CAMBRIDGE

Learning Vocabulary in Another Language

I. S. P. Nation

Second Edition

CAMBRIDGE APPLIED LINGUISTICS

Series Editors: Carol A. Chapelle and Susan Hunston

Learning Vocabulary in Another Language

THE CAMBRIDGE APPLIED LINGUISTICS SERIES

The authority on cutting-edge Applied Linguistics research

Series Editors 2007–present: Carol A. Chapelle and Susan Hunston
1988–2007: Michael H. Long and Jack C. Richards

For a complete list of titles please visit: www.cambridge.org/elt/cal

Recent titles in this series:

Narrative Research in Applied Linguistics
Edited by Gary Barkhuizen

Teacher Research in Language Teaching
A Critical Analysis
Simon Borg

Figurative Language, Genre and Register
Alice Deignan, Jeannette Littlemore and Elena Semino

Exploring ELF
Academic English Shaped by Non-native Speakers
Anna Mauranen

Genres across the Disciplines
Student Writing in Higher Education
Hilary Nesi and Sheena Gardner

Disciplinary Identities
Individuality and Community in Academic Discourse
Ken Hyland

Replication Research in Applied Linguistics
Edited by Graeme Porte

The Language of Business Meetings
Michael Handford

Reading in a Second Language
Moving from Theory to Practice
William Grabe

Modelling and Assessing Vocabulary Knowledge
Edited by Helmut Daller, James Milton and Jeanine Treffers-Daller

Practice in a Second Language
Perspectives from Applied Linguistics and Cognitive Psychology
Edited by Robert M. DeKeyser

Feedback in Second Language Writing
Edited by Ken Hyland and Fiona Hyland

Task-Based Language Education
From Theory to Practice
Edited by Kris van den Branden

Second Language Needs Analysis
Edited by Michael H. Long

Insights into Second Language Reading
A Cross-Linguistic Approach
Keiko Koda

Research Genres
Exploration and Applications
John M. Swales

Critical Pedagogies and Language Learning
Edited by Bonny Norton and Kelleen Toohey

Exploring the Dynamics of Second Language Writing
Edited by Barbara Kroll

Understanding Expertise in Teaching
Case Studies of Second Language Teachers
Amy B. M. Tsui

Criterion-Referenced Language Testing
James Dean Brown and Thom Hudson

Corpora in Applied Linguistics
Susan Hunston

Pragmatics in Language Teaching
Edited by Kenneth R. Rose and Gabriele Kasper

Cognition and Second Language Instruction
Edited by Peter Robinson

Research Perspectives on English for Academic Purposes
Edited by John Flowerdew and Matthew Peacock

Computer Applications in Second Language Acquisition
Foundations for Teaching, Testing and Research
Carol A. Chapelle

Learning Vocabulary in Another Language

Second Edition

I. S. P. Nation

Victoria University of Wellington

The first edition of this title was
published under the series editorship
of Michael H. Long and Jack C. Richards

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107623026

First published 2001
Second edition 2013

Printed in the United Kingdom by Short Run Press, Exeter

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Nation, I. S. P. Learning vocabulary in another language / I. S. P. Nation. –
Second Edition.
pages cm. – (Cambridge applied linguistics)
ISBN 978-1-107-04547-7 (hardback) – ISBN 978-1-107-62302-6 (pb)
1. Vocabulary – Study and teaching. 2. Language and languages – Study and teaching. I. Title.
P53.9.N29 2013
418.0071 – dc23 2013021172

ISBN 978-1-107-62302-6 Paperback
ISBN 978-1-107-04547-7 Hardback

Contents

Series editors' preface *page* xiii
Acknowledgements xiv

Introduction 1
Learning goals 1
The four strands 2
Main themes 4
The audience for this book 4
The first and the second editions 4
Changes in the second edition 5
References 8

1 *The goals of vocabulary learning* 9
Counting words 9
How much vocabulary do learners need to know? 11
How many words are there in the language? 12
How many words do native speakers know? 13
How much vocabulary do you need to use another language? 14
Frequency-based word lists 16
Specialised vocabulary 19
Frequency levels in a large corpus 20
High-frequency words 22
Mid-frequency words 25
Low-frequency words 28
Specialised vocabulary 30
Zipf's law 32
Testing vocabulary knowledge 35
Training learners in choosing which words to learn 37
References 41

2 *Knowing a word* 44
Learning burden 44
Do L1 and L2 words share the same lexical store? 45
The receptive/productive distinction 46
Aspects of knowing a word 58
Levelt's process model of language use 61
Spoken form 65
Written form 70
Word parts 72
Connecting form and meaning 73
Concept and referents 75
Associations 79
Grammatical functions 82
Collocations 83
Constraints on use 84
Item knowledge and system knowledge 85
References 86

3 *Teaching and explaining vocabulary* 92
What is involved in vocabulary teaching and what role should teaching play? 92
What are the features of good vocabulary-teaching techniques? 95
Which vocabulary learning activities are the best ones? 98
What learning conditions help vocabulary learning? 102
How can we apply technique analysis to improving learning? 114
How should we do direct vocabulary teaching? 117
How should teachers (or writers) explain words? 126
How should teachers deal with words in intensive reading? 128
What are the different kinds of vocabulary activities and procedures? 131
Spoken form 131
Written form 131
Word parts 133
Strengthening the form–meaning connection 133
Concept and reference 134
Associations 136
Grammar 138
Collocation 138
Constraints on use 138
Vocabulary teaching procedures 139
How can we develop learners' fluency? 141

Should vocabulary-learning activities be matched to an individual's learning style? 143
How can computer-assisted vocabulary learning help? 144
References 152

4 *Vocabulary and listening and speaking* 161
What vocabulary knowledge is needed for listening? 161
How can we provide vocabulary support for listening? 164
What helps vocabulary learning from listening to stories? 167
How much vocabulary learning occurs through listening to lectures? 172
How much vocabulary learning occurs through interactive tasks? 172
Is there a special vocabulary of speaking? 180
How can learners develop fluency with spoken vocabulary? 182
How can a teacher use input to increase vocabulary knowledge? 185
How can a teacher use labeled diagrams to increase vocabulary knowledge? 187
How can a teacher use cooperative tasks to focus on vocabulary? 189
How can a teacher design activities to help incidental vocabulary learning? 190
Designing the worksheets 194
An adapted activity 197
References 199

5 *Vocabulary and reading and writing* 204
Reading and the four strands 204
How much vocabulary is needed for reading? 205
How much reading is needed for vocabulary development? 211
Is input by itself all that is needed for vocabulary development? 216
How can learners increase their reading vocabulary? 218
Is reading unsimplified text a good way to help vocabulary growth? 223
Will extensive reading of texts written for young native speakers make reading easier for non-native speakers? 226
How can learners be supported when reading unsimplified text? 230

Glossing 238
Is reading simplified texts a good way to help vocabulary growth? 247
What are the features of a good graded reading scheme? 250
How can we design a good graded reading programme? 254
Is extensive reading a good way to improve language proficiency? 258
Vocabulary and writing 262
How much does vocabulary use affect the quality of writing? 263
How can we measure the quality of vocabulary use in writing? 263
How can we improve vocabulary use in writing? 268
References 276

