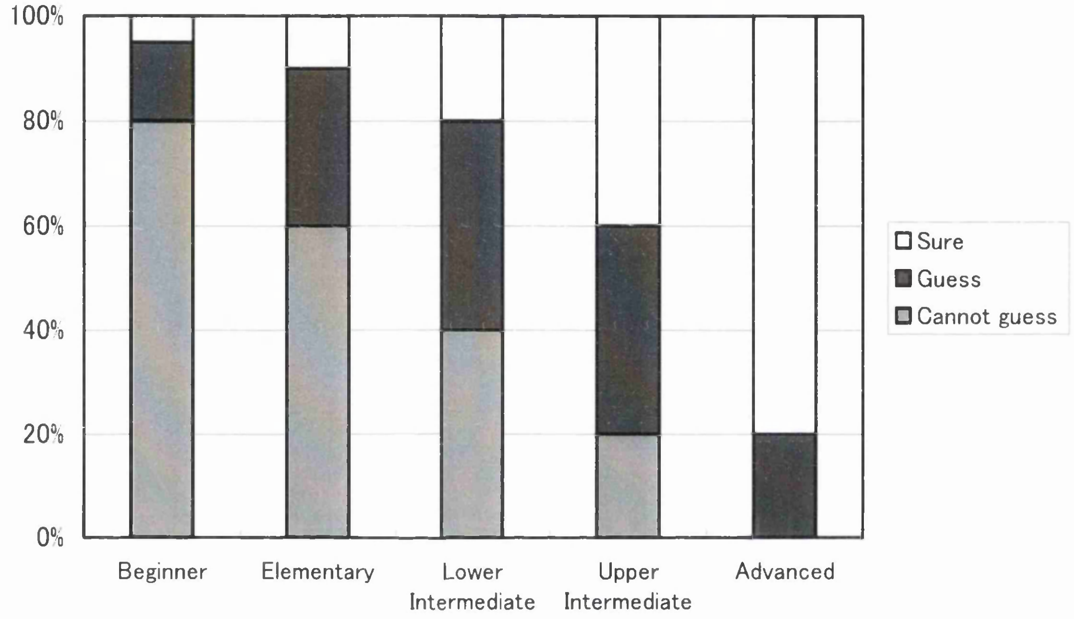
To the depth or extent of a language learner's comprehension of information in a dictionary entry may be added the dimension of confidence: the extent to which an MLD user is confident of having accurately understood the information that was sought in the entry for a word.

The reason that we can neither assume complete comprehension of MLD entries nor absolute confidence about having understood them is due to the central role of guessing in much MLD use. For language learners to be able to guess at the meaning of information in an MLD, they need to understand a certain amount of the definition or other content of the dictionary entry consulted. At one extreme, if nothing is understood of the information sought, then even guessing is not possible. At the other extreme, with higher levels of proficiency in the target language, there will be fewer occasions on which MLD users consciously guess at the meaning of looked-up information since they are more often sure of its meaning; for these users, the meaning written in the

MLD is transparent. Figure 11.2 provides a simplified illustration of the varying importance of guessing with MLD use at different levels of L2 proficiency.

Figure 11.2

A model for the role of guessing in MLD use at different L2 proficiency levels



According to this model, language learners who consult an MLD will be faced with one of three circumstances:

a) They are unable to understand enough of the relevant content of the entry to even guess at the information sought (Cannot guess);

b) They can only understand some of the content of the entry, and so use this limited content to guess at the information sought (Guess);

c) They understand enough of the relevant content to be sure of the information sought in the MLD (Sure).

As shown in Figure 11.2, circumstance a) will be encountered most frequently by low proficiency level language learners, circumstance b), guessing, most frequently by intermediate proficiency level learners, and circumstance c) most frequently by more advanced level L2 learners.

This model of the situation regarding the role of guessing in MLD is a simplification in three main ways: in the lack of gradation within each circumstance, in the lack of gradation from each circumstance to those bordering it, and in the absence of any representation of partial comprehension and partial guessing. Neither does it represent the reality that being sure of information is no guarantee of being correct; we may be sure of something but mistaken. The model does, however, give an indication of the importance of guessing in MLD use at all levels, especially at the intermediate levels of L2 proficiency at which language learners may begin to use an MLD.

11.3.1.2 The nature of guessing

One example of a definition, with the defined word deleted, may illustrate MLD users' varying needs to guess at different levels of L2 proficiency:

A _________ is a large insect that lives in hot countries and makes a loud high-pitched noise.

Assuming no prior knowledge of the word, and no information inferable from the word form or from the context in which the word was encountered, we may imagine that a beginner in the language of the MLD, if at a level to read the text at all, would not be able to understand enough of the vocabulary in the 18-word definition to guess at the meaning of the word. Many elementary or lower intermediate level learners, perhaps unfamiliar with the words *insect* and *pitched*, may only guess that this definition refers to some kind of living creature. Others may know the word *insect*; some of these might guess that the definition refers to a cicada, some might guess, for example, that it refers to some other kind of insect, and others would feel that they had insufficient information to guess. Upper intermediate and advanced level learners may understand every word of the definition, but may still not be sure of which particular insect it refers to. They may guess that this definition refers to a cicada, provided that they know of this insect and know of no other insect that matches this description. However, they may not be sure that this guess is correct; with their knowledge of the world they may not know that a cicada is large, they may not know whether there are cicadas in the country of the target language, or they may know of insects other than cicadas which this definition would describe equally well.

11.3.2 Confidence

The example of *cicada* illustrates two types of guessing which may lead to a lack of confidence. We may term an MLD user's confidence regarding information guessed at in this way as general confidence. We will also consider two further types or aspects of confidence relevant to MLD use: medium-inherent confidence and instrumental confidence.

11.3.2.1 General confidence

MLD users' confidence regarding a looked-up word is their confidence in the accuracy of their comprehension of information obtained through looking up a word in the MLD. A lack of such confidence may have two main causes. The first is where crucial parts of the content providing the information cannot be understood: specific words in a definition or example sentence or, for example, the grammar or other information in the entry. Confidence in these circumstances depends on the users' belief that they have understood the information correctly, or have understood enough of it to make an accurate guess as to the intended meaning of the looked-up information.

The second factor affecting general confidence arises in situations where the MLD user has been able to understand all the language used to convey the meaning of the information sought in the dictionary. In many such cases MLD users will be sure of their accurate comprehension of the information they looked up. On other occasions, as we considered for *cicada*, even full understanding of the definition or other information in an MLD entry may provide only limited confidence regarding accuracy of comprehension of the actual meaning of that information. This may be because of a lack of confidence as to whether the MLD user's interpretation of the meaning of the information corresponds to the meaning intended by the dictionary. Alternatively, the meanings language learners gain from an MLD entry may differ from the meanings which they had previously believed, with the result that the learner is unsure whether his or her understanding of the information in the MLD is correct.

11.3.2.2 Instrumental confidence

As we saw in the case studies reported in Chapters Eight to Ten, a further type of confidence regarding comprehension of information in an MLD appears to be closely related to the proposed application of the information looked up in the MLD. We can term this instrumental confidence: the possession of sufficient confidence about word knowledge to use that knowledge in a particular way. There are various "low-risk" applications of information for which only a low level of confidence about the information is required. These include, for example, information for words looked up while reading for pleasure; even if our comprehension of the meaning of a looked up word is mistaken, our enjoyment of the book may be unimpaired, and we may in any case encounter the word further on in our reading and so gain a clearer understanding of its meaning. A high-risk application for reading might be comprehension of the instructions on a bottle of pills; miscomprehension could lead to illness or death. There are also low-risk and high-risk speaking environments and situations. We may, for example, feel confident enough to use a looked up L2 word with fellow L2 learners in a casual conversation but not confident to use the same word in an academic presentation to native speakers of the language.

In the case studies reported in Chapters Eight to Ten, a low-risk situation was the self-assessment of L2 word knowledge that would not be followed by any requirement to provide evidence of that knowledge. A higher risk situation, resulting in lower confidence levels, was the self-assessment of L2 word knowledge that was followed immediately by a test in which participants had to provide translation equivalents for the words they had rated as definitely known. Instrumental confidence would, then, be a reflection of the perceived consequences for a language learner of making a mistake in understanding or using a particular word in the L2.

11.3.2.3 Medium-inherent confidence

A more fundamental, and general, lack of confidence about comprehension of information in an MLD entry may not be related to individual words and our knowledge of the world but related to the language through which the information is conveyed. If the dictionary user holds the belief, consciously or unconsciously, that "believing is seeing – in my own language", then the MLD will be insufficient to give the user total confidence about what is read and understood in an MLD entry. For such language learners, the true meaning of an L2 word is only found in its L1 equivalent or description, as found in a bilingual dictionary. In other words, although such MLD users may be in no doubt as to the meaning contained in an MLD entry, they may not feel confident that they really know the word until they see the meaning for it in their L1.

Following MLD use for targeted words, the participant in the study reported in Chapter Nine rated almost all 67 previously unknown looked-up words as known to some degree, with the majority rated as probably known, but he rated only four of the items as "definitely known". It is very likely that for many more than four of the words the participant would have had no doubt about the accuracy of his understanding gained through the MLD entries. However, it does appear that the L2 medium used for comprehension of the items disqualified the majority of them from becoming viewed, and rated, as definitely known. In other words, he appears to have had low medium-inherent confidence regarding the lexical information encountered through MLD use.

11.3.3 Confidence and retention

Confidence about lexical information obtained through MLD use will affect both whether a looked-up word is used at all and the particular circumstances in which it is used. In addition, we may expect confidence about the accuracy of our understanding of lexical information to affect its retention in various ways. First, if we accept that our actual use of words assists in their retention then we must recognise that the non-use or restricted use of words about which we are not confident will deny or restrict any benefit in terms of retention that would otherwise have been gained through their use. Secondly, the less sure language learners are of the accuracy of their comprehension of lexical information in an MLD entry, the less likely they are to expend any conscious effort on the learning of this lexical information. Since retention is related to effort expended on learning a word, we may expect lower retention rates for looked-up words about which MLD users do not feel confident.

Thirdly, we might expect lower levels of unconscious learning for words about which a language learner is unsure than for words about which the learner has no doubts. Just as a word may be unconsciously tagged by a language learner as informal or as American English, so too may a word be tagged as not properly understood and, as a result, unconsciously neglected. For learners for whom "properly understood" is equated with having seen the meaning or other required information in their L1, such as the participant reported in Chapter Nine, this tagging may be applied to all words for which information was obtained only from an MLD.

11.3.4 Factors affecting retention

The confidence of a language learner regarding his or her comprehension of information looked up in an MLD is an important factor in the learner's retention of such information. It does not, however, exist in isolation from two other factors which will also have an important influence on the retention of information about looked-up words: the amount of effort expended by the MLD user to understand the information sought, and MLD users' success in understanding this information correctly. As a contrast to MLD use, we will also refer to bilingual dictionary (BD) use as we consider the effect, and interaction, of these factors in L2 vocabulary acquisition. Table 11.3 illustrates the interaction of these three elements and suggests the likely effect that that their different combinations may have on L2 vocabulary acquisition rates for looked-up words.

Table 11.3

Effort, success, and confidence in the retention of looked-up words

Effort		Success		Confidence		Retention	Typical Dictionary
Yes	No	Yes	No	Yes	No		
Y		Y		Y		1.	(MLD)
Y		Y			N	2.	(MLD)
	N	Y		Y		2.	(BD)
	N	Y			N	3.	(BD)
	N		N		N	4.	(BD)
Y			N		N	5.	(MLD)
	N		N	Y		5.	(BD)
Y			N	Y		6.	(MLD)

Here we will focus mainly on the beneficial effect of dictionary use that is successful in terms of comprehension, represented in the top half of the table. Mistaken comprehension and the retention of this mistaken information, as shown in the bottom half of the table, is also an important issue but as it largely represents a mirror image of successful dictionary use, we will not discuss it in detail at this stage.

The table suggests that particular combinations of the three elements may be more typically experienced by users of either monolingual or bilingual dictionaries. It should be stressed, however, that all combinations may be found for users of either dictionary type. Also, as illustrated in Figure 11.2, L2 proficiency level would also be a major factor in determining levels of dictionary user effort, success rates, and confidence levels and, consequently, retention of looked up information in different dictionary types.

For successful dictionary use, we may expect the greatest retention to occur where more effort is required for comprehension of information and where dictionary users are confident of having understood the information correctly. This combination is probably most likely to occur with MLD use, either with advanced level L2 users or for easier entries with intermediate level users. The next highest levels of retention may be expected from either of two combinations of elements. One, perhaps typical for bilingual dictionary use, is where there is less effort involved but greater confidence about successful comprehension. The other combination, more typical of MLD use, is where there is greater effort involved but less confidence about successful comprehension of the information sought. Finally, lowest levels of retention of information gained through dictionary use may be expected when there is less effort involved in understanding information together with lower levels of confidence about the accuracy of comprehension; this might be an outcome of bilingual dictionary use in which, for example, the user is not sure of the meaning of the L1 equivalent provided in the entry.