6 *Specialised uses of vocabulary* 289
What is academic vocabulary? 289
Why is academic vocabulary important? 291
How can you make an academic vocabulary list? 296
How can you sequence the introduction of academic vocabulary? 298
What is the nature and role of academic vocabulary? 300
How can you test academic vocabulary? 301
How can you learn academic vocabulary? 301
What is technical vocabulary? 303
How can you distinguish technical vocabulary from other vocabulary? 303
How can you learn technical vocabulary? 305
What roles can vocabulary play in discourse? 307
How is vocabulary related to the information content of a text? 308
How is vocabulary related to the organisation of a text? 311
How does vocabulary signal the relationship between the writer or speaker and reader or listener? 314
Words in discourse 315
How well does content-based learning support the learning of vocabulary? 318
References 321

7 *Vocabulary-learning strategies* 326
What vocabulary learning strategies are there? 326
Planning vocabulary learning 329
Sources: Finding information about words 330

Processes: Establishing vocabulary knowledge 331
Skill in use: Enriching knowledge 332
How can we train learners in strategy choice and use? 332
How well do learners use strategies? 334
What are the effects of training learners in strategy use? 342
Procedures that integrate strategies 343
References 345

8 *Learning words from context* 348
Which is best: intentional or incidental learning? 348
How are reading and vocabulary growth related to each other? 350
What proportion of unknown words can be guessed from context? 352
How much vocabulary is learned from context? 355
What can be learned from context? 358
What clues does a context provide and how effective are they? 360
What are the causes of poor guessing? 364
Do different learners approach guessing in the same way? 366
How can learners be trained to guess from context? 369
Does drawing attention to words help learning from context? 370
Do glossing and dictionary use help vocabulary learning? 371
What formats should be used for testing or practising guessing? 372
What are the steps in the guessing-from-context strategy? 375
How should we plan the training of learners in the strategy of guessing from context? 380
References 381

9 *Word parts* 389
Is it worthwhile learning word parts? 390
Where do English words come from? 390
How many words fit into a word family? 391
What are the most useful prefixes and suffixes? 394
Do language users see words as being made of parts? 396
What are the most useful word stems? 400
What knowledge is required for learners to use word parts? 402

What word-building skills should the teacher monitor and test? 404
What is the word part strategy for remembering new words? 408
References 411

10 *Using dictionaries* 414
Do learners use dictionaries well? 415
Do dictionaries help learners understand and produce text? 416
Does dictionary look-up help with vocabulary learning? 417
What skills are needed to use a dictionary? 419
What dictionaries are the best? 423
How can we evaluate dictionaries? 427
Dictionary use and learning 432
References 433

11 *Deliberate learning from word cards* 437
Is learning from word cards a useful activity? 438
What is the most effective way of learning from word cards? 445
Is flashcard software useful for vocabulary learning? 468
How can we train learners in the use of word cards? 469
References 471

12 *Finding and learning multiword units* 479
What are multiword units? 479
Why are multiword units important? 481
How can we identify and count multiword units? 485
Where can we find lists of multiword units? 495
What is the range of ways to teach and learn multiword units? 497
References 508

13 *Testing vocabulary knowledge and use* 514
Purposes of tests 515
How can we test to see where learners need help? 515
How can we measure if learners have control of the important vocabulary learning strategies? 517
How can we test vocabulary to place learners in classes at the right level? 518
How can we test whether a small group of words in a course has been learned? 521

How can we test whether the total vocabulary of the course has been learned? 521
How can we measure native speakers' and non-native speakers' total vocabulary size? 522
What kind of vocabulary test item is the best? 535
Is it enough to ask learners if they know the word? 539
Should choices be given? 540
Should translations be used? 544
Should words be tested in context? 545
Should vocabulary be tested in a communicative context? 548
How can depth of knowledge about a word be tested? 549
How can we measure words that learners don't know well? 557
How can we measure how well learners actually use words? 561
References 562

14 *Designing the vocabulary component of a language course* 569
Goals 570
Needs analysis 570
Environment analysis 572
Principles of vocabulary teaching 572
Content and sequencing 573
Format and presentation 577
Monitoring and assessment 578
Evaluation 580
Autonomy and vocabulary learning 583
The goals of vocabulary learning 584
What should be learned and in what order? 584
Learning procedures 588
Checking learning 593
References 596

Appendices 598
1. Headwords of the Academic Word List 598
2. Vocabulary levels dictation test 603
3. Function words 604
References 605

Subject index 606

Author index 613

Series editors' preface

Over ten years ago, when the first edition of *Learning Vocabulary in Another Language* was published, vocabulary learning was characterised by the then series editors as an area studied by only a few pioneers, Paul Nation being one of them. In part due to the tremendous impact of the first edition of Nation's book, today research and teaching of second language vocabulary learning is no longer the preoccupation of just a few. On the contrary, throughout applied linguistics, vocabulary, formulaic expressions, word patterns and lexical bundles are centre stage in the study of how learners develop the ability to make meaning. With the importance of the lexical dimension of language development recognised, the research basis for understanding vocabulary teaching and learning has grown to be substantial. A second edition of Paul Nation's seminal work was needed.

The second edition of *Learning Vocabulary in Another Language* possesses the same qualities that made the first edition so popular. It is organised around issues relevant to readers needing a solid understanding of vocabulary in order to improve practices in second language vocabulary teaching and assessment. For example, chapters outline the goals of vocabulary learning, teaching and explaining vocabulary, vocabulary and listening and speaking, as well as vocabulary and reading. The book presents and interprets a comprehensive pool of research on second language vocabulary acquisition, and in so doing it provides research-based recommendations for practice. Relevant research appears across the domains of linguistics, second language acquisition, assessment and technology; Nation has culled the pertinent findings to address important questions such as whether or not learners actually acquire new word meanings from context and how learners use dictionaries. The style of writing is direct and engaging for readers at a range of levels. The book begins with the basics (that is, knowing a word), and it builds to the real world challenges educators face, such as assessing vocabulary knowledge and use, and developing the vocabulary component of a language course.

We are very happy to welcome this new edition of *Learning Vocabulary in Another Language* to the Cambridge Applied Linguistics Series.

Carol A. Chapelle and Susan Hunston

Acknowledgements

The authors and publishers acknowledge the following sources of copyright material and are grateful for the permissions granted. While every effort has been made, it has not always been possible to identify the sources of all the material used, or to trace all copyright holders. If any omissions are brought to our notice, we will be happy to include the appropriate acknowledgements on reprinting.

Parts of Chapter 4 appeared in Joe, A., Nation, P. and Newton, J. (1996) Speaking activities and vocabulary learning. *English Teaching Forum* **34**, 1: 2–7; Parts of Chapter 5 appeared in Nation, I. S. P. (1997) The language learning benefits of extensive reading. *The Language Teacher* **21**, 5: 13–16; Parts of Chapter 8 appeared in Nation, I. S. P. (1982) Beginning to learn foreign vocabulary: A review of the research. *RELC Journal* **13**, 1: 14–36; Table on p. 488 taken from Nation/Webb (2011) Researching and Analyzing Vocabulary, 1E. Copyright © 2011 Heinle/ELT, a part of Cengage Learning, Inc., www.cengage.com/permissions. Reproduced by permission; Parts of Chapter 14 appeared in Nation, I. S. P. (1998) Helping learners take control of their vocabulary learning. *GRETA* **6**, 1: 9–18.

Introduction

This book is about the teaching and learning of vocabulary, but the teaching and learning of vocabulary is only a part of a language development programme. It is thus important that vocabulary teaching and learning is placed in its proper perspective.

Learning goals

Vocabulary learning is only one sub-goal of a range of goals that are important in the language classroom. The mnemonic LIST is a useful way of remembering these goals that are outlined in Table 0.1. L = Language, which includes vocabulary; I = Ideas, which cover content and subject matter knowledge as well as cultural knowledge; S = Skills; and T = Text or discourse, which covers the way sentences fit together to form larger units of language.

Although this book focuses on the vocabulary sub-goal of language, the other goals are not ignored. However, they are approached from the

Table 0.1 *Goals for language learning*

General goals	Specific goals
Language items	pronunciation vocabulary grammatical constructions
Ideas (content)	subject matter knowledge cultural knowledge
Skills	accuracy fluency strategies process skills or subskills
Text (discourse)	conversational discourse rules text schemata or topic type scales

viewpoint of vocabulary. There are chapters on vocabulary and the skills of listening, speaking, reading and writing. Discourse is looked at in Chapter 6 on specialised uses, and pronunciation, spelling and grammar are looked at in relation to vocabulary knowledge in Chapter 3.

The four strands

The approach taken in this book rests on the idea that a well-balanced language course should consist of four major strands (Nation, 2007; Nation and Yamamoto, 2011). These strands can appear in many different forms, but they should all be there in a well-designed course.

Firstly, there is the strand of learning from comprehensible meaning-focused input. This means that learners should have the opportunity to learn new language items through listening and reading activities where the main focus of attention is on the information in what they are listening to or reading. As we shall see in the following chapter, learning from meaning-focused input can best occur if learners are familiar with at least 98 per cent of the running words in the input they are focusing on. Put negatively, learning from meaning-focused input cannot occur if there are lots of unknown words.

The second strand of a course is the strand of meaning-focused output. Learners should have the chance to develop their knowledge of the language through speaking and writing activities where their main attention is focused on the information they are trying to convey. Speaking and writing are useful means of vocabulary development because they make the learners focus on words in ways they did not have to while listening and reading. Having to speak and write encourages learners to listen like a speaker and read like a writer. This different kind of attention is not the only contribution that speaking and writing activities can make to language development. From a vocabulary perspective, these productive activities can strengthen knowledge of previously met vocabulary.

The third strand of a course is one that has been subject to a lot of debate. This is the strand of language-focused learning, sometimes called form-focused instruction. There is growing evidence (Ellis, 2005; Williams, 2005) that language learning benefits if there is an appropriate amount of usefully focused deliberate teaching and learning of language items. From a vocabulary perspective, this means that a course should involve the direct teaching of vocabulary and the direct learning and study of vocabulary. As we shall see, there is a very large amount of research stretching back to the late 19th century which shows that the gradual cumulative process of learning a word can be given a strong boost by the direct study of certain features of the word.

The fourth strand of a course is the fluency development strand. In the activities which put this strand into action learners do not work with new language items. Instead, they become more and more fluent in using items they already know. A striking example of this can be found in the use of numbers. Learners can usually quickly learn numbers in a foreign language. But if they go into a post office and the clerk tells them how much the stamps they need are going to cost, they might not understand because the numbers were said too quickly for them. By doing a small amount of regular fluency practice with numbers (the teacher says the numbers, the learners write the figures), the learners will find that they can understand one-digit numbers said quickly (1, 7, 6, 9) although they have trouble with two-digit numbers said quickly (26, 89, 63, 42) or three-digit numbers (126, 749, 537, 628). A little further practice will make these longer numbers fluently available for comprehension. If a course does not have a strong fluency strand, then the learning done in the other three strands will not be readily available for normal use.

In a language course, these four strands should get roughly the same amount of time. That means that no more than 25 per cent of the learning time in and out of class should be given to the direct study of language items. No less than 25 per cent of the class time should be given to fluency development. If the four strands of a course are not equally represented in a particular course, then the design of the course needs to be looked at again.

These four strands need to be kept in mind while reading this book. Where recommendations are made for direct vocabulary learning, these should be seen as fitting into that 25 per cent of the course which is devoted to language-focused learning. Seventy-five per cent of the vocabulary development programme should involve the three meaning-focused strands of learning from input, learning from output and fluency development.

The four strands apply generally to a language course. In this book we will look at how vocabulary fits into each of these strands. It is worth stressing that the strands of meaning-focused input and output are only effective if the learners have sufficient vocabulary to make these strands truly meaning focused. If activities which are supposed to be meaning focused involve large amounts of unknown vocabulary, then they become language focused because much of the learners' attention is taken from the message to the unknown vocabulary. Similarly, fluency development activities need to involve little or no unknown vocabulary or other language items, otherwise they become part of the meaning input and output strands, or language-focused learning.

Main themes

A small number of major themes run through this book, and these are first dealt with in Chapters 2, 3 and 4. Firstly, there is the cost/benefit idea based on the results of word frequency studies. Its most important application is in the distinction between high-frequency and mid- and low-frequency vocabulary and the different ways in which teachers should deal with these types of vocabulary. The cost/benefit idea also applies to individual words in that the amount of attention given to an item should be roughly proportional to the chances of it being met or used again, that is, its frequency.

Secondly, there is the idea that learning a word is a cumulative process involving a range of aspects of knowledge. Learners thus need many different kinds of meetings with words in order to learn them fully. There is to date still little research on how vocabulary knowledge grows and how different kinds of encounters with words contribute to vocabulary knowledge. In this book, knowing a word is taken to include not only knowing the formal aspects of the word and knowing its meaning, but also being able to use the word.

Thirdly, there is the idea that teachers and learners should give careful consideration to how vocabulary is learned, in particular, the psychological conditions that are most likely to lead to effective learning. Because these conditions are influenced by the design of learning tasks, quite a lot of attention is given to the analysis and design of vocabulary-learning activities.

The audience for this book

This book is intended to be used by second and foreign language teachers. Although it is largely written from the viewpoint of a teacher of English, it could also be used by teachers of other languages.

This book is called *Learning Vocabulary in Another Language* partly in order to indicate that most of the suggestions apply to both second and foreign language learning. Generally the term **second language** will be used to apply to both second and foreign language learning. In the few places where a contrast is intended, this will be clear from the context.

The first and the second editions

"I've got the first edition. Is it worth buying the second edition?" – this is a question I expect to be asked, so here is my answer.

Yes. Most of the changes in the second edition are the result of a large amount of research which has appeared since the first edition was published in 2001. By my rough calculation, over 30 per cent of the research on vocabulary that has appeared in the last 110 years was published in the last eleven years. Teaching and learning vocabulary, particularly for foreign and second language learners, is no longer a neglected aspect of language learning. So, if you don't buy the second edition you will be out of date by eleven years and at least 30 per cent of the field. On a rough estimate, at least one-fifth of the book is new material.

There were also errors in the first edition, largely because of a lack of research on the relevant areas. Some of that research has now been done, much of it by my students, colleagues and friends, and a few people who fit two or all of those categories.