As we have considered the role of confidence regarding L2 knowledge gained through MLD use, we have come to recognise two things. One is that our understanding of the nature of L2 vocabulary acquisition may need to be changed in recognition of the role of confidence as a factor that will either impede or promote vocabulary acquisition. Secondly, we need to recognise that vocabulary testing instruments that fail to address the factor of learner confidence about word knowledge will continue to limit our understanding of L2 vocabulary acquisition. As we go on to consider methodological issues concerning the acquisition of L2 words through MLD use, one question we will address is how to investigate, or at least take account of, the role of confidence in vocabulary acquisition.

11.4 Methodological issues

A range of methodological issues concerning the investigation of L2 vocabulary acquisition through dictionary use have been considered in the review of previous research in Chapter Two and addressed through the studies reported in Chapters Three to Ten. We will now focus on three issues that are central to learning through dictionary use and have been in the background of many of the studies discussed, but which have not yet been considered in any depth: how to investigate changing vocabulary knowledge; how to deal with the wide range of types of lexical items for investigation; and how to test for both lexical knowledge and confidence about that knowledge.

11.4.1 Investigating changing vocabulary knowledge

Within the word *acquisition* itself is a recognition of a changing state: from having nothing to having something, or from having less of something to having more of it. With this understanding, all studies in L2 vocabulary acquisition can be said to investigate changing vocabulary knowledge in some way. However, the vast majority of studies, certainly among those concerned with dictionary use, restrict their investigations to one framework, addressing one or two aspects of change for a small set of words. Studies are usually comprised either of stages i) to iii) or of stages i) to v):

i) the establishment of a prior state of knowledge for each of a set of words, usually determined through a pre-test;

ii) a learning condition for the targeted words, either presented openly or disguised within another task such as reading comprehension;

iii) an immediate post-test to determine the effect of the learning condition;

iv) a "forgetting condition" of an interval in which participants have no contact with the targeted words;

v) a delayed post-test to determine the effect of the "forgetting condition" on the state of the targeted words.

The aim of this framework is to obtain two or three "snapshots" of the state of a set of target words and thereby to gauge the effect on these words of one or two experimental learning or forgetting conditions. However, snapshots of moving things may either be blurred and out of focus because of the movement, or clear and in focus but fail to capture the movement that is an essential part of the matter under investigation. "Snapshot tests", too, may either fail to capture the fluctuation or movement of lexical items within our mental lexicon, or may produce data that are unclear and difficult to interpret.

To continue this photographic analogy, many studies investigating L2 vocabulary acquisition through dictionary use have employed instruments that only record black and white images, data for items that are either known or unknown. Data for grey, partially known, items are not recorded or not taken account of. Where such partial knowledge is recorded in some way, as for example with the partially correct items in the studies reported in Chapters Three to Six, such data are either discounted and ignored or combined with data for fully known items. Yet it is among these "grey" items, where the instability of the language learner's L2 lexicon is most evident, that we may expect to see the greatest effect of different learning conditions.

A further respect in which "grey" data are ignored is the almost exclusive attention paid in studies to average or typical response rates. From this perspective, differences between individual participants or between individual words become largely invisible, only seen as a problem when they result in high standard deviation rates or results that fail to provide evidence of significant differences between experimental groups.

A film may be defined as a series of snapshots that, together, tell a changing story. In the same way, instruments that are capable of recording the state of words a number of times within any given condition may be better able to record the story of changing word knowledge. In order to obtain a picture that shows the situation with more clarity, one further requirement is for a more sensitive instrument: one that is able to record shades of lexical knowledge, that will both provide a truer picture of the mental lexicon in each "snapshot" and that, through series of snapshots, also will reveal more of the change that takes place among targeted words as a result of different learning conditions.

11.4.2 Targeting words

Much research into L2 vocabulary acquisition through MLD use has focused on small sets of targeted words, with the assumption that these will be typical of words that might be unknown to the participants, typical of dictionary entries in an MLD, or typical of L2 words that language learners look up in dictionaries. Despite this aim of typicality, or representativeness, the selection of items appears to have been either rudimentary, such as including a few words from each major word class, or haphazard, expecting words selected from a chosen text to be somehow representative of the vocabulary of the target language or of the words that language learners may look up.

One way of addressing this issue is to use a focused set of targeted items; either to focus on a specific word or entry type or to identify and focus on entry difficulty levels. Among the studies reported in the first part of this thesis, considerable care was taken in the selection of target words: focusing on words from single word classes and on sets of words within a particular, limited, word frequency band. However, as we have noted, there is a very wide range of word types, entry types, and difficulty levels other than those of word class and word frequency; even sets of words chosen carefully according to one or two aspects may fail to be typical or representative of many other types and levels of word or dictionary entry.

A focus on dictionary entries for targeted words may provide a better measure of parity than word frequency alone, especially when our aim is to compare learning for sets of words of different word classes. For example, sets of targeted words could consist of only words with monosemous dictionary entries, with a set definition length, or of words for which the same defining style is used. Alternatively, word sets could be selected with reference to the volume and type of lexical information contained within dictionary entries; for example, whether entries contain, or do not contain, information regarding pragmatic, register, or regional variety.

The last two factors, defining style and whether or not an entry contains only meaning-related information in the definition, suggest that it may also be possible to establish dictionary entry difficulty levels and for a study to focus on one of these. The classification of entry difficulty levels could be based on information types (e.g., meaning, register, or pragmatics) found within the entry, on definition length, or on the presence or absence of words outside the basic defining vocabulary of the dictionary.

Through focusing on a type of word, a type of entry, or an entry difficulty level, it would be possible for a reasonably small set of target words to represent that type or level much more adequately than a small set of target words claiming to be representative of all the words in the dictionary or all the words a language learner may encounter or look up. Research based on focused sets of words such as these would be much more robust than that for a small general set of items, and much more easily defended against accusations that research with different sets of words alone would be likely to produce different outcomes.

An alternative approach to using focused sets of items is to use very large sets of targeted words; this was the approach adopted in the four case studies reported in this thesis. Sets of twenty or thirty target words, even if chosen with great care, cannot be representative of each type and level of word, of each type of dictionary entry, and of each combination of different types and levels. A study in which there are 300 items, with around 100 of these looked up, will contain reasonable numbers of a wide range of word and entry types; sufficient to be described as typical, at least, of words that might be looked up for a particular text type, and sufficient also for data for some word or entry types to be extracted and examined as required.

11.4.3 Investigating confidence in lexical knowledge

Confidence regarding comprehension of information looked up in an MLD is both an important facet of L2 word knowledge and an important factor in the retention of that information. However, learner confidence about word knowledge has been largely ignored in studies of comprehension and retention of information in MLD entries. As a result, our understanding of the nature of L2 learners' retention of lexical information has remained limited. As we focus on this issue we will consider means by which learner confidence about lexical knowledge may be measured, and suggest ways in which this may increase our understanding of L2 vocabulary acquisition.

As we suggested above, there may be three different types of L2 learner confidence regarding information looked up in an MLD:

i) General confidence: confidence regarding successful comprehension or accurate retention of information from an MLD;

ii) Instrumental confidence: the amount of confidence required to use the information for different purposes;

iii) Medium-inherent confidence: confidence regarding lexical knowledge that is affected by the medium through which this information was obtained.

The three types of confidence listed above, and their effect on L2 vocabulary acquisition, are suppositions for which there is as yet little direct evidence. This lack of evidence is a consequence of the issue of learner confidence about word knowledge having remained largely unexplored. However, recognising its potential importance, we will now go on to consider means by which we may measure the impact that each of these confidence types has on L2 vocabulary acquisition through MLD use.

11.4.3.1 Measuring general confidence

The evaluation of general confidence regarding comprehension of information in an MLD is simply concerned with the questions "Are you sure you understand?" or "How sure are you that you understand?" These may be expressed, unambiguously, as "Are you sure?" or "How sure are you?" Regarding retention, the same questions may be asked but, reflecting the greater complexity of retention as opposed to comprehension, may be interpreted in different ways. "Are you sure?" may be understood to mean "Are you sure that you remember correctly?" or "Are you sure that you understood correctly in the first place?", or a combination of these two. This distinction may not, however, be important unless our concern is with the basis for confidence regarding retained lexical information.

To investigate general confidence, the questions "Are you sure?" or "How sure are you?" can be added to various kinds of vocabulary test items, whether in "supply answer" tests, multiple-choice tests, or some kind of yes/no tests. The addition of these confidence-related questions should be relatively simple to add to these tests and add only a few seconds per item to the time required for the administration of the tests.

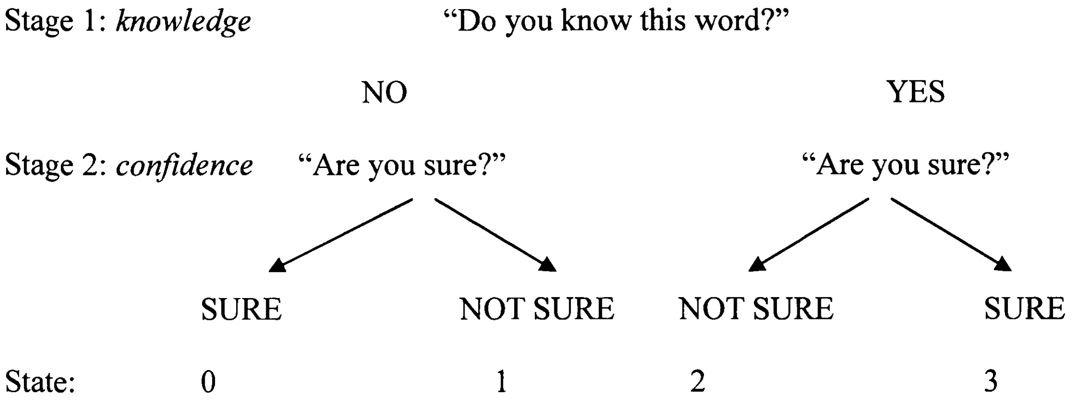
Investigations of confidence may also be either implicitly or explicitly contained within tests such as V_States, used in the studies reported in Chapters Eight to Ten. In the V_States self-evaluation task, the main concern is not confidence about word knowledge but word knowledge itself. However, as we noted in Chapter Ten, participant responses and the descriptions of the states themselves both suggest that confidence about word knowledge is also evaluated and rated. The descriptions of the four states include the following confidence-related language: *I'm not sure*; *I think*; and *I definitely.*

One problem with the implicit inclusion of confidence about knowledge in an instrument such as this is that different participants may interpret the states differently, with some mainly recording the extent of their knowledge of the items and others recording their confidence about this knowledge. There is one further problem with an instrument that combines participant evaluation of acquired word knowledge with confidence about that knowledge without distinguishing between these two aspects of retention. It is that fluctuating confidence levels, especially when affected globally by external circumstances such as impending vocabulary tests, will result in apparent falls in vocabulary levels. When this happens, if we view the instrument as a measure of word knowledge alone, we are compelled to try to explain apparent falls or rises in word knowledge for which there is no discernible cause.

An instrument that distinguishes between participant knowledge of targeted words and confidence about the accuracy of that knowledge instrument may be obtained by asking about word knowledge and confidence about word knowledge in two different stages, as described briefly in Chapter 10. A model for such an instrument is shown in Figure 11.4.

Figure 11.4

A self-evaluation instrument for word knowledge and confidence



As with V_States, data from this instrument would still show words as being in one of four states according to their answers for the Stage 1 and Stage 2 questions:

0: *don't know, sure*; 1: *don't know, not sure*; 2: *know, not sure*; 3: *know, sure*.

Using these two stages of word knowledge followed by confidence about that knowledge, the meaning of each state would be clearer, and data for the two elements of word knowledge and confidence could be more easily extracted for analysis.

11.4.3.2 Measuring instrumental confidence

The impact of instrumental confidence on vocabulary self-evaluation scores was observed in case studies reported in Chapters Eight, Nine and Ten. In these studies, the same self-evaluation task was conducted a number of times, with a supply-meaning test for words rated "definitely known" following only the final self-evaluation task. In one study, the participant was unaware that this final self-evaluation task would be followed by a meaning test and in this final task rated the greatest number of items as definitely known. However, in subsequent studies the three participants were aware that their final self-evaluation task would be followed by a meaning test, and for two of these there was a sizeable drop in numbers of items rated definitely known.

These latter studies unintentionally recorded the effect of instrumental confidence regarding knowledge of the targeted words: "Do you know this word?" compared with "We will test your knowledge of words you claim to know – do you know them?". These results suggest one possible method for investigating instrumental confidence: participants would do the same self-evaluation task twice, once knowing that the task would be followed by a test of the vocabulary and once knowing that no test would follow.

An alternative method of investigating instrumental confidence would be to describe various contexts for using a set of target words: while reading a novel; talking with a native speaker; or giving instructions about a medical operation to be undertaken. With these situations in mind, participants would do a self-evaluation task for the set of words.