I am also pleased to note that my thinking has changed on some issues in the teaching and learning of vocabulary, largely as a result of research findings and my own experience and thinking. These include the idea of mid-frequency vocabulary, largely as a result of research on word lists and testing native speaker vocabulary size. I also now feel that I am beginning to understand what collocations are. I am also becoming more sceptical of the value of vocabulary teaching, largely because of its necessarily limited scope and limited effectiveness.

When working on this second edition, I often wondered if the field of teaching and learning vocabulary is now so vigorous and large that it is beyond the scope of any one book and certainly one person. If you have already bought this book, then I hope I am wrong and you have got your money's worth.

Changes in the second edition

One of the changes in Chapter 1 is because of Chung and Nation's (2003, 2004) research on technical vocabulary. In the first edition I got this completely wrong, saying that about five per cent of the running words in a technical text would be technical vocabulary. In fact research showed that it was closer to 20 to 30 per cent of the running words. The second major change in Chapter 1 is as a result of the development of the lower-frequency word family lists based on the British National Corpus. At the time of writing, these lists now go up to the 24th one-thousand word lists, and the development of these lists has meant that we can do much more detailed analysis of texts and their vocabulary demands, as well as develop more soundly based vocabulary size tests. This research has highlighted the idea of mid-frequency words (Schmitt and Schmitt, 2012). At the time of writing the first edition, Coxhead's (2000) work on the Academic Word List was just being completed. The research just made it into the first edition, but in this second edition it is given the additional attention it deserves.

Chapter 2, 'Knowing a word', includes recent research on the relationship between first language (L1) and second language (L2) vocabulary storage. Chapter 3 includes a description of Technique Feature Analysis first introduced in Nation and Webb (2011a). Chapter 4 on listening and speaking includes recent work on vocabulary learning through lectures and learning in interactive activities. Chapter 5 on reading and writing is largely reorganised, and there is much more on glossing because of the growth in research on electronic glossing. It also includes recent corpus and experimental work on text coverage as well as recent studies of learning from graded readers and reading fluency. Chapter 6, 'Specialised uses', now has critiques of the academic word list, recent work on technical vocabulary and a section on content-based vocabulary teaching. Chapter 7, 'Vocabulary-learning strategies', includes recent research on strategy training and strategy use. The research on strategy use is now becoming more rigorous with less dependence on questionnaires. Chapter 8 contains recent research on guessing. Chapter 9, 'Word parts', has only very few changes. The changes in Chapter 10 are largely due to the growth in electronic dictionaries. Chapter 11, 'Deliberate learning from word cards', includes recent research on whether expanded spacing is better than even spacing within a learning session. It also includes criteria for evaluating flashcard programmes (Nakata, 2011). It also includes what I consider to be the most significant recent research finding in the field of vocabulary learning, namely that rote learning results in both implicit and explicit knowledge (Elgort, 2011) and thus the learning/acquisition distinction is not relevant for vocabulary. Chapter 12 on finding and learning collocations is almost completely rewritten. For me this was the most unsatisfactory chapter in the first edition. I now feel I am beginning to see how the work on collocations fits together, largely by separating the types of criteria used to classify collocations into criteria of form, meaning, function and storage. I have kept a few small sections but have taken a new approach to the chapter. There has been a large amount of research on collocations and some of it is very innovative. However there is still a need for clear definitions of what kind of units are being investigated and following these definitions closely when doing the research. Chapter 13 on testing changes the table of test sensitivity to agree with Laufer and Goldstein's (2004) findings. There is now more on the Word Associates Test, and there is also recent research on vocabulary size, including (Biemiller, 2005) findings with

L1 learners and research on the *Vocabulary Size Test*. Chapter 14 on planning has very few changes.

There is now an international community of vocabulary researchers and I am grateful to them for the knowledge, support and encouragement they have given me in the preparation of this book and in my research and writing.

Since I wrote my first book, *Teaching and Learning Vocabulary* (Nation, 1990) and the first edition of this book, another generation of vocabulary researchers has appeared. Although this is still a relatively small group, it is made up of very productive researchers who have identified a range of useful research focuses and who persist in exploring and refining research in those chosen areas. It is also notable that recently two books focusing on the research methodology of vocabulary studies have appeared (Nation and Webb, 2011; Schmitt, 2010). Research on vocabulary is clearly alive and well.

I am very grateful to Norbert Schmitt, Pavel Szudarski, Suhad Sonbul and Laura Vilkaite for comments on a draft of this book. Their insightful comments led to significant improvements in the book.

References

Biemiller, A. (2005). Size and sequence in vocabulary development, in Hiebert, E. H. and Kamil, M. L. (eds.), *Teaching and Learning Vocabulary: Bringing Research into Practice*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 223–42.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, **15**, 2, 103–16.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, **32**, 2, 251–63.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, **34**, 2, 213–38.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language, *Language Learning*, **61**, 2, 367–413.

Ellis, R. (2005). Principles of instructed language learning, *System*, **33**, 209–24.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, **54**, 3, 399–436.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software, *Computer Assisted Language Learning*, **24**, 1, 17–38.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Boston: Heinle.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, **1**, 1, 1–12.

Nation, I. S. P. and Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.

Nation, I. S. P. and Yamamoto, A. (2011). Applying the four strands to language learning, *International Journal of Innovation in English Language Teaching and Research*, **1**, 2, 1–15.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N. and Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, doi:10.1017/S0261444812000018.

Williams, J. (2005). Form-focused instruction, in Hinkel, E. (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 671–91.

1 *The goals of vocabulary learning*

The idea behind this chapter is that it is helpful to use frequency and range of occurrence to distinguish several levels of vocabulary. Distinguishing these levels helps ensure that learners learn vocabulary in the most useful sequence and thus gain the most benefit from the vocabulary they learn. Making the high-frequency/mid-frequency/low-frequency distinction ensures that the teacher deals with vocabulary in the most efficient ways.

Counting words

There are several ways of counting words, that is, deciding what will be counted.

Tokens

One way is simply to count every word form in a spoken or written text and if the same word form occurs more than once, then each occurrence is counted. So, the sentence, *It is not easy to say it correctly*, would contain eight words, even though two of them are the same word form, *it*. Words which are counted in this way are called **tokens**, and sometimes **running words**. If we try to answer questions like 'How many words are there on a page or in a line?', 'How long is this book?', 'How fast can you read?' or 'How many words does the average person speak per minute?', then our unit of counting will be the token.

Types

We can count the words in the sentence *It is not easy to say it correctly* another way. When we see the same word occur again, we do not count it again. So the sentence of eight tokens consists of seven different words or **types**. We count words in this way if we want to answer questions like 'How large was Shakespeare's vocabulary?', 'How many words do you need to know to read this book?' or 'How many words does this dictionary contain?'

Lemmas

Counting *book* and *books* as two different words to be learned seems a bit strange. So, instead of counting different types as different words, closely related words could be counted as members of the same word or **lemma**. A lemma consists of a headword and its inflected forms and reduced forms (*n't*). Usually, all the items included under a lemma are all the same part of speech (Francis and Kučera, 1982). The English inflections consist of plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative, possessive (Bauer and Nation, 1993). The Thorndike and Lorge (1944) frequency count used lemmas as the basis for counting, and the computerised count on the Brown corpus produced a lemmatised list (Francis and Kučera, 1982). In the Brown count the comparative and superlative forms were not included in the lemma, and the same form used as a different part of speech (*walk* as a noun, *walk* as a verb) are not in the same lemma. Variant spellings (*favor*, *favour*) are usually included as part of the same lemma when they are the same part of speech. Leech et al. (2001) used similar criteria in their count of the British National Corpus (http://ucrel.lancs.ac.uk/bncfreq).

Lying behind the use of lemmas as the unit of counting is the idea of **learning burden** (Swenson and West, 1934). The learning burden of an item is the amount of effort required to learn it. Once learners can use the inflectional system, the learning burden of *mends*, if the learner already knows *mend*, is negligible. One problem to be faced in forming lemmas is to decide what will be done with irregular forms such as *mice*, *is*, *brought*, *beaten* and *best*. The learning burden of these is clearly heavier than the learning burden of regular forms like *books*, *runs*, *talked*, *washed* and *fastest*. Should the irregular forms be counted as a part of the same lemma as their base word or should they be put into separate lemmas? Lemmas also separate closely related items, such as the adjective and noun uses of words like *original*, and the noun and verb uses of words like *display*. An additional problem with lemmas is to decide what is the headword of the lemma – the base form or the most frequent form? (Sinclair, 1991: 41–2).

Using the lemma as the unit of counting greatly reduces the number of units in a corpus. Bauer and Nation (1993) calculated that the 61,805 tagged types (or 45,957 untagged types) in the Brown corpus become 37,617 lemmas, which is a reduction of almost 40% (or 18% for untagged types). Nagy and Anderson (1984) estimated that 19,105 of the 86,741 types in the Carroll et al. (1971) corpus were regular inflections.

Word families

Lemmas are a step in the right direction when trying to represent learning burden in the counting of words. However, there are clearly other affixes which are used systematically and which greatly reduce the learning burden of derived words containing known base forms, for example *-ly*, *-ness* and *un-*. A **word family** consists of a headword, its inflected forms and its closely related derived forms.

The major problem in counting using word families as the unit is to decide what should be included in a word family and what should not. Learners' knowledge of the prefixes and suffixes develops as they gain more experience of the language. What might be a sensible word family for one learner may be beyond another learner's present level of proficiency. This means that it is usually necessary to set up a scale of word families, starting with the most elementary and transparent members and moving on to less obvious possibilities (Bauer and Nation, 1993). Ward and Chuenjundaeng (2009) warn that we need to be cautious in assuming that learners know the family members of word families. Their study of low-proficiency Thai university students showed that the students' ability to see the relationship between stems and derived forms was very limited. Neubacher and Clahsen (2009) found that less proficient non-native speakers of German were more influenced by the morphological structure involving regular affixes than high proficiency non-native speakers. Non-native speakers seemed more likely to store words as unanalysed wholes.

Which unit we use when counting will depend on our reason for counting. Whatever unit we use, we need to make sure it is the most suitable one for our purpose. We need to make this decision when working out how much vocabulary our learners need to know.

How much vocabulary do learners need to know?

Whether we are designing a language course or planning our own course of study, it is useful to be able to set learning goals that will allow us to use the language in the ways we want to. When we plan the vocabulary goals of a long-term course of study, we can look at three kinds of information to help decide how much vocabulary needs to be learned: the number of words in the language, the number of words known by native speakers, and the number of words needed to use the language.

How many words are there in the language?

The most ambitious goal is to know all of the language. This is very ambitious because native speakers of the language do not know all the vocabulary of the language. There are numerous specialist vocabularies, such as the vocabulary of nuclear physics or computational linguistics, which are known only by the small groups of people who specialise in these areas. Still, it is interesting to have some idea of how many words there are in a language. This is not an easy question to answer because there are numerous other questions which affect the way we answer it. They involve considerations like the following.

What do we count as a word? Do we count *book* and *books* as the same word? Do we count *green* (the colour) and *green* (a large grassed area) as the same word? Do we count people's names? Do we count the names of products like *Fab*, *Pepsi*, *Vegemite*, *Chevrolet*? One way to answer these questions and the major question 'How many words are there in English?' is to count the number of words in very large dictionaries. *Webster's Third New International Dictionary* is one of the largest non-historical dictionaries of English. It contains around 54,000 base word families excluding proper names (Goulden et al., 1990: 322–3). This is a very large number and is well beyond the goals of most first and second language learners. Another way is to look at very large collections of texts and see how many words occur in those texts. Nagy and Anderson (1984) projected from their analysis of part of the data from Carroll et al.'s (1971) *Word Frequency Book* that, excluding proper names, foreign words, formulae, numbers and non-words, there were between 54,000 and 88,500 different word families in printed school English, depending on what is included in a word family. The *Word Frequency Book* is based on a corpus of 5 million running words. An analysis of the British National Corpus using the Range program comes up with similar figures.

There are 272,782 word types in the British National Corpus that are not in the first 20,000 word family lists and the accompanying proper name, marginal words, transparent compounds and abbreviations lists. Almost half of the 272,782 different word families are proper nouns. Four per cent are foreign words and six per cent are low-frequency members of word families already in the 20 one-thousand-word lists. Ideally, these family members should be added to the families in the existing lists.

The new words not yet in the lists plus the 20,000 in the word lists total around 70,000 word families which is a figure within Nagy and

Anderson's (1984) estimates, and the number of words in most reasonably sized non-historical dictionaries.

A major reason for trying to see how many words there are in English is to set the boundaries for measures of learners' vocabulary size. Early studies of vocabulary size using faulty methodology (Diller, 1978; Seashore and Eckerson, 1940) reached estimates that were well beyond the number of words in the language.

How many words do native speakers know?

Instead of considering how many words there are in the language, a less ambitious way of setting vocabulary learning goals is to look at what native speakers of the language know. Unfortunately, research on measuring vocabulary size has generally been poorly done (Nation, 1993), and the results of the studies stretching back to the late nineteenth century are often wildly incorrect. We will look at the reasons for this in Chapter 13.

More reliable studies (Goulden et al., 1990; Zechmeister et al., 1995) suggest that educated adult native speakers of English know under 20,000 word families. These estimates are rather low because the counting unit is word families which have several derived family members, and proper nouns are not included in the count. A very rough rule of thumb would be that for each year of their early life, starting at the age of three and probably up to 25 years old or so, native speakers add on average 1,000 word families a year to their vocabulary (Biemiller and Slonim, 2001). Learning 1,000 word families a year is an ambitious goal for non-native speakers of English, especially those learning English as a foreign rather than second language. In one important respect however the learning burden of English words for learners of English as a foreign language is becoming easier. This is because a large number of English words exist as loanwords in the learner's first language. For example, Daulton (2008) estimates that about half of the first 3,000 words of English exist in Japanese in some form or other, and Japanese learners know the meanings of these loanwords. The existence of these loanwords makes the learning of their English forms easier.