11.4.3.3 Measuring medium-inherent confidence

Language learners may have different levels of confidence about L2 word knowledge inherent to the medium through which the information about the words was obtained. There are two obvious media the effect of which could be investigated in this context: bilingual and monolingual dictionary entries. However, since our concern is not to compare comprehension rates, target words would have to be selected on the basis of monolingual definitions being easy to understand, and upon bilingual and monolingual entries for a word containing the same information. A test of comprehension of the target words may or may not be necessary to confirm that participants had understood the information correctly, but participants would be asked how confident they were of their understanding of each of the words. A delayed retention test, again with confidence rating for each item, would also provide valuable data regarding the effect of this type of confidence on the retention of words learned through the different media.

11.5 Conclusion

In this chapter we have reviewed four aspects of the empirical work reported in the thesis: productive MLD use by language learners; learner comprehension rates for MLD definitions; the retention of lexical information gained through MLD use; and variation in comprehension and retention both among participants and among target words. In the light of this review, and of the body of empirical research reported in the thesis, we have also discussed three issues that are central to research into L2 vocabulary acquisition through MLD use: the relationship between language learners and MLDs; the roles of comprehension, guessing, and confidence in MLD users' retention of looked-up lexical information; and various methodological issues concerning the investigation of vocabulary acquisition through dictionary use.

Regarding the relationship between language learners and MLDs, we began by focusing on the contexts in which MLD use, or indeed any dictionary use, may take place. We suggested that two important aspects of these contexts are encounters with L2 words in social and linguistic contexts coupled with incomplete or uncertain knowledge of these words. Where a language learner looks up a word, this can usually be taken to indicate that neither any prior knowledge of the word nor the contexts in which it was encountered provide all the information about the word that the learner requires.

Encounters with L2 words for which a learner has incomplete or unsure knowledge brought us to the third aspect of learner dictionary use: a need to know lexical information which the learner lacks, or about which the learner lacks confidence. We suggested that this need to know that leads to dictionary use could be driven by immediate context-driven requirements for lexical information or by less immediate motivations such as learner interest or for specific language learning reasons. As for the language learner's choice to use an MLD rather than only a bilingual dictionary, we suggested that reasons for this may often be specifically learning motivated: either in terms of a desire for clearer, more accurate information or with the perception of the MLD entry not only as a receptacle for lexical information but also as a medium for language learning itself. We then considered the processes involved within language learner MLD use, from the process of use with its various stages of location of information, comprehension of that information, and application of the information.

Our next focus was on factors involved in the retention of lexical information looked up in an MLD. Specifically, we considered learner comprehension of MLD entries, the role of guessing in this comprehension, and the consequent degrees of confidence about looked up information that may in turn affect its retention by a language learner. We took specific examples of definitions for English words and reflected on the varying role of guessing in comprehension of these definitions for learners of varying L2 proficiency levels.

Our focus on guessing brought us to the issue of learner confidence about the accuracy of information obtained through MLD entries. We suggested that there may be three types of confidence about this information: general confidence, instrumental confidence, and language medium inherent confidence. This in turn led us to consider the contribution of the two factors of learner effort and learner confidence in affecting successful retention of lexical information encountered through dictionary use.

Finally, we looked at three methodological issues at the core of research into L2 vocabulary acquisition in the context of MLD use: how to investigate changing vocabulary knowledge; how to conduct lexical research in a field in which there is such a wide range of types of word or dictionary entry; and how to conduct research that is able to take account of the factor of learner confidence in the retention of L2 lexical information. We suggested that instruments that are capable of investigating the effect of repeated encounters with targeted words may be most suited to the investigation of continuing L2 vocabulary development. As for the wide range of word types that exist, we rejected the idea of focusing on smaller assorted sets of items, as has been the focus of much research in the field to date, in favour of either small focused sets of words of a particular type or very large sets of items that may be more representative of the range of word types that exist. Regarding confidence about L2 word knowledge, we suggested that some instruments do already measure confidence about word knowledge as well as word knowledge retention itself, but that they do not produce data that differentiate between or measure these two aspects of word knowledge independently. We suggested that these instruments could be adapted so that we would have access to separate data both for word knowledge retention and for confidence about retained word knowledge.

Chapter Twelve: Conclusion

The aim throughout this thesis has been to investigate the effect of foreign language learners' MLD use on their knowledge of looked-up words. This aim has been pursued first through an in-depth review of previous research in the field and then by two series of empirical studies into a range of different aspects of L2 vocabulary acquisition through MLD use.

In the first series of experiments we investigated the success of participants' consultation of MLD entries or definitions for a set of target words in terms of comprehension and retention as compared with inference and retention of meaning of the words when encountered in various written contexts. These studies showed a clear advantage for participants with access to information from MLDs over participants who encountered the target words in written contexts, a general advantage for targeted verbs over targeted adjectives, and an overall advantage for sentence style definitions with targeted verbs. These studies also served to highlight some enduring problems that had also been observed in previous research. There were problems stemming from the time participants required per item, resulting in small numbers of targeted words in the studies, with the consequent difficulty of establishing representative sets of items. There were also problems with developing sensitive yet reliable tests of vocabulary retention, especially instruments that would be capable of recording incremental growth in knowledge of targeted items.

The second series of studies was composed of four individual case studies, each conducted over a number of weeks. These studies differed from the first series of studies, and much previous research which rely on standard objective vocabulary tests in two important ways. They all involved repeated encounters with very large numbers of targeted words in the context of extensive reading, with or without the use of an MLD; they involved the repeated evaluation, by the participants, of knowledge of the targeted items; and they involved projections of changing vocabulary knowledge derived from transitional probability matrices. While it is true that the methods employed, the sets of targeted words, the reading materials, and the instruments employed all required some adjustments or change through this series of studies, these studies were successful in producing large amounts of valuable data concerning the development of L2 vocabulary in the two learning circumstances under investigation. Perhaps more importantly, they confirmed the value of the methods and instruments employed to provide projections of, and so comparisons between, the effects of the two learning conditions of L2 reading without and with an MLD.

Our understanding of issues surrounding language learner use of monolingual learner dictionaries is growing and the picture we have of L2 vocabulary acquisition by this means is becoming more complete. As our understanding increases, so too does our appreciation of the complexity involved in certain aspects of this field, such as the range of word and dictionary entry types that MLD users may encounter, together with varying difficulty levels, and the element of confidence as a factor affecting word knowledge retention. In some respects, then, we might say that it is becoming a more difficult area to investigate. However, as new instruments are developed, tested, refined and put into service, useful data regarding the effect of learner dictionary use will increasingly be able to provide dictionary makers, foreign language teachers, and learner dictionary users with information that will aid them in their respective tasks of making useful language learning tools, of providing informed guidance in the use of these tools, and of using these tools effectively.

Table of Contents

Bibliography

Aizawa, K. (1999). *A study of incidental vocabulary learning through reading by Japanese EFL learners.* Unpublished PhD thesis, Tokyo Gakugei University, Tokyo.

Ard, J. (1982). The use of bilingual dictionaries by ESL students while writing. *International Review of Applied Linguistics, 58*, 1-27.

Atkins, B.T.S. (ed.) (1998). *Using dictionaries: Studies of dictionary use by language learners and translators,* (Lexicographica. Series maior, 88). Tübingen: Max Niemeyer Verlag.

Atkins, B.T.S. and Knowles, F. (1990). Interim report on the EURALEX / AILA research project into dictionary use. In T. Magay and J. Zigány (eds.), *BudaLEX '88, Proceedings from the 3rd international EURALEX congress* (381-392). Budapest: Akadémiai Kiadó.

Atkins, B.T.S. and Varantola. K. (1998). Monitoring dictionary use. In Atkins, B.T.S. (ed.), *Using dictionaries: Studies of dictionary use by language learners and translators* (83-122), (Lexicographica. Series maior, 88). Tübingen: Max Niemeyer Verlag.

Atkins, B.T.S. and Varantola, K. (1998). Language learners using dictionaries: the report on the EURALEX / AILA Research Project on Dictionary Use. In Atkins, B.T.S. (ed.), *Using dictionaries: Studies of dictionary use by language learners and translators* (21-81), (Lexicographica. Series maior, 88). Tübingen: Max Niemeyer Verlag.

Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: A study of language needs and reference skills. *Applied Linguistics 2*, (3), 207-222.

Black, A. (1986). The effect on comprehension and memory of providing different types of defining information for new vocabulary: a report on two experiments conducted for Longman Dictionaries and Reference Division. Cambridge: MRC Applied Psychology Unit (unpublished internal report).

Bogaards, P. (1991). Dictionnaires pédagogiques et apprentissage du vocabulaire, *Cahiers de lexicologie, 59*, 93-167.

Bogaards, P. (1992). French dictionary users and word frequency. In: H. Tommola, K.Varantola, T. Salami-Tononen and J.Schopp (eds.). *EURALEX '92 Proceedings* (51-59). Tampere: Department of Translation Studies, University of Tampere.

Bogaards, P. (1998). Des dictionnaires au service de l'apprentissage du français langue étrangère. *Cahiers de Lexicologie, 71* (1), 127-167.

Bradley, I. and Meek, R. L. (1986). *Matrices and society.* Harmondsworth: Penguin.

Brown, C. (1993). Factors affecting the acquisition of vocabulary: Frequency and saliency of words. In T. Huckin, M. Haynes and J. Coady (eds.), *Second language reading and vocabulary learning* (263-286). Norwood, NJ: Ablex.

Carduner, J. (2003). Productive dictionary skills training: What do language learners find useful? *Language Learning Journal*, *28*, 70-76.

Chikamatsu, N. (1996). The effects of L1 orthography on L2 word recognition. *Studies in Second Language* Acquisition, *18* (4), 403-432.

Christianson, K. (1997). Dictionary use by FL writers: What really happens? *Journal of Second Language Writing, 6* (1), 23-43.

Craik, F.I.M. and Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behaviour*, *11*, 671-684.

Craik, F.I.M. and Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104,* 268-294.

Crowther, J. et al. (1995). *Oxford advanced learner's dictionary, 5th edition.* Oxford: Oxford University Press.

de Groot, A.M.B. (1993). Word type effects in bilingual processing tasks: Support for a mixed representational system. In R Schreuder and B Weltens (eds.) *The Bilingual Lexicon* (27-51). Amsterdam: Benjamins.

Ellis, N.C. and Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*, 559-617.

Ellis, N. and Beaton, A. (1995). Psycholinguistic determinants of foreign language vocabulary learning. In B. Harley (ed.), *Lexical issues in language learning* (107-165). Ann Arbor, MI: John Benjamins.

Fischer, U. (1994). Learning words from context and dictionaries: An experimental comparison, *Applied Psycholinguistics, 15*, (4), 551-574.

Fowles, J. (1981). *The Collector*. New York: Dell.

Fraser, C.A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition, 21* (2), 225-241.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (ed). *Language development (Vol. 2): Language, thought, and culture* (301-334). Hillsdale, NJ: Lawrence Erlbaum.

Goodale, M. (1995). *Collins COBUILD English dictionary workbook.* London: HarperCollins.

Grace, C.A. (1998). Retention of word meanings inferred from context and sentence-level translations: Implications for the design of beginning level CALL software. *The Modern Language Journal 80* (3), 327-339.

Grinstead, W.J. (1915). An experiment in the learning of foreign words. *Journal of Educational Psychology 6*, 242-245.

Hanks, P. (1987). Definitions and explanations. In J.M. Sinclair, (ed.), *Looking up. An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English Language Dictionary* (116-136). London: Collins.

Harvey, K. and Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics, 18* (3), 253-278.

Hatch, E. and Brown, C. (1995). *Vocabulary, semantics and language education.* Cambridge: Cambridge University Press.

Hoey, M. (2005). *Lexical priming.* London: Routledge.

Hornby, A. S. (ed.) (1980). *Oxford advanced learner's dictionary of current English, 3rd edition.* Oxford: Oxford University Press.

Horst, M. and Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review, 56* (2), 308-328.

Hu, M. and Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13* (1), 403-430.

Hulstijn, J. H. (1993). Retention of inferred and given word meanings: experiments in incidental vocabulary learning. In P.J.L. Arnaud and H. Béjoint (eds.), *Vocabulary and applied linguistics* (113-125). London: Macmillan.

Hulstijn, J.H., Hollander, M. and Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign-language students: The influence of marginal glosses, dictionary use, and reoccurrence of unfamiliar words, *The Modern Language Journal, 80* (3), 327-339.

Iwai, Y. (2000). *Dictionary use in context and vocabulary acquisition*, Unpublished MA dissertation. University of Birmingham.

Johns, T. (1991). Should you be persuaded – two samples of data-driven learning materials. In T. Johns and P. King (eds.) *Classroom Concordancing* (1-16). (ELR Journal 4) Birmingham: Birmingham University.

Kipfer, B.A. (1984). Methods of ordering senses within entries. In R.R.K. Hartmann, (ed.) *LEXeter '83 Proceedings*, 101-108. Tübingen: Max Niemeyer Verlag. Reprinted in R.R.K. Hartmann, (ed.), *Lexicography: Critical concepts in lexicography.* (Volume 3, 182-190). London: Routledge.

Knight, S. (1994). Dictionary use while reading: the effects on comprehension and vocabulary acquisition for students of different verbal abilities. *Modern Language Journal, 78* (3), 285-299.

Komuro, Y. (2004). *Macmillan essential dictionary workbook for learners of English* [in Japanese]. Tokyo: Macmillan Languagehouse Ltd.

Koyama, T. and Takeuchi, O. (2004). Comparing electronic and printed dictionaries: How the difference affected EFL learning. *Jacet Bulletin 38*, 33-46.

Krantz, G (1990). *Learning vocabulary in a foreign language: A study of reading strategies.* Göteborg: Acta Universitatis Gothoborgensis.

Laufer, B. (1997). The lexical plight in second language reading. In J. Coady and T. Huckin (eds.) *Second language vocabulary acquisition* (20-34). New York: Cambridge University Press.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P.A. Arnaud and H. Bějoint (eds.) *Vocabulary and applied linguistics.* London: Macmillan.

Laufer, B. and Hill, M. (2000) What lexical information do L2 learners select in a CALL dictionary and how does it affect word recognition? *Language Learning and Technology* 3 (2), 58-76.

Lewis, C.S. (1950). *The lion, the witch and the wardrobe.* London: Geoffrey Bles.

Luppescu, S. and Day, R. (1993). Reading, dictionaries, and vocabulary learning. *Language Learning 43* (2), 263-287.

MacFarquhar, P.D. and Richards, J.C. (1983). On dictionaries and definitions. *RELC Journal: A Journal of Language Teaching and Research in Southeast Asia 14* (1), 111-124.

MacWhinney, B. (1997). *Cognitive approaches to language learning*. Michigan: MIT Press.

Matsuda, T. (ed.) (1992). *Kenkyusha's English-Japanese dictionary for the general reader.* Tokyo: Kenkyusha.

Matsuda, T. (Ed.) (1994). *An encyclopedic supplement to the dictionary for the general reader.* Tokyo: Kenkyusha.

McKeown, M. (1993). Creating effective definitions for young word learners. *Reading Research Quarterly 28* (1), 16-31.

Meara, P. (2001). *V_States, v.0.3.* [computer programme] Swansea: University of Wales Swansea.

Meara, P. (2005). Reactivating a dormant vocabulary. In S.H. Foster-Cohen, M. del Pilar and J. Cenoz (eds.), *EuroSLA yearbook 5* (269-280). Amsterdam: John Benjamins.

Meara, P. and Buxton, B. (1987). An alternative to multiple choice vocabulary tests, *Language Testing 4* (2), 142-151.

Meara, P. and English, F. 1988. Lexical errors and learners' dictionaries. ERIC DOC. Ed. 654 321, 2-16.

Meara, P. and Rodriguez Sanchez, I. (1994). Matrix models of vocabulary acquisition: an empirical assessment. *CREAL Occasional Paper, 1.*

Miller, A. and Gildea, P.M. (1985). How to misread a dictionary, *AILA Bulletin* (1985), 13-26.

Miller, A. and Gildea, P.M. (1987). How children learn words. *Scientific American 257* (3), 94-99.

Milton, J. (2001). Introspecting on vocabulary learning from an informal task. In *Vocabulary Acquisition Research Group conference proceedings*. Retrieved from http://www.swan.ac.uk/cals/calsres/events/02_contents/02_Milton.htm.

Mochizuki, M. (2003). JACET 8000 compared with other vocabulary lists. In M. Murata, S. Yamada and Y. Tono (eds.), *Proceedings of ASIALEX 2003: How can dictionaries help human and machine learning* (378-383)? Urayasu: The Asian Association for Lexicography.

Mondria J.-A. and De-Wit Boer, M. (1991) The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics. 12* (3), 249-267.

Mullich, H. (1990). *Die Definition Ist Blod! Herubersetzen mit dem einsprachigen worterbuch das franzosische und Englische Lernerworterbuch in der hand der Deutschen Schuler.* (Lexicographica. Series maior 37). Tübingen: Max Niemeyer Verlag.

Nagy, W.E., Herman, P.A. and Anderson, R.C. (1985). Learning words from context. *Reading research quarterly 20*, 233-253.

Nagy, W.E. and Herman, P.A. (1985). Incidental vs. instructional approaches to increasing reading vocabulary. *Educational Perspectives 23*, 16-21.

Nakamura, T. (2001) Japanese EFL learners' cognitive processing of English regular and irregular words. *Bulletin of the Faculty of Education, Hiroshima University 50* (1), 153-162.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Boston: Heinle and Heinle.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P. and Coady, J. (1988). Vocabulary and reading. In R. Carter and M. McCarthy (eds.), *Vocabulary and language teaching* (97-110). London: Longman.

Nesi, H. (1998). Defining a shoehorn: the success of learners' dictionary entries for concrete nouns. In Atkins B.T.S (ed.) *Using dictionaries: Studies in the use of L2 dictionaries (*Lexicographica. Series maior), (159-178). Tübingen: Max Niemeyer Verlag.

Nesi, H. (2000). *The use and abuse of EFL dictionaries. How learners of English as a foreign language read and interpret dictionary entries* (Lexicographica. Series maior, 98). Tübingen: Max Niemeyer Verlag.

Nesi, H. and Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. *System, 22* (1), 1-15.

Neubach, A. and Cohen, A.D. (1988). Processing strategies and problems encountered in the use of dictionaries. *Dictionaries, 10,* 1-20.

Nist, S.L. and Olejnik, S. (1995). The role of context and dictionary definitions on varying levels of word knowledge. *Reading Research Quarterly, 30* (2), 172-193.

Nuccorini, S. (1994). On dictionary misuse. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (eds.), *Euralex 1994 proceedings* (586-597). Amsterdam: Amsterdam New University.

Palmberg, R. (1987). Patterns of vocabulary development in foreign language learners. *Studies in second language acquisition, 9*, 201-220.

Proctor, P. et al. (1995). *Cambridge international dictionary of English.* Cambridge: Cambridge University Press.

Ringbom, H. (1987). *The role of the first language in foreign language learning.* Clevedon: Multilingual Matters.

Rodgers, T.S. (1969). On measuring vocabulary difficulty: an analysis of item variables in learning Russian-English vocabulary pairs. *IRAL, 7*, 327-343.

Ronald, J. and Tajino, A. (2005). A comparison of paper and electronic monolingual dictionaries: Location, comprehension, and retention of secondary senses. In

V.B.Y. Ooi, A. Pakir, I. Talib, L. Tan, P.K.W. Tan and Y.Y. Tan (eds.) *Proceedings of ASIALEX 2005: Words in Asian cultural contexts* (255-261). Singapore: National University of Singapore.

Rundell, M. (2006). Learners' dictionaries. In K. Brown (ed.) *The Elsevier encyclopedia of language and linguistics, 2nd edition* (volume 6, 739-743). New York: Elsevier.

Schatz, E. and Baldwin, R. (1986). Context clues are unreliable predictors of word meanings. *Reading Research Quarterly, 21* (4), 439-453.

Schouten-van Parreren, C. (1985). *Woorden leren in het vreemde-talenonderwijs.* [Teaching vocabulary in a foreign language]. Apeldoorn: Van Walraven.

Scott, M. (1999). *Wordsmith tools*, v.3. [Computer software]. Oxford: Oxford University Press.

Seibert, L. C. (1930). An experiment on the relative efficiency of studying French vocabulary in associated pairs versus studying French vocabulary in context. *Journal of Educational Psychology, 21* (4), 297-314.

Sharwood Smith, M. (1994) *Second language learning: Theoretical foundations*. London: Longman.

Sheldon, S. (1990) *Memories of midnight*. New York: Morrow.

Shinmura, I. (ed.) (1976). *Kojien dainihan hoteiban* [*Kojien* revised and updated 2nd edition]. Tokyo: Iwanami Shoten.

Sinclair, J. et al. (1995), *Collins COBUILD English dictionary*, 2nd edition. London: HarperCollins.

Singleton, D. (1999). *Exploring the second language mental lexicon.* Cambridge: Cambridge University Press.

Summers, D. et al. (1995). *Longman dictionary of contemporary English, 3rd edition.* Harlow: Longman.

Tomaszczyk, J. (1979). Dictionaries: Users and uses, *Glottodidactica,* 12, 103-119.

Tono, Y. (2001). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension.* (Lexicographica. Series maior, 106).Tübingen: Max Niemeyer Verlag.

Uryu, Y. (ed.) (1993). *Eigo tango 2001* [English words 2001]. Tokyo: Kawai Publishing.

van Daalen-Kapteijns, M.M., and Elshout-Mohr, M. (2001) The acquisition of word meanings as a cognitive process. *Journal of Verbal Learning and Verbal Behavior*, 20, 386 – 389.

Waring, R. (1999). *Tasks for assessing receptive and productive second language vocabulary.* Unpublished Ph.D. Thesis. University of Wales, Swansea.

Wesche, M. and Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13-40.

White, W.H. (1988). *Vocabulary acquisition from reading and the ESL learner.* Dissertation, University of Southern California, Los Angeles.

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), 303-325.

Wingate, U. (2002). *The effectiveness of different learner dictionaries: An investigation into the use of dictionaries for reading comprehension by intermediate learners of German* (Lexicographica. Series maior, 112). Tübingen: Max Niemeyer Verlag.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Appendix 3.1 Reading text

Tony Rizzoli had grown up in a place called Hell's Kitchen, in New York. Geographically, it was located in the middle of the West Side of Manhattan, between 8th Avenue and the Hudson River, and its northern and southern boundaries ran from 23rd to 38th Streets. But psychologically and emotionally Hell's Kitchen was a city within a city. The streets were ruled by (1) *armed* gangs, and murder contracts were sold at a hundred dollars.

The residents of Hell's Kitchen lived in dirty apartment blocks, full of lice, rats and cockroaches. There were no bathtubs, and the youths solved the shortage in their own way; they jumped naked into the water off the Hudson River docks, where the dirty water from the Kitchen's streets emptied into the river. The docks stank of the (2) *stagnant* mass of dead, (3) *swollen* cats and dogs.

The street scene provided an endless variety of action. A fire engine answering an alarm… a gang fight on the roofs, a wedding procession… a ball game on the sidewalk… a chase after a runaway horse… a shooting… The only playground the kids had were the streets, the roofs, the trash-filled vacant lots and – in the summer time – the (4) *noisome* waters of the river. And over everything, the (5) *acrid* smell of poverty. That was the atmosphere in which Tony Rizzoli had grown up.

Tony Rizzoli's earliest memory was of being knocked down, and having his milk money stolen. He was seven years old. Older and bigger boys were a constant threat. The route to school was dangerous, and the school itself was a (6) *relentless* battleground. By the time Rizzoli was fifteen years old he had developed a strong body and (7) *unnerving* skill as a fighter. He enjoyed fighting, and because he was good at it, it made him feel (8) *invincible.* He and his friends put on boxing matches at Stillman's Gym.

From time to time, some of the gangsters dropped in to keep an eye on the fighters they owned. Frank Costello appeared once or twice a month, along with Joe Adonis and Lucky Luciano. They were amused by the boxing matches that the youngsters put on, and they found it (9) *diverting* to bet on their fights. Tony Rizzoli was always the winner, and he quickly became a favorite of the gangsters.

One day while Rizzoli was changing in the locker room the young boy overheard a conversation between Frank Costello and Lucky Luciano. 'The kid's making me rich,' Luciano was saying. 'I won five thousand on him last week.'

'You going to put a bet on his fight with Lou Domenic?'

'Sure. I'm betting ten thousand.'

'What odds will you get?'

'Ten to one. But what does it matter? Rizzoli's a certainty.'

Tony was not certain what the conversation meant. He went to his older brother, Gino, and told him about it.

'Wow!' his brother exclaimed. 'Those guys are betting big money on you.'

'But why? I'm not a professional.'
Gino thought for a moment. 'You've never lost a fight, have you, Tony?'
'No.'
'What probably happened is that they made a few small bets for fun, and then when they saw what you could do they began betting seriously.'
The younger boy replied. 'It doesn't mean anything to me.'
Gino took his arm and said, 'It could mean a lot to you. To both of us. Listen to me, kid...'

The fight with Lou Domenic took place at Stillman's Gym on a Friday afternoon and all the big boys were there – Frank Costello, Joe Adonis, Albert Anastasia, Lucky Luciano and Meyer Lansky. They enjoyed watching the (10) *budding* young fighters, but what they enjoyed even more was the fact that they had found a way to make money on the kids.

Lou Domenic was seventeen, a year older than Tony and five pounds heavier. But he was not ready for Tony Rizzoli's boxing skills an (11) *unerring* killer instinct. The fight was five rounds. The first round went easily to young Tony. The second round also went to him. And the third. The gangsters were already counting their money. 'The kid's going to grow up to be a world champion,' Lucky Luciano said. 'How much did you bet on him?'
'Ten thousand,' Frank Costello replied. 'The best odds I could get were fifteen to one. The kid's already got a reputation.'