We need to be careful when seeing native speakers' language proficiency as a goal for L2 learners. Mulder and Hulstijn (2011) looked at the Dutch language proficiency of native speakers of Dutch. They tested native speakers across a wide range of ages (18–76 years old) and with a wide range of educational backgrounds and in a wide range of professions. Lexical fluency and lexical memory span declined with age while lexical knowledge increased. High education and a high profession level positively affected lexical knowledge and lexical memory span. There was a large variability in native speakers' language knowledge and skills. This variability has also been noted in studies of the vocabulary size of young native speakers of English (Biemiller and Slonim, 2001). This variability raises the question of what type of native speakers we should use when comparing them with non-native speakers.

How much vocabulary do you need to use another language?

Studies of native speakers' vocabulary suggest that second language learners need to know very large numbers of words. While this may be useful in the long term, it is not an essential short-term goal. This is because studies of native speakers' vocabulary growth see all words as being of equal value to the learner. Frequency-based studies show very strikingly that this is not so, and that some words are much more useful than others (see Schmitt, 2008, for a very useful discussion of vocabulary-learning goals). Thus, another way of setting vocabulary-learning goals is to work out how many really useful words learners need to know.

Table 1.1 shows part of the results of a frequency count of just under 500 running words in the Ladybird version of the children's story, *The Three Little Pigs*. It contains 124 different word types.

The most frequent word is *the* which occurs 41 times in the book. Note the large proportion of words occurring only once and the very high frequency of the few most frequent words. Note also the quick drop in frequency of the items.

When we look at texts our learners may have to read and conversations that are like ones they may be involved in, we find that a relatively small amount of well-chosen words can allow learners to do a lot. An analysis of various kinds of texts using 1,000-word family lists made from the British National Corpus (Nation, 2006) shows that between 3,000 to 4,000 word families are needed to get 95% text coverage, and between 6,000 and 9,000 word families are needed to gain 98% coverage (see Table 1.2). A coverage of 98% is chosen as the goal because a small amount of research supports this figure (Hu and Nation, 2000; Schmitt et al., 2011; van Zeeland and Schmitt, 2012), and because this represents a manageable amount of unknown vocabulary. If 2% of the running words are not known, this equates to one word in 50, or one word in every five lines (assuming 10 words

Table 1.1 *An example of the results of a frequency count*

the	41	*than*	4	*by*	2	*him*	1
little	25	*very*	4	*care*	2	*houses*	1
pig	22	*asked*	3	*chin*	2	*huff*	1
house	17	*carrying*	3	*day*	2	*knocked*	1
a	16	*eat*	3	*does*	2	*live*	1
and	16	*gave*	3	*huffed*	2	*long*	1
said	14	*give*	3	*let*	2	*mother*	1
he	12	*his*	3	*m*	2	*must*	1
i	10	*in*	3	*no*	2	*my*	1
me	10	*it*	3	*puffed*	2	*next*	1
some	9	*ll*	3	*strong*	2	*off*	1
wolf	9	*met*	3	*take*	2	*once*	1
build	8	*myself*	3	*then*	2	*one*	1
t	8	*not*	3	*time*	2	*puff*	1
third	8	*on*	3	*too*	2	*road*	1
was	8	*pigs*	3	*along*	1	*set*	1
of	7	*please*	3	*are*	1	*so*	1
straw	7	*pleased*	3	*ate*	1	*their*	1
to	7	*shall*	3	*blow*	1	*them*	1
you	7	*soon*	3	*but*	1	*there*	1
man	6	*stronger*	3	*came*	1	*took*	1
second	6	*that*	3	*chinny*	1	*up*	1
catch	5	*they*	3	*come*	1	*upon*	1
first	5	*three*	3	*door*	1	*us*	1
for	5	*want*	3	*down*	1	*walked*	1
will	5	*who*	3	*fell*	1	*we*	1
bricks	4	*with*	3	*go*	1	*went*	1
built	4	*won*	3	*grew*	1	*were*	1
himself	4	*yes*	3	*had*	1	*which*	1
now	4	*yours*	3	*hair*	1	*your*	1
sticks	4	*big*	2	*here*	1	*yourselves*	1

per line), or six unknown words per 300-running word page, or around 1,200 unknown words in a 200-page book. There is research that shows that 95% coverage may be sufficient for spoken narrative texts (van Zeeland and Schmitt, 2012).

The figures in Table 1.2 assume that vocabulary is learned in the order of its frequency. That is, that the first 1,000 words are learned before the second 1,000 words, and the second 1,000 words are learned before the third 1,000 words, and so on. This is a reasonable assumption for the high- and mid-frequency levels of the language.

Table 1.2 *English vocabulary sizes needed to get 95% and 98% coverage (including proper nouns) of various kinds of texts (Nation, 2006)*

Texts	95% coverage	98% coverage	Proper nouns
Novels	4,000 word families	9,000 word families	1–2%
Newspapers	4,000 word families	8,000 word families	5–6%
Children's movies	4,000 word families	6,000 word families	1.5%
Spoken English	3,000 word families	7,000 word families	1.3%

Webb and Macalister (forthcoming) show that texts written for young native speakers (the very popular New Zealand School Journals) have the same vocabulary size demands as texts written for native-speaking adults. Predictably, graded readers provide a much more favourable vocabulary load.

Frequency-based word lists

We can usefully distinguish three kinds of vocabulary based on frequency levels. Let us look at a written academic text and examine the different frequency levels of vocabulary it contains. The text is from Neville Peat's (1987) *Forever the Forest. A West Coast Story* (Hodder and Stoughton, Auckland).

The vocabulary is divided into three groups according to frequency lists of word families. The high-frequency words (the most frequent 2,000 word families) are unmarked in the text, the mid-frequency words (7,000 word families from the 3rd to the 9th 1,000-word lists inclusive) are in *italics*, and the low-frequency words (10th 1,000-word list onward) are in **bold**.

> Sustained-*yield* management ought to be long-term government policy in *indigenous* forests *zoned* for production. The adoption of such a policy would represent a break through the boundary between a *pioneering*, extractive phase and an *era* in which the

timber industry adjusted to living with the forests in **perpetuity**. A forest sustained is a forest in which harvesting and *mortality* combined do not exceed *regeneration*. Naturally enough, faster-growing forests produce more *timber*, which is why attention would tend to swing from **podocarps** to *beech* forests regardless of the state of the **podocarp** resource. The colonists cannot be blamed for *plunging* in without thought to whether the resource had limits. They brought from *Britain* little experience or understanding of how to maintain forest structure and a *timber* supply for all time. Under *German* management it might have been different here. The *Germans* have practised the sustained approach since the seventeenth century when they faced a *timber* shortage as a result of a series of wars. In *New Zealand* in the latter part of the twentieth century, an anticipated shortage of the most valuable native *timber*, *rimu*, prompts a similar response – no more *contraction* of the *indigenous* forest and a balancing of yield with *increment* in selected areas.

This is not to say the idea is being *aired* here for the first time. Over a century ago the first *Conservator* of Forests proposed sustained harvesting. He was cried down. There were far too many trees left to bother about it. And yet in the *pastoral* context the dangers of **overgrazing** were appreciated early in the piece. *New Zealand geography* students are taught to this day how **overgrazing** causes the *degradation* of the soil and hillsides to slide away, and that with them can go the *viability* of hill-country sheep and cattle farming. That a forest could be **overgrazed** as easily was not widely accepted until much later – so late, in fact, that the *counter* to it, sustained-yield management, would be forced upon the industry and come as a shock to it. It is a simple enough concept on paper: balance harvest with growth and you have a natural *renewable* resource; forest products forever. Plus the social and economic benefits of regular work and income, a regular *timber* supply and relatively stable markets. Plus the environmental benefits that *accrue* from minimising the impact on soil and water qualities and wildlife.