And suddenly, something happened. In the middle of the fifth round, Lou Domenic knocked out Tony Rizzoli. The referee began to count... very slowly, looking nervously at the audience.
'Get on your feet, you little fool,' Joe Adonis screamed. 'Get up and fight!'
The counting went on, and even at that slow pace, it finally reached ten. Tony Rizzoli was still on the mat.
'I don't believe it! One lucky punch!'

The men began to add up their losses. It was a lot of money. Tony Rizzoli was carried to one of the dressing rooms by Gino. Tony kept his eyes tightly closed, (12) *petrified* that they would find out he was conscious and do something terrible to him.

Appendix 3.2 Entries for target words

1. armed

1 Someone who is **armed** is carrying a weapon, usually a gun. *A third man escaped and police believe he may be armed. ...a barbed wire fence patrolled by armed guards... The rebels are well organized, disciplined, and very well armed.*

ADJ

2. stagnant

2 **Stagnant** water is not flowing, and is therefore often dirty, smelly, and unhealthy.

ADJ- GRADED

3. swollen

1 If part of your body is **swollen**, it is larger and rounder than usual, usually as a result of injury or illness. *My eyes were so swollen I could hardly see.*

ADJ-GRADED

4. noisome

If you describe something or someone as **noisome**, you mean that you find them extremely unpleasant: a literary word. *Noisome vapours arise from the mud left in the docks... His noisome reputation for corruption had already begun to spread.*

ADJ-GRADED:
usu ADJ n
= noxious

5. acrid

An **acrid** smell or taste is strong and sharp, and usually unpleasant. *The room filled with the acrid smell of tobacco. The plant has an unpleasant odour and an acrid taste.*

ADJ-GRADED:
usu ADJ n
= pungent,
bitter

6. relentless

1 Something bad that is **relentless** never stops or never becomes less intense. *The pressure now was relentless.* **relentlessly** *The sun is beating down relentlessly.*

ADJ-GRADED
= remorseless

7. unnerving

If you describe something as **unnerving**, you mean that it is startling or very worrying. *It is very unnerving to find out that someone you see every day is carrying a potentially deadly virus. ... her unnerving habit of continuously touching people she was speaking to.*

ADJ-GRADED
= disconcerting

8. invincible

1 If you describe an army or sports team as **invincible**, you believe that they cannot be defeated. *When Sotomayor is on form he is virtually invincible.*

ADJ-GRADED
= unbeatable

9. diverting

If you describe something as **diverting**, you mean that it is amusing or entertaining; an old-fashioned word.

ADJ-GRADED
= enjoyable

10. budding

1 If you describe someone as, for example, a **budding** business man or a **budding** artist, you mean that they are starting to succeed or become interested in business or art. *The forum is now open to all budding entrepreneurs. Budding linguists can tune into the activity cassettes in French, German, Spanish and Italian.*

ADJ:
ADJ n

11. unerring

If you describe someone's judgement or ability as **unerring**, you mean that they are always correct and never mistaken. *These designs demonstrate her unerring Eye for colour and detail... Paul is a thoroughly likeable man with an unerring sense of comedy. She has an unerring instinct for people's weak spots.*

ADJ
usu ADJ n
= unfailing

12. petrified

1 If you are **petrified**, you are extremely frightened, perhaps so frightened that you cannot think or move. *I've always been petrified of being alone... Most people seem to be petrified of snakes.*

ADJ-GRADED
oft ADJ *of* n/-ing,
= terrified

Appendix 3.3 Pre-test and post-test

次の単語について意味を知っているものには○、わからないものにはXをつけてください。意味がわかる単語については日本語か英語で説明してください。

[For the following words, write a circle for the words that you know and an X for words you do not know. Give the meaning of the words you do know, either in Japanese or English.]

1. armed ....

2. stagnant ....

3. swollen ....

4. noisome ....

5. acrid ....

6. relentless ....

7. unnerving ....

8. invisible ....

9. diverting ....

10. budding ....

11. unerring ....

12. petrified ....

Appendix 3.4 Learning task

次の単語を使って英文を上の行に書いてください。　番号________　氏名___________

[On the top lines, write English sentences using the words. Student number, name]

1. armed
2. stagnant
3. swollen
4. noisome
5. acrid
6. relentless
7. unnerving
8. invisible
9. diverting
10. budding
11. unerring
12. petrified

[Instructed after completion] 終わったら下の行に英文を日本語に翻訳してください

[When you have finished, write a Japanese translation of the English sentences]

Appendix 4.1 Entries for target words

Abbreviate

abbreviate **abbreviates, abbreviating,** VERB **abbreviated** If you **abbreviate** something, especially a word or a piece of writing, you make it shorter.

abbreviate *v* [T] to make a word or expression shorter by missing out letters or using only the first letter of each word.: **be abbreviated to**

Appal

appal BrE, **appall** *AmE v* [T] to shock someone by being very bad or unpleasant

appal **appals, appalling, appalled**; spelled VERB **appall** in American English. If something appals you, it disgusts you because it seems so bad or unpleasant.

Amputate

amputate **amputates, amputating,** VERB **amputated**. If a surgeon **amputates** someone's arm or leg, he or she cuts all or part of it off in an operation because it is diseased or badly damaged.

amputate *v* [I,T] to cut off someone's arm, leg, finger, etc. during a medical operation.

Blab

blab *v* [I] *informal* to tell secret information to someone who is not supposed to know it: [+ **to**]

blab **blabs, blabbing, blabbed**. If someone VERB **blabs** about something secret, they tell people about it: an informal word.

Bode

bode *v*
2 bode well/ill (for) *especially literary* to be a good or bad sign for the future.

bode **bodes, boding, boded**. If something VERB **bodes** ill, it makes you think that something bad will happen in the future. If something **bodes** well, it makes you think that something good will happen: a formal word.

Cackle

cackle **cackles, cackling, cackled**. If someone VERB **cackles**, they laugh in a loud unpleasant way, often at someone else's misfortune.

cackle *v* [I] to laugh in a loud, unpleasant way, making short high sounds.

Cadge

cadge *v* [I,T] *BrE informal* to ask someone for food or cigarettes because you do not have any or do not want to pay; MOOCH *AmE*: **cadge sth from/off**

cadge **cadges, cadging, cadged**. If someone VERB **cadges** food, money, or help from you, they ask you for it and succeed in getting it: used mainly in informal British English.

Canoodle

canoodle canoodles, canoodled, canoodling. VERB In British English, if two people are **canoodling**, they are kissing and cuddling each other a lot; an informal word.

canoodle *BrE old-fashioned* if two people canoodle, they kiss and hold each other in a sexual way.

Chomp

chomp *v* [I + **away/on**] to bite food noisily.

chomp **chomps, chomping, chomped**. If a VERB person or animal **chomps** their way through food or **chomps** on food, they chew it noisily: an informal use.

Coerce

coerce **coerces, coercing, coerced**. If you VERB **coerce** someone into doing something, you make them do it, although they do not want to: a formal word.

coerce *v* [T] to force someone to do something they do not want to do by threatening them: **coerce sb into doing sth**

Dilate

dilate v [I, T] if a part of your body dilates or if something dilates it, it becomes wider.

dilate **dilates, dilating, dilated**. When V-ERG things such as blood vessels or the pupils of your eyes **dilate** or when something **dilates** them, they become wider or bigger.

Elope

elope **elopes, eloping, eloped**. When two V-RECIP people **elope** they go away secretly to get married.

elope v [I] to leave your home secretly in order to get married.

flunk *v informal* 1 [I, T] to fail a test [T] to give someone low marks on a test so that they fail it

flunk out *phr v* [I] *informal especially AmE* to be forced to leave a school or college because your work is not good enough.

Jilt

jilt *v* [T] to end a relationship with someone

jilt **jilts, jilting, jilted**. If someone is VERB **jilted** by the person who they are having a romantic relationship with, that person ends the relationship suddenly in a way which is surprising and upsetting: an informal use.

Perspire

perspire **perspires, perspiring, perspired**. VERB When you **perspire**, a liquid comes out on the surface of your skin, because you are hot or frightened: a formal word.

perspire *v* [I] to become wet on parts of your body, especially because you are hot or have been doing hard work; SWEAT

Pooh-pooh

pooh-pooh *v* [T] to say that you think that an idea, suggestion, effort etc is silly or not very good

pooh-pooh pooh-poohs, pooh- poohing, pooh- VERB **poohed**. If someone **pooh-poohs** an idea or suggestion, they say or imply that it is foolish, impractical,or unnecessary: an informal word.

Raze

raze **razes, razing, razed**. If buildings, VERB villages or towns are **razedor razed** to the ground, they are completely destroyed.

raze *v* [T] to completely destroy a town or building; **raze sth to the ground** (= destroy it so that nothing is left)

suss **susses, sussing, sussed**. In British English, if you **suss** a person or a situation, you realize or work out what their real character or nature is: an informal word.

Trounce

trounce **trounces, trouncing, trounced**. If you **trounce** someone in a competition or contest, you defeat them easily by a large score; an informal word.

trounce *v* [T] to defeat someone completely

Waft

waft *v* [I always + adv/prep] to mov through the air [+ **up/along/off etc**]

waft **wafts, wafting, wafted**. If sounds, scents, or smoke **waft** through the air, or if something such as a light wind **wafts** them, they move gently through the air.

Whinge

whinge **whinges, whingeing, whinged**. If you say that someone **is whingeing** about something, you think that they are complaining in an annoying way about something unimportant; used in informal British English.

whinge *v* [I] *BrE* or *AustrE* to keep compla an annoying way

Wince

wince *v* [I] **1** to suddenly change the expre your face as a reaction to something pa upsetting **2** to suddenly feel very uncomfo embarrassed because of something that something you remember etc **wince** **thought/idea/memory**

wince **winces, wincing, winced**. If you **wince**, the muscles of your face tighten suddenly because you have felt a pain or because you have just seen, heard, or remembered something unpleasant.

She had to have both legs amputated following the crash in April 1993.

Appal

He said he was appalled at the way animals were treated.

Mr Peters said he had been appalled by the conditions in which the housemaids live.

His ignorance appals me.

Blab

Her mistake was to blab about their affair.

But one of the gang has blabbed to police, starting a violent witch-hunt to find the traitor.

She'll start blabbing it out to the whole class.

Blare

It had a loudspeaker that would blare out the latest hits.

Continuous country music blared from the radio.

Outside, car horns and trumpets blared.

Bode

The perilous state of their economy does not bode well for their future.

These developments boded ill for the religious peace within the armed forces.

Coupled with last year's hot summer, this bodes well for British tourism.

Cackle

Scowcroft, Gates, and others cackled with laughter.

Cackling joyously, he fetched out a bottle.

Tinker was falling about cackling.

She was shocked when she saw him canoodling with R

Chomp

I chomped hungrily through the large steak.

Miguel chomped on his fresh stick of gum.

He took the ice from his cup and began chomping on i

Coerce

One of them said he was coerced into signing a confes

They said the statements were coerced, and the four after police failed to find any physical evidence.

They believed that coercing children to leave widespread.

Comply

Foreign companies could operate in Britain as long as with British law.

His failure to comply with treatment created a li situation.

He said NATO would respond with great force if the comply.

Cringe

I still cringe when I remember my ignorance.

My boyfriend cringes at the things I say but I'm ju fun-loving.

They all thought it was hilarious, but I was embarrassment.

Dilate

When we're with people we find attractive our eyes di

His pupils were dilated almost to the rims of his iris w

Your partner will be examined internally to check ho has dilated.

I looked up, feigning surprise.

Flunk

If you don't go back to school, you'll flunk out.

He flunked the subject.

If you don't start studying math he's going to flunk you.

Jilt

Billy Ing had developed affections for a Mexican girl who eventually jilted him.

Plenty of girls get jilted.

He never showed emotion except for a flicker when he talked about Kathryn jilting him.

Perspire

I perspire a lot during a race but I have a tube through which I can drink.

Re-apply sunscreen every two hours, especially if you have been swimming or perspiring.

Hot food cools you off because it lets you perspire.

Pooh-pooh

He pooh-poohed suggestions that he rest.

He had pooh-poohed their warnings.

It's easy to pooh-pooh the whole idea of cultural diplomacy and royal visits.

Raze

They'd kill everyone then raze the village to the ground.

The Serbs had razed the police station during the night.

They were buried alive for more than 36 hours in an apartment block razed by Friday's Greek earthquake.

Maggie likes to organize people and tends to sulk if sh her own way.

Suss

I can suss out the smell of garlic at twenty feet.

American keeper Brad Friedel had sussed the danger.

Haven't you sussed that one out yet?

Trounce

Jeremy Bates, Britain's No 1 tennis player, was trounce No 4 seed Cedric Pioline of France.

Australia then trounced England 24-4.

In the 1980 election, Reagan trounced Carter.

Waft

A faint, very aromatic scent of fire smoke wafted towar

Soft, romantic music wafted through the luxurious hote

She complained after smoke from Mr Legg's cigarette her house.

Whinge

He told members to stop whingeing about workers' righ

She was always whingeing she was short of cash.

Ministers used to phone up and whinge.

Wince

Dragging herself upright, she winced at the stab of pai

Gordon now winces at the mention of Frame.

Nicole winced as a sharp pain shot through her.

appal ....

....

blab ....

....

bode ....

....

cackle ....

....

cadge ....

....

canoodle ....

....

chomp ....

....

coerce ....

....

dilate ....

....

elope ....

....

feign ....

....

flunk ....

....

....

sulk ....

....

suss ....

....

trounce ....

....

waft ....

....

whinge ....

....

wince ....

....

[Instructed after completion] 終わったら単語の下の行に単語の意味を書いてください
[When you have finished, write the meaning of each word under the line for the wor

2.

If a surgeon __________ someone's VERB arm or leg, he or she cuts all or part of it off in an operation because it is diseased or badly damaged.

v [I,T] to cut off someone's arm, leg, finger, etc. during a medical operation.

3.

If you _________ something, VERB especially a word or a piece of writing, you make it shorter.

v [T] to make a word or expression shorter by missing out letters or using only the first letter of each word.: **be ________ to**

v [I] *informal* to tell secret infor someone who is not supposed to know it: [+ t

If someone ________ about something secret, they tell people about it: ar informal word.

6.

If someone_________, they laugh in a loud unpleasant way, often at someone else's misfortune.

v [I] to laugh in a loud, unplea making short high sounds.

Example Sentences group

1.

I can probably ________ a lift later.

On the rare occasions that Liam was in the house when we were working, he would ________ cigarettes off my labourers.

None of them will walk anywhere if they can ________ a ride.

2.

Doctors think they might have to ________ her foot.

His left arm had been ________ below the elbow.

She had to have both legs ________ following the crash in April 1993.

3

We changed our name to the "Zaire River Expedition", quickly ________ by the military to ZRE.

Write these in your book, ________ or simplifying as necessary.

The James Taylor Quartet have ________ their name at last.

4.

In 1977, my girlfriend Lynn and I ________ to get married.

While on holiday in Mexico he ______ 15-year-old Mexican girl.

Two young lovers - a street boy and an uppe fresh from convent school - meet, fall in _______.

5

Her mistake was to ________ about their aff

But one of the gang has ________ to police violent witch-hunt to find the traitor.

She'll start ________ it out to the whole clas

6

Scowcroft, Gates, and others ________ with

________joyously, he fetched out a bottle.

Tinker was falling about ________.

4.	faze	elope	mosey	waddle	whinge
5.	hobnob	wince	blab	saunter	swat
6.	smirk	cackle	denigrate	mumble	dilate
7.	douse	atone	evict	sulk	jilt
8.	pooh-pooh	gobble	waft	capitulate	abate
9.	feign	comply	meddle	eavesdrop	blab
10.	revere	raze	denigrate	suss	spatter
11.	perspire	chug	chomp	gloat	baffle
12.	lacerate	abbreviate	teem	cavort	wince
13.	pucker	canoodle	pester	bode	bonk
14.	incur	elope	chomp	atone	fend
15.	saunter	flunk	gobble	cadge	revere
16.	kowtow	coerce	feign	faze	smirk
17.	cavort	perspire	mosey	canoodle	pucker
18.	eavesdrop	gloat	placate	trounce	sulk
19.	abate	droop	waft	coerce	incur
20.	appal	fend	raze	capitulate	irk
21.	jilt	cuss	dilate	meddle	chug
22.	placate	mumble	trounce	amputate	cuss
23.	baffle	hobnob	bode	comply	whinge
24.	spatter	swat	kowtow	pooh-pooh	suss

My both arms have to be amputated because of the traffic accident.

Unless the verb requires two people to be stated, reference to one person is acceptable:

My sister might elope soon.

Guidance for categorizing sentences as unacceptable was given on the following:

An incorrect subject or object of the verb was used:

My brain has to be amputated by the doctors.
I jilted my best friend.

A transitive verb was treated as intransitive, or vice versa:

As the next subject was gymnastics, I feigned.
He was cackling his brother when his parents came back.

Verbs that have no passive form are presented in a passive form.

I was cackled by that guys.

The verb is accompanied by extraneous words:

I jilted with my lover.

No sensible meaning can be drawn from the sentence:

I am abbreviated to your kindness.
My dog razes my garage.
I think she is coerced into dancing a confession.

The sentence demonstrates misunderstanding of the verb:

Hashimoto trounced Obuchi. (In fact, Obuchi trounced Hashimoto.)
I abbreviated some sentences to leave only important ones.
NY is abbreviated to New York.

Finally, for the Questionable category, rater guidelines were simply to keep down the r sentences for this category and for it to be reserved for sentences which were illogic unusual:

I amputated my leg because my leg damaged by the accident.
If our parents agree with our marriage, let's elope soon.
I'm going to canoodle with my boyfriend at six.

Word	O or X	Meaning
abundant	....	....
affluent	....	....
afoot	....	....
akin	....	....
archaic	....	....
assertive	....	....
astounding	....	....
astute	....	....
audacious	....	....
averse	....	....
barbaric	....	....
bereft	....	....
bewildering	....	....
blatant	....	....
bogus	....	....
brash	....	....
callous	....	....
cardiac	....	....
cheeky	....	....
classy	....	....
colossal	....	....
corrugated	....	....
crass	....	....
daft	....	....
defunct	....	....
diffident	....	....
dilapidated	....	....
dogmatic	....	....

fascist ....

fleeting ....

fraudulent ....

futile ....

furtive ....

galore ....

garish ....

gaudy ....

genial ....

gigantic ....

gleeful ....

glum ....

grumpy ....

halcyon ....

haphazard ....

harrowing ....

hoarse ....

holistic ....

hyper ....

idyllic ....

illicit ....

intricate ....

inviolate ....

jittery ....

judicious ....

languid ....

lax ....

lenient ....

neutral ....

nonchalant ....

obese ....

obnoxious ....

ornate ....

ostensible ....

oval ....

palpable ....

pelvic ....

pivotal ....

poignant ....

pretentious ....

quaint ....

rampant ....

randy ....

satirical ....

tangy ....

ubiquitous ....

vehement ....

wan ....

Thank you!

Galore
In large amounts or numbers.

You use galore to emphasise that something you like exists in very large quantities: an informal word used in written English.

Morbid
If you describe a person or their interest in something as morbid, you mean that they are very interested in unpleasant things, especially death, and you find this strange or unwise.

Having a strong or unhealthy interest in unpleasant subjects, especially death.

Gaudy
Clothes, colours, etc that are gaudy are too bright and look cheap.

If something is gaudy, it is very bright-coloured and showy: often used to express disapproval and to suggest that it is vulgar.

Callous
A callous person or action is very cruel and shows no concern for other people or their feelings.

Not caring that other people are suffering.

Eerie
Strange and frightening.

If you describe something as eerie, you mean that it seems strange and frightening, and makes you feel nervous.

Colossal
If you describe something as colossal, you are emphasizing that it is very large.

Extremely large.

Illicit
Not allowed by laws or rules, or strongly disapproved of by society.

An illicit activity or substance is not allowed by law or the social customs of a country.

If something is inviolate, it has not been
be harmed or affected by anything: a form

Lenient
When someone in authority is lenient, th
as strict or severe as expected.

Not strict in the way you punish someone
their behaviour.

Poignant
Making you feel sad or full of pity.

Something that is poignant makes you fee
because it reminds you of something
happened in the past, or because somethin
wanted to happen did not happen.

Defunct
If something is defunct, it no longer exi
stopped functioning or operating.

Not existing anymore or not useful anymo

Fleeting
Lasting for only a short time.

Fleeting is used to describe something
lasts a very short time.

Bereft
If a person or thing is bereft of somethin
longer have it: a formal word.

bereft of hope, meaning, life, etc: c
without any hope etc.

Obese
(Technical) Very fat in a way that is unhea

If someone is obese, they are extremely c
or extremely fat.

The shops are doing their best to entice us with bargains galore.

The games has already brought you chances galore to get rich.

You'll see wildflowers galore in this country – and they're beautiful.

Morbid

More onlookers might have been expected, if only out of morbid curiosity.

You should get away from all these morbid imaginings.

He is not one to get morbid.

Gaudy

He wore shiny shoes and a gaudy Hawaiian shirt.

The interior is gaudy rather than practical.

His brother Ronnie lay in gaudy splendour in a horse-drawn carriage.

Callous

A police officer said the killing was a cowardly and callous act.

I know it sounds awfully cold and callous.

He was caged in a narrow wooden box for five nights by a callous kidnapper.

Eerie

Naomi Wallace's play is eerie, scary, and full of pent-up emotional violence.

They paused, hearing an eerie sound echoing through the woods.

Her works as a young adult were dark, eerie, abstract paintings of people.

Colossal

He won both events by colossal margins.

How did you face up to this colossal task?

The distance between him and them seemed colossal.

She was watching me from under ey
very furtive way.

Mr Dobson accused the government
underhand and furtive over the matter.

Inviolate

Even the Supreme Leader is not inviola

And this secret purpose remains inviola

He has given up some of his previousl
authority.

Lenient

His parents condemned the sentence, v
felt was too lenient.

They were going to be tried by Columb
and they will be accorded lenient treatm

Poignant

Ali described a poignant moment when
few years old.

It is a poignant love story starring Juli
and Robert Carlyle.

The remainder of his story is both po
disturbing.

Defunct

Rosie Boycott was one of the founders
defunct Spare Rib magazine.

One successful scheme has turned
department store into a children's muse

Nevertheless, the Montreal Protocol i
defunct.

Fleeting

The Iraqi Foreign Minister, Tariq Az
fleeting visit to Jordan at the weekend.

For a fleeting moment it appeared as tl
had made a breakthrough.

All this was as beautiful and fleeting as obese.

When the filming was over, I asked the producer if I looked obese.

Averse

The coach is not averse to using the most basic ploys for motivation.

Powell is not averse to the idea.

They are seldom averse to seizing any opportunity to make economics.

Quaint

All the shops are shut due to a quain tradition.

I am aware of a number of quaint past are performed in rural parts of Britain.

Fingleton, in one of his many books, clear how quaint he viewed all this stuffi

A.	haphazard	quaint	defunct	idyllic
B.	galore	nautical	morbid	tangy
C.	gaudy	jittery	obese	hoarse
D.	satirical	illicit	archaic	blatant
E.	lenient	meagre	languid	averse
F.	fleeting	rampant	haphazard	inviolate
G.	akin	cardiac	dreary	bereft
H.	palpable	quaint	adrift	euphoric
I.	obese	hoarse	nautical	morbid
J.	colossal	gleeful	lenient	pivotal
K.	elusive	galore	satirical	furtive
L.	eerie	idyllic	dreary	callous
M.	gaudy	illicit	archaic	languid
N.	rampant	poignant	inviolate	cardiac
O.	bogus	ostensible	defunct	furtive
P.	palpable	fleeting	lenient	tangy
Q.	jittery	akin	gleeful	bereft
R.	callous	eerie	ostensible	pivotal
S.	meagre	elusive	afoot	colossal
T.	poignant	blatant	euphoric	bogus

In large amounts or numbers.

You use ________ to emphasise that something you like exists in very large quantities: an informal word used in written English.

C

(Technical) Very fat in a way that is unhealthy.

If someone is ________, they are extremely overweight or extremely fat.

If you say that you are not ________ to
you mean that you quite like it or quite
it: a formal word.

Not to be ________ to: used to say th
likes to do something sometimes,
something that is slightly wrong or bad f

Example Sentences group

A

All the shops are shut due to a ________ Roman tradition.

I am aware of a number of ________ pastimes that are performed in rural parts of Britain.

Fingleton, in one of his many books, made it clear how ________ he viewed all this stuffiness.

B

The shops are doing their best to entice us with bargains ________.

The games has already brought you chances ________ to get rich.

You'll see wildflowers ________ in this country – and they're beautiful.

C

Fasts and very low-calorie diets do not work for ________ people; they put the weight back on.

In 1980, 6 per cent of men aged 16 – 64 were ________.

When the filming was over, I asked the producer if I looked ________.

D

Marijuana remains the most comm
________ drug in the country.

It prohibited nudity, interracial love, a
________ sex.

Will Jack's marriage survive an _____

E

The coach is not ________ to usin
basic ploys for motivation.

Powell is not ________ to the idea.

They are seldom ________ to s
opportunity to make economics.

C – correct P – partially correct W – wrong

Most answers will be very easy and straightforward to answer – either c wrong. The following provides guidelines to help judge between Cor Partially correct and between Partially correct and Wrong.

Rating for example words are shown on the attached sheet. Please look at t look at the guidelines below:

a) Wrong part of speech

If it's the wrong part of speech – noun instead of adjective or adjective i verb – that counts as P.

b) Meaning missing

If an essential part of the meaning is missing, that counts as P.

c) Too general

If the answer is too general or vague, that counts as P.

Combinations of a) and b), and c)

a) + b) or a) + c) still count as P.
But b) + c) is probably W – see below.

Much too general

大きい [*“okii”– big*] for *obese* counts as W. It is a case of b) + c): missing th of *too fat* (or *unhealthily fat*) and it is too general.