In practice, however, sustainability depends on how well the dynamics of the forest are understood. And these vary from area to area according to forest make-up, soil *profile*, *altitude*, *climate* and factors which forest science may yet discover. *Ecology* is deep-felt.

High-frequency words

In the example text, high-frequency words, including the function words *in*, *for*, *the*, *of*, *a*, and so on, are not marked at all. Appendix 3 contains a complete list of English function words. The high-frequency words also include many content words: *government*, *forests*, *production*, *adoption*, *represent*, *boundary*. The classic list of high-frequency words is Michael West's (1953) *A General Service List of English Words*, which contains around 2,000 word families, although these are not solely frequency based. Almost 80% of the running words in this text are high-frequency words. Schmitt and Schmitt (2012) argue for having a 3,000-word family high-frequency vocabulary list. Such a number, plus proper nouns, transparent compounds and marginal words, typically provides 95% coverage of a text.

Mid-frequency words

The second group of words are the mid-frequency words. They include words like *zoned*, *pioneering*, *aired* and *pastoral*, and are marked in italics in the text. There are 6,000 to 7,000 of them (depending on how many high-frequency word families one assumes there are) and they range from the third 1,000 words to the ninth 1,000 words (note that the ninth 1,000 words start with the word family number 8,001 and end with word family number 9,000). Mid-frequency words include generally useful, moderately frequent words, including many that almost got into the high-frequency word list. The mid-frequency words are distinguished from the low-frequency words because, together with the high-frequency words, they represent the amount of vocabulary needed to deal with English without the need for outside support. They are also largely general-purpose vocabulary. Schmitt and Schmitt (2012) cover a good range of reasons for distinguishing mid-frequency vocabulary from low-frequency vocabulary. One important such reason is that it highlights this vocabulary and clearly sets it as a learning goal.

Low-frequency words

Only three word families in the text are not mid-frequency words but are low frequency, beyond the first 9,000 words of English: *perpetuity*, *overgraze* and *podocarp*. They are marked in bold. Low-frequency words make up about 1 per cent of the words in this text, and there are thousands of them in the language. By far the biggest group of words, they make up only a very small proportion of the running words.

These words consist of technical terms for various subject areas and words that we rarely meet in our use of the language. Mid-frequency readers (see Paul Nation's website, www.victoria.ac.nz/lals/staff/paul-nation.aspx) are designed to provide opportunities to incidentally learn mid-frequency words by lightening the vocabulary load of the text, and this is done primarily by replacing the low-frequency words with high- or mid-frequency words.

Specialised vocabulary

For certain kinds of text, particularly academic text, there may be shortcuts that learners can take by focusing on the vocabulary which is particularly important in such texts.

Academic words

Academic texts contain many words that are common in different kinds of academic texts, *policy*, *phase*, *adjusted*, *sustained*. Typically these words make up about 9% of the running words in the text. The best-known list of academic words is the Academic Word List (Coxhead, 2000). Appendix 1 contains the 570 headwords of this list. This small list of words is very important for anyone using English for Academic Purposes (see Chapter 6). In the text above they include the words *sustained*, *policy*, *extractive*, *phase*, *adjusted*, *exceed* and so on. Davies and Gardner have developed a useful academic vocabulary list that does not build on a particular high-frequency list (www.academicwords.info).

Technical words

The text above contains some words that are very closely related to the topic and subject area of the text. These words include *indigenous*, *regeneration*, *overgraze*, *podocarp*, *beech*, *rimu* (a New Zealand tree), *timber* and *forest*. These words are reasonably common in this topic area but are not so common elsewhere. As soon as we see them we know what topic is being dealt with. Technical words like these typically cover a large proportion of the running words in a text. They differ from subject area to subject area. If we look at technical dictionaries, such as dictionaries of economics, geography or electronics, we usually find about 1,000 entries in each dictionary. Technical words however can consist of high-frequency words, mid-frequency words and what in another text would be classified as low-frequency words. As we shall see in a later chapter, technical words can make up between 20% and 30% of the running words in a text. Words from the *Academic Word List* may also be technical words in some texts.

For academic purposes, learning the *Academic Word List* and the technical vocabulary of the relevant field is an efficient way for a second language learner to cope with the vocabulary of an academic text. Figure 1.1 provides data for an academic textbook where technical words have been distinguished from the other levels of words.

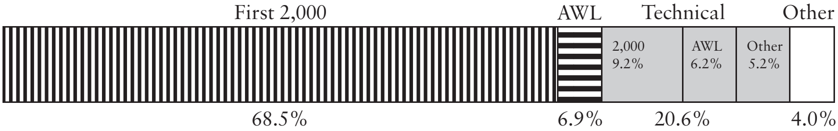

Figure 1.1 Coverage of academic text by the **General Service List**, *academic words, technical words and other vocabulary in an applied linguistics text*

In Figure 1.1, technical words have been taken out of the first 2,000 words (represented in this figure by the *General Service List*) and the *Academic Word List* (AWL) and this of course reduces their coverage to 68.5% and 6.9% respectively. As the divisions under 'Technical' in the figure show, the *General Service List* would otherwise cover 68.5% plus 9.2% (77.7%). 'Other' includes both mid- and low-frequency words. Note the large text coverage by the technical words (20.6%), and the relatively large coverage by the technical words from the first 2,000 and the *Academic Word List*.

Frequency levels in a large corpus

We have just looked at an example of a short text. Let us now look at a longer text. Table 1.3 gives figures for a collection of texts consisting of 100 million running words, namely the British National Corpus.

Each of the 20 word lists contains 1,000 word families. The proper nouns list contains over seventeen 1,000 word families, but does not include every proper noun in the British National Corpus. The compounds list contains transparent compounds like *forever*, *aftershave* and *ashtray* where the meaning of the compound is transparently related to the meaning of the parts. The marginal words list contains items like *er*, *ooh*, *aah*, *gosh*, *sshh* which are common in spoken language but are not dictionary entry words. Note the very fast drop in

Table 1.3 *Coverage of the British National Corpus by word family lists made from the corpus*

Lists	% coverage of tokens	% cumulative coverage of tokens including proper nouns, marginal words and transparent compounds
1st 1,000	77.96	81.14
2nd 1,000	8.10	89.24
3rd 1,000	4.36	93.60
4th 1,000	1.77	95.37
5th 1,000	1.04	96.41
6th 1,000	0.67	97.08
7th 1,000	0.45	97.53
8th 1,000	0.33	97.86
9th 1,000	0.22	98.08
10th 1,000	0.28	98.23
11th 1,000	0.15	98.38
12th 1,000	0.11	98.49
13th 1,000	0.09	98.58
14th 1,000	0.07	98.65
15th 1,000	0.06	98.71
16th 1,000	0.04	98.75
17th 1,000	0.04	98.79
18th 1,000	0.03	98.83
19th 1,000	0.02	98.85
20th 1,000	0.01	98.86
Proper nouns	2.57	
Marginal words	0.31	
Compounds	0.30	
Not in the lists	1.02	99.08

percentage of text coverage for the higher frequency lists. Also note in Column 3 that it takes around 4,000 word families plus proper nouns, marginal words and transparent compounds to get to 95% coverage and 9,000 word families to get to 98% coverage.