I hope these guidelines will help, but they won't make it all automatic – ot could do it, or a computer could. Use your own judgment. Where answers raters are different, you will need to make a joint decision.

Once again, thanks a lot for your help – I couldn't do it without you!

Suss

To realize something: suss (that); Suss sb/sth out: to understand the important things about someone or something, especially things they are trying to hide.

In British English, if you suss a person or situation, you realize or work out what their real character or nature is; an informal word.

Galore

You use galore to emphasize that something you like exists in very large quantities; an informal word used in written English.

In large amounts or numbers.

Coerce

To force someone to do something they do not want to do by threatening them.

If you coerce someone into doing something, you make them do it, although they do not want to; an informal word.

Morbid

If you describe a person or their interest in something as morbid, you mean that they are very interested in unpleasant things, especially death, and you find this strange or unwise.

Having a strong and unhealthy interest in unpleasant subjects, especially death.

Appal

To shock someone by being very bad or unpleasant.

If something appals you, it disgusts you because it seems so bad or unpleasant.

Perspire

To become wet on parts of your body, especially because you are hot or have been doing hard work.

When you perspire, a liquid comes out on the surface of your skin, because you are hot or frightened; a

Bode

To be a good or bad sign for the future.

If something bodes ill, it makes you
something bad will happen in the future. If
bodes well, it makes you think that some
will happen.

Gaudy

If something is gaudy, it is very bright-co
showy; often used to express disappro
suggest that it is vulgar.

Clothes, colours, etc that are gaudy are too
look cheap.

Dilapidated

A dilapidated building, vehicle, etc is old
bad condition.

A building that is dilapidated is old and in
bad condition.

Abbreviate

If you abbreviate something, especially a
piece of writing, you make it shorter.

To make a word or expression shorter by
letters or only using the first letter of each

Sulk

To show that you are annoyed about so
being silent and having an unhappy ex
your face.

If you sulk, you are silent and bad-tem
while because you are annoyed about so
informal word used showing disapproval.

Callous

A callous person or action is very cruel ar
concern for other people or their feelings.

Not caring that other people are suffering.

Raze
If buildings, villages, or towns are razed, or razed to the ground, they are completely destroyed.

To completely destroy a town or building.

Colossal
If you describe something as colossal, you are emphasizing that it is very large.

Extremely large.

Trounce
To defeat someone completely.

If you trounce someone in a competition or contest, you defeat them easily or by a large score; an informal word.

Illicit
An illicit activity or substance is not allowed by law or the social customs of a country

Not allowed by laws or rules, or strongly disapproved of by society.

Dilate
If a part of your body dilates, or if something dilates it, it becomes wider.

When things such as blood vessels or the pupils of your eyes dilate or when something dilates them, they become wider or bigger.

Furtive
If you describe someone's behaviour as furtive, you disapprove of them behaving as if they want to keep something secret or hidden.

Behaving as if you want to keep something secret.

Whinge
To keep complaining in an annoying way.

If you say that some is whingeing about something, you think that they are complaining in an annoying ways about something unimportant; used in informal British English.

Defunct
If something is defunct, it no longer e
stopped functioning or operating.

Not existing anymore or not useful anymor

Amputate
To cut off someone's arm, leg, finger e
medical operation.

If a surgeon amputates someone's arm or le
cuts all or part of it off in an operation b
diseased or badly damaged.

Fleeting
Fleeting is used to describe something that
very short time.

Lasting for only a short time.

Pooh-pooh
To say that you think that an idea, sugge
etc is silly or not very good.

If someone pooh-poohs an idea or suggest
or imply that it is foolish, impractical, or
an informal word.

Bereft
If a person or thing is bereft of something,
longer have it: a formal word.

Bereft of hope, meaning, life etc: comple
any hope etc.

Ostensible
Seeming to be the reason or purpose of so
usually hiding the real reason or purpose.

Ostensible is used to describe something t
be true or is officially stated to be tru
which you or other people have doubts; a f

If you say that you are not averse to something, you mean that you quite like it or quite want to do it: a formal word.

Not to be averse to: used to say that someone likes to do something sometimes, especially something that is slightly wrong or bad for them.

Suss

I can suss out the smell of garlic at twenty feet.

American keeper Brad Friedel had sussed the danger.

Haven't you sussed that one out yet?

Galore

The shops are doing their best to entice us with bargains galore.

The game has already brought you chances galore to get rich.

You'll see wildflowers galore in this city – and they're beautiful.

Coerce

One of them said he was coerced into signing a confession.

They said the statements were coerced, and the four were released after police failed to find any physical evidence.

They believed that coercing children to leave school was widespread.

Morbid

More onlookers might have been expected, if only out of morbid curiosity.

You should get away from all these morbid imaginings.

He is not one to get morbid.

Appal

He said he was appalled at the way animals were treated.

Mr Peters said he had been appalled by the conditions in which the housemaids live.

His ignorance appals me.

Feign

The referee accused Durie in his feigning injury.

They cannot bear to be alone and may or feign illness to get their own way.

I looked up, feigning surprise.

Bode

The perilous state of their economy doe well for their future.

These developments boded ill for th peace within the armed forces.

Coupled with last year's hot summer, th well for British tourism.

Gaudy

He wore shiny shoes and a gaudy Haw

The interior is gaudy rather than practic

His brother Ronnie lay in gaudy sple horse-drawn carriage.

Dilapidated

We ended up with a dilapidated farmer

The searchers suddenly came upon a dilapidated building.

It was a bit dilapidated but not i condition.

Abbreviate

We changed our name to the "Z Expedition", quickly abbreviated by t to ZRE.

Write these in your book, abbre simplifying as necessary.

The James Taylor Quartet have abbre name at last.

nights by a callous kidnapper.

Jilt
Billy Ing had developed affections for a Mexican girl who eventually jilted him.

Plenty of girls get jilted.

He never showed emotion except for a flicker when he talked about Kathryn jilting him.

Eerie
Naomi Wallace's play is eerie, scary, and full of pent-up emotional violence.

They paused, hearing an eerie sound echoing through the woods.

Her works as a young adult were dark, eerie abstract paintings of people.

Raze
They'd kill everyone then raze the village to the ground.

The Serbs had razed the police station during the night.

They were buried alive for more than 36 hours in an apartment block razed by Friday's Greek earthquake.

Colossal
He won both events by colossal margins.

How did you face up to this colossal task?

The distance between him and them seemed colossal.

Trounce
Jeremy Bates, Britain's No 1 tennis player, was trounced 6-1, 6-1 by No 4 seed Cedric Pioline of France.

Australia then trounced England 24-4.

In the 1980 election, Reagan trounced Carter.

Your partner will be examined internall
how far her cervix has dilated.

Furtive
He cast a furtive glance to left and right

She was watching me from under her e
very furtive way.

Mr Dobson accused the government
underhand and furtive over the matter.

Whinge
He told members to stop whingeing abo
workers' rights.

She was always whingeing she was sho

Ministers used to phone up and whinge

Poignant
Ali described a poignant moment when
few years old.

It is a poignant love story starring Juli
and Robert Carlyle.

The remainder of his story is both po
disturbing.

Cackle
Scowcroft, Gates, and others cac
laughter.

Cackling joyously, he fetched out a bot

Tinker was falling about cackling.

Defunct
Rosie Boycott was one of the founders
defunct Spare Rib magazine.

One successful scheme has turned
department store into a children's muse

Nevertheless, the Montreal Protocol i
defunct.

All this was as beautiful and fleeting as a dream.

Pooh-pooh
He pooh-poohed suggestions that he rest.

He had pooh-poohed their warnings.

It's easy to pooh-pooh the whole idea of cultural diplomacy and royal visits.

Bereft
Without the CND the public would be bereft of truthful facts and figures about nuclear issues.

The U.S. is not a land of angry workers bereft of opportunity.

Jakarta appears bereft of ideas or initiatives that might break the cycle of violence.

Ostensible
The ostensible aim of his visit was to attend a meeting.

That was the ostensible purpose of the controversial trip to Peking.

Farm policy has four ostensible objectives.

Elope
In 1977, my girlfriend Lynn and I eloped and tried to get married.

While on holiday in Mexico he eloped with a 15-year-old Mexican girl.

Two young lovers - a street boy and an upper-class girl fresh from convent school - meet, fall in love and elope.

Obese
Fasts and very low-calorie diets do not work for obese people; they put the weight back on.

In 1980, 6 per cent of men aged 16 – 64 were obese.

When the filming was over, I asked the producer if I looked obese.

on it.

Afoot
Plans are already afoot to expand next y

There are moves afoot to give schools the Internet.

He sensed that something new was afoo

Waft
A faint, very aromatic scent of fire sm towards us.

Soft, romantic music wafted thr luxurious hotel suite.

She complained after smoke from cigarettes wafted into her house.

Akin
I think Mr Lee Kwan Yew is akin t parent – very disciplined and very dem

In most offices the employees are akin unit.

It's neither a play nor a musical but akin to a melodrama.

Blab
Her mistake was to blab about their aff

But one of the gang has blabbed starting a violent witch-hunt to find the

She'll start blabbing it out to the whole

Hoarse
Her own voice came out in a hoarse wh

His voice was scratchy and hoarse, b determined to communicate.

She's hoarse, as if she's been shouting

ostensible illicit averse

poignant afoot galore

callous dilapidated blatant

furtive

1. ....
2. ....
3. ....
4. ....
5. ....
6. ....
7. ....
8. ....
9. ....
10. ....

Page 2

gaudy bereft obese

eerie fleeting morbid

akin defunct hoarse

colossal

11. ....
12. ....
13. ....
14. ....
15. ....
16. ....
17. ....
18. ....
19. ....
20. ....

trounce elope fo

waft suss p

coerce abbreviate d

bode

21. ....
22. ....
23. ....
24. ....
25. ....
26. ....
27. ....
28. ....
29. ....
30. ....

Page 4

whinge amputate c

appal raze c

jilt perspire s

blab

31. ....
32. ....
33. ....
34. ....
35. ....
36. ....
37. ....
38. ....
39. ....
40. ....

2
You use __________ to emphasize that something you like exists in very large quantities; an informal word used in written English.

In large amounts or numbers.

3
Seeming to be the reason or purpose of something but usually hiding the real reason or purpose.

__________ is used to describe something that seems to be true or is officially stated to be true, but about which you or other people have doubts; a formal word.

5
You use __________ to describe some
that is done in an open or very obvious
order to emphasize your shock or surpr
is done in such an open or obvious wa

Something bad that is __________ is v
and easy to see, but the person respons
does not seem embarrassed or ashamed

Page 3 (verbs)

21
When two people __________, they go away secretly together to get married.

To leave your home secretly in order to get married.

22
To force someone to do something they do not want to do by threatening them.

If you __________ someone into doing something, you make them do it, although they do not want to; an informal word.

23
If you __________ something, especially a word or a piece of writing, you make it shorter.

To make a word or expression shorter by missing out letters or only using the first letter of each word.

24
To defeat someone completely.

If you __________ someone in a con
contest, you defeat them easily or
score; an informal word.

25
To realize something: __________
__________ sb/sth out: to unde
important things about someone or
especially things they are trying to hid

In British English, if you __________
situation, you realize or work out wh
character or nature is; an informal wor

2

The shops are doing their best to entice us with bargains __________.

The game has already brought you chances __________ to get rich.

You'll see wildflowers __________ in this city – and they're beautiful.

3

The __________ aim of his visit was to attend a meeting.

That was the __________ purpose of the controversial trip to Peking.

Farm policy has four __________ objectives.

5

Their election promises have, as I turned out to be __________ lies.

Adams was dismissed for a _ professional foul on Neil Shipperley.

This was not as __________ a aggression in Belgium.

Page 3 (verbs)

21

In 1977, my girlfriend Lynn and I __________ and tried to get married.

While on holiday in Mexico he __________ with a 15-year-old Mexican girl.

Two young lovers - a street boy and an upper-class girl fresh from convent school - meet, fall in love and __________.

22

One of them said he was __________ into signing a confession.

They said the statements were __________, and the four were released after police failed to find any physical evidence.

They believed that __________ children to leave school was widespread.

23

We changed our name to the "Z Expedition", quickly __________ by to ZRE.

Write these in your book, _____ simplifying as necessary.

The James Taylor Quartet have _____ name at last.

24

Jeremy Bates, Britain's No 1 tennis __________ 6-1, 6-1 by No 4 seed Ce of France.

Australia then __________ England 2

In the 1980 election, Reagan ________

25

I can __________ out the smell c twenty feet.

American keeper Brad Friedel had _ the danger.

Haven't you __________ that one out

hoarse	28	15	13
dilapidated	22	11	11
ostensible	11	0	11
fleeting	16	9	7
averse	7	1	6
eerie	14	11	3
callous	11	11	0
morbid	0	0	0
poignant	0	0	0
gaudy	17	18	−1
afoot	13	17	−4
blatant	1	10	−9

Verbs	Dictionary Definitions *(N = 35)*	Example Sentences *(N = 37)*	Advantage for Dic. Def
blab	28	0	28
elope	35	10	25
cackle	27	3	24
coerce	22	1	21
perspire	24	6	18
amputate	32	16	16
bode	15	0	15
dilate	13	0	13
jilt	12	1	11
pooh-pooh	15	5	10
sulk	11	1	10
chomp	28	21	7
appal	5	0	5
whinge	12	9	3
raze	14	12	2
suss	19	17	2
feign	10	10	0
trounce	23	23	0
abbreviate	20	23	−3
waft	19	24	−5

appeased
array
assaulting
assume
badgers
bait
banner
bargaining
barrel
bats
beaks
bearing
beckoned
bellowing
blame
blast
blaze
blister
bluebells
bluebottle
boggles
bold
boughs
brambles
brat
briar
briskly
brood

can't abide
centaur
chamber
charred
chimed
chirped
clashing
cloven
clung
cluster
clutching
cocking
concealed
consorts
copper
counsel
cramped
craves
crisp
crockery
crocuses
curse
curtsey
decent
decoy
deformed
delay
dense

disposal
dispute
disheveled
dodging
dominions
drained
dratted
drooped
duke
dunces
dungeons
eerie
engraved
enrage
eternal
evidence
extremely
feats
fetch
fiddling
fidgets
filthy
flask
flick
floated
flurry
flushed
fluttering

frisked
frowsty
fumbling
fusty
gaiety
gasped
giddy
glanced
glaring
gleam
glittering
gloating
gloriously
gluttony
goose
gorse
grant
grate
grip
groan
guise
hares
harm
harp
hastily
haunting
hawks
hawthorn

hinges
hoist
hunk
hurrahs
imperial
incantatio
inclined
indigo
indulgenc
interferers
invent
jaws
kingfisher
knight
knuckles
laburnums
larder
lashing
leering
legend
limbs
lithe
lulling
lurking
marvel
meadow
melancho
mended

occupant
onset
overwhelming
pact
pax
peacocks
peered
perched
perish
phew
pitter-patter
pleaded
plumage
popped
pounced
pounding
premises
prey
primroses
prigs
reckoned
remnants
renounced
repulsive
resumed
revelry
ripe
rippled
robe
romp
roused
rubble
rug
saccharine
salute
scarlet
scornfully
scouts
sensation
shafts
sip
sire
sizzling
slab
slacking
slave
sluice
smeared
sneered
sniffing
snigger
soldier
sorcerers
spades
spires
spirits
spitting
splendid
splutter
spoils
stratagem
struggled
stumbled
stunning
swaying
swiftest
swirling
swish
tackled
taken aback
tallow
tame
tapping
tassel
thee
thoroughly
thudding
thrush
tiptoe
token
twitter
valiant
vicious
vile
vultures
wading
wailed
warrior
whet
wireless
wreaths
wreck
wrenched
wringing
wrinkled
yawn
ye
yew

Contemporary English. (No E-J dictionary or other E-E dictionary)

2. Put a tab in the dictionary for each word you look up, and bring the dictionary to my room when you have your next test.

Guidelines:

1. You cannot look up every word – focus on words which you need to know to understand the story better.

2. For each word you look up, follow these four steps: a) if the word is an inflected form (e.g., *-er, -s, -ing*), remove the inflections to find the form to look up.
b) from the context, guess the possible type of thing the word means (e.g., *a type of animal/food,* etc.)
c) keeping a) and b) in mind, look up the word.
d) read the definition <u>and</u> examples to understand better.

Don't look up too many words – it will take too much time.

DON'T FORGET TO LEAVE A TAB EACH TIME:
Use one colour tab for words in the outside column – and another colour tab for words looked up in the inside column.

01322223:fond
12222223:fetch
00100011:din
00000000:curse
01112133:crockery
00000112:boggles
33333333:evidence
01112112:pedlars
00001101:twitching
01222333:device
00000100:saccharine
01110110:bruised
00000112:dunces
11333333:bluebottle
01000000:gaiety
00000000:stout
00000012:knight
00000101:pleaded
00000001:guise
00000001:thudding
33333333:ankles
01001211:slave
01000011:array
01101011:aisle
00122222:tame
12121112:spires
01233333:crocuses
00000010:lurking
02222212:trace

01000011:rellos
12232333:trial
01000001:jaws
01111110:inclined
02232333:filthy
00001122:moths
00000011:beaks
00100000:yawn
00122233:bellowing
00012223:perched
00000011:whet
00000111:hastily
11112222:disposal
33333333:extremely
12211123:fluttering
23223333:minister
01010122:renounced
01122012:mercy
01011122:shoved
00111222:leering
01111012:drained
01122212:bold
01112233:forth
00222121:bargaining
00000021:ye
12222122:thrush
02201011:blame
01002101:peacocks
00100000:tallow

01112221:occupant
00000021:nagging
02111223:craves
22333233:delay
00000121:reckoned
01121212:decent
11222233:sniffing
00212333:foe
11002223:struggled
00002222:wreck
01000111:vile
01001122:swirling
00022212:brood
01333333:bluebells
00111203:sorcerers
01000011:thoroughl
01000101:pounced
01010211:ghouls
01011112:briskly
12222333:squatted
00000110:phew
01010101:blast
01110133:glaring
01222233:swaying
01002012:taken ab
00001011:clung
00000011:assaultin
12232323:splutter
12132233:tossing

01211132:prey
12222233:marvel
00000002:wading
00101011:haystack
00011122:seaweed
00000001:centaur
01001011:muck
23323333:shuffled
01001112:rattle
00000012:ripe
23223333:paisley
01000010:fumbling
00001101:sill
00111111:burring
00000101:intruding
01112232:disobey
00000011:enrage
00000111:lithe
12121232:peered
02011122:spindle
00002112:barrel
12332222:indigo
00001121:appeased
01111111:mended
32333333:glanced
01000001:meadow
00000122:dratted
11222222:pact

22223233:concealed
10221122:haunting
02222333:scarlet
01001111:premises
01101010:frugal
01001011:gasped
00001011:cloven
01022222:lashing
01010011:eerie
00000000:thee
00000110:selvage
21221222:bait
33333333:harm
00011011:consorts
00000101:hunk
00000100:laburnums
01111211:dense
11011213:contumely
12332223:clutching
00000111:gluttony
00000010:stale
01012033:tongs
11222223:sprites
01012112:spurs
01113232:heady
02001112:fidgets
01000011:smeared
00211333:sluice

01010121:fondled
00100111:drooped
00000111:dispute
01112121:bearing
01101222:popped
00000001:badgers
00011120:hares
12222222:champing
12322233:soldier
00010031:decoy
01001111:chimed
02222123:loge
02022112:brat
02222222:shafts
12223333:harp
00001112:engraved
13333333:melancho
01112221:prigs
00000000:beckoned
33333333:splendid
00001210:wreaths
01001210:propped
12222222:wrenche
01112322:startled
00000000:heave
01102232:warrior
01112222:dim
00001111:limbs

11223233:rug
22322333:floated
00100010:extenuate
01112113:goose
00000123:sizzling
01000111:gloating
01111111:hoist
13232333:mortar
00001011:remnants
01111121:cramped
01002112:wrinkled
00012111:fortress
01000111:incantation
00000000:can't abide
00000001:bugle
00000012:feats
01001121:gleam
33333333:eternal
00000000:wailed
11223223:stead
00000000:groan
01112112:scouts
00000111:prodigious
01122213:revelry
01000001:slab
00000000:camphor
00000000:stunning
00000110:brambles
01010001:treason

01122122:cluster
00000001:blaze
11223233:clashing
00021211:disheveled
00000113:perish
00000001:flurry
00000000:yew
12322223:shovel
01011000:hinges
11110111:ain't
01110112:rippled
01211212:duke
01112222:twinkling
00000111:slacking
22333333:spirits
01011011:dominions
01111112:salute
01112122:shreds
00001111:cocking
33333333:apes
00000111:fiddling
00122212:grant
00000001:ransacking
23333333:legend
01111111:swiftest
01000010:trowels
01001102:bunks
12232233:grip
01000113:stratagem

00000001:pan
01011112:hawks
13333333:imperial
00000003:dodging
11111111:interferers
00000010:indulgenc
00001012:moan
00000000:token
01010001:hawthorn
00000011:siege
00001011:curtsey
00011001:quarry
01011112:kingfisher
01011221:flick
01110222:flask
01111112:tassel
01212223:puddles

```
event 6: 86 102 80 42
event 7: 59 113 72 66
event 8: 37 108 75 90
building matrix 2:3
building matrix 2:3
183   50   43   34
185   44   42   39
186   41   41   42
184   40   41   45
181   40   42   47
180   39   42   49
178   39   43   50
177   39   43   51
175   39   44   52
174   39   44   53
173   39   44   54
171   39   45   55
170   39   45   56
```

```
event 6: 125 39 106 31
event 7: 108 32 120 41
event 8: 83 62 113 43
event 9: 87 62 114 38
building matrix 2:3
167   45   62   27
164   50   64   23
164   52   64   21
164   53   64   20
163   54   64   20
163   54   64   20
163   54   64   20
163   54   64   20
163   54   64   20
163   54   64   20
building matrix 4:5
124   33   95   49
108   32   103   58
95   31   108   67
85   29   112   75
76   28   115   82
69   27   117   88
```

awesome
banner
barbed
barren
battalions
beggar
benefactors
betray
bewildered
bitchy
blazing
blessing
blinking
bloom
boasting
bound
brassieres
briefer
brittle
burdens
bursting
cannibals
capture
carousel
caused
cease
cellar
chandelier
chanting
checkers
chiffon
circuits
citadel
clandestine
claws
coal
commodities
compelled
complaint

craze
creation
crime
cruel
crushed
currents
darn
deceiving
decline
defender
deny
deserted
deserve
desperate
disgrace
distress
downtown
dragonfly
drifting
drooling
drowsy
eagle
echo
edged
embrace
endures
extent
extinguished
fades
faith
faltered
favour
fence
ferns
filthy
flame
fling
flooded
flow

gladly
glance
glistening
gloomy
glory
gonna
gospel
governing
grace
grant
greed
grips
guilty
gut
hesitate
holy
hostage
humble
humming
hunger
hush
huts
indulgent
injustice
inscribed
inspiration
interrupt
ivory
ivy
jails
jars
jaw
jive
judges
knees
lantern
larceny
lay
liberation

nervous
noble
obsession
orchid
outskirts
pad
pane
passion
patriarchs
perils
perishing
petitions
petrified
pews
pity
plantations
poison
possessed
pour
praise
preacher
precious
prescribed
presume
proclaim
prophet
punished
purring
quandary
racking
rapture
realise
rebel
recognize
refined
reflected
regret
reigns
sacrifice
saints
savage
saviour
scarlet
scars
scorched
scribbled
scuffle
seams
sensation
sentimental
serpent
serve
shed
shifting
shoulder
sill
skinny
sin
slithers
smoldering
smuggled
snug
sophomore
soul
source
spinning
squinting
standards
staring
steal
steeples
stolen
stony
stricken
string stubborn
subterranean
thief
thirst
thorns
thrilling
throat
throne
tights
timber
tomb
torch
tourniquet
treat
tremble
tricks
trust
tumble
twisted
ugly
unspeakably
upon
uprising
vain
valley
vapors
vengeful
victory
vines
virtue
voice
wanna
weak
weary
whisper
wipe
worth
wrap
wreath
yell

```
00000000:thespian                00000000:thespian
00000000:newt                    00000000:newt
01110112:automat                 00113213:automat
00000000:denouement              00000000:denouement
00000000:smarmy                  00000000:smarmy
00000000:dubbin                  00000000:dubbin
11211112:maestro                 00000000:maestro
00000002:requisite               00000000:requisite
00000000:tortilla                00000000:tortilla
00000000:chillum                 00000000:chillum
00000000:stash                   00000000:stash
00000000:sullen                  00000000:sullen
00000000:inebriate               00000000:inebriate
00000000:uppity                  00000001:uppity
00000000:bivouac                 00000000:bivouac
00000000:farcical                00000000:farcical
01002000:glib                    00000000:glib
00001000:yonks                   00000000:yonks
00000000:perplexed               00000000:perplexed
00001000:razzle                  00000000:razzle
00000001:precinct                00000000:precinct
00001001:gully                   00003100:gully
21222123:sensor                  12223333:sensor
00000000:condoned                00000000:condoned
00000000:fragment                00000000:fragment
00000000:credulous               00000000:credulous
```

```
event 3: 124 56 47 74
event 4: 116 65 52 68
event 5: 92 79 54 76
event 6: 90 79 58 74
event 7: 66 73 62 100
event 8: 60 66 67 108

building matrix 2:3
131   51   45   74
135   49   44   73
138   48   43   72
141   47   42   71
142   47   42   70
144   47   41   69
146   46   41   68
147   46   41   67
149   46   40   66
150   46   40   65
```

```
event 3: 101 24 27 149
event 4: 103 22 21 155
event 5: 79 23 17 182
event 6: 68 24 27 182
event 7: 64 29 22 186
event 8: 64 27 36 174

building matrix 2:3
92   24   29   156
85   23   30   163
78   22   31   170
72   21   31   177
67   20   31   183
```