Looking at Column 2, we can see that there are 43 words per 1,000 running words from the 1,000 words at the third 1,000 level. From the 1,000 words at the ninth 1,000 level there will be around two words per 1,000 tokens, roughly around one word per 500-word page. From the 1,000 words at the twentieth 1,000 level, there will be one word in every 10,000 running words, or one in every 200 pages.

Table 1.4 *Coverage of the British National Corpus by high-, mid- and low-frequency words*

Type of vocabulary	% coverage
High-frequency (2,000 word families)	86%
Mid-frequency (7,000 word families)	9%
Low-frequency (tenth 1,000 word level onwards)	1–2%
Proper nouns, exclamations etc.	3–4%
Total	100%

Table 1.4 uses the figures in Table 1.3 to show the rough proportions of words at the high-, mid- and low-frequency word levels.

The figures in Table 1.4 are approximate and include the 1.02% of tokens indicated as *Not in the lists* in the last row of Table 1.3, which are distributed between the low-frequency words and the proper nouns and so on.

Figure 1.2 presents the data in Table 1.4 in a diagrammatic form. Proper nouns, exclamations and so on have been included with high-frequency words in the figure. The size of each of the sections indicates the proportion of the text taken up by each type of vocabulary.

There are some very important generalisations that can be drawn from Table 1.4 and the other information that we have looked at. We will look at these generalisations and at the questions they raise. Brief answers to the questions will be given here with little explanation, but the questions and their answers will be examined much more closely in later chapters.

High-frequency words

There is a small group of high-frequency words which are very important because these words cover a very large proportion of the running words in spoken and written texts and occur in all kinds of uses of the language.

How large is this group of words? The usual way of deciding how many words should be considered as high-frequency words is to look at the text coverage provided by successive frequency-ranked groups of the words (see Nation, 2001, for the effect of using a variety of criteria to decide on the boundary between high- and low-frequency words). The teacher or course designer then has to decide where the coverage gained by spending teaching time on these words is no longer worthwhile. The rapid drop in Table 1.4 shows that the group

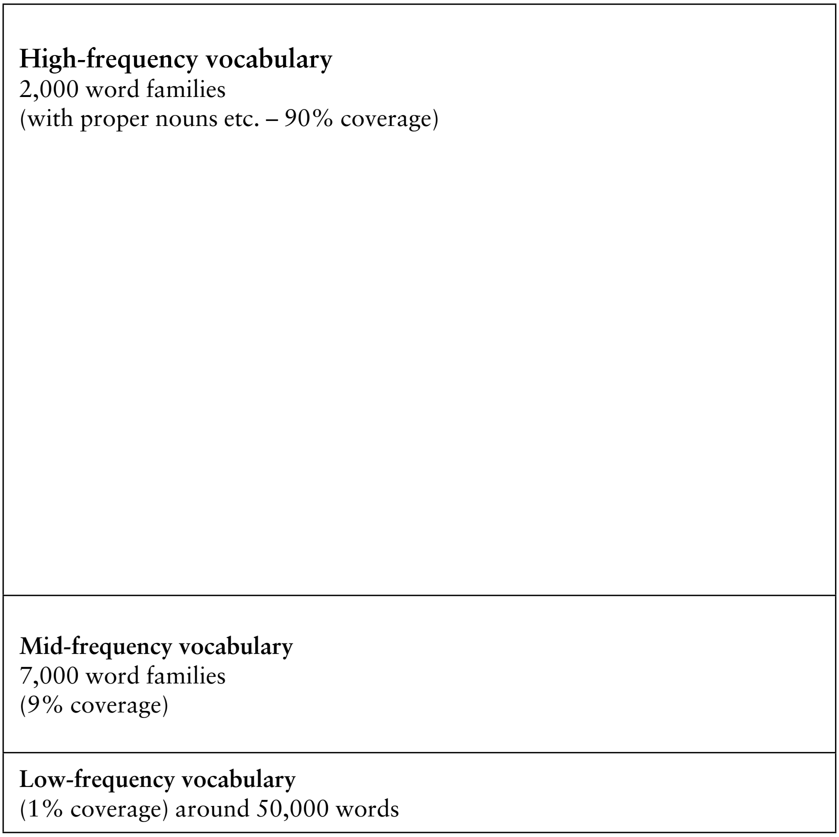
High-frequency vocabulary
2,000 word families
(with proper nouns etc. – 90% coverage)
Mid-frequency vocabulary
7,000 word families
(9% coverage)
Low-frequency vocabulary
(1% coverage) around 50,000 words

Figure 1.2 Coverage of the British National Corpus by high, mid- and low-frequency word family lists

of high-frequency words is relatively small. Schmitt and Schmitt (2012) suggest that 3,000 word families is a suitable size for the group of high-frequency words. In this book, we will stay with 2,000 but this has clearly become a matter of debate and will be affected by the reason for distinguishing high-frequency words from mid-frequency words.

What are the words in this group? The classic list of high-frequency words is Michael West's (1953) *A General Service List of English Words*, which contains around 2,000 word families. About 165 word families in this list are function words, such as *a*, *some*, *two*, *because* and *to* (see Appendix 3). The rest are content words, that is nouns, verbs, adjectives and adverbs. The older series of graded readers are based on this list. Because of its age and because it was made using other criteria besides frequency, this list is falling out of favour. The major problem with replacing it is the difficulty of making a list that is suitable for learners in the school system, and that takes account of both spoken and written language which is relevant to the learners using the lists. The first 2,000 words of the BNC/COCA lists on Paul Nation's website are also not solely frequency based but include complete lexical sets of numbers, days of the week, months and seasons. Research is continuing on the feasibility and construction of high-frequency word family lists. Making such a list is a much more difficult task than it seems (Nation and Webb, 2011: 131–55).

How stable are the high-frequency words? In other words, does one properly researched list of high-frequency words differ greatly from another? Frequency lists may disagree with each other about the frequency rank order of particular words but if the research is based on a well-designed corpus there is generally about 80% agreement about what particular words should be in the list of high-frequency words. Nation and Hwang's (1995) research on the *General Service List* showed quite a large overlap between the this and more recent frequency counts. Replacing some of the words in the *General Service List* with other words from a more recent frequency count resulted in an increase in coverage of only 1%. It is important to remember that the 2,000 high-frequency words of English consist of some words that have very high frequencies and some words that are frequent but are only slightly more frequent than others not in the list. The first 1,000 words cover about 74% and the second 1,000 about 4% of the running words in an applied linguistics academic text. When making a list of high-frequency words, both frequency and range must be considered. Range is measured by seeing how many different texts or subcorpora each particular word occurs in. A word with wide range occurs in many different texts or subcorpora.

How should teachers and learners deal with these words? The high-frequency words of the language are so important that considerable time should be spent on these words by both teachers and learners. The words are a small enough group to enable most of them to get attention over the span of a long-term English programme. They need to be met across the four strands of meaning-focused input, meaning-focused output, language-focused learning and fluency development. This attention thus should be in the form of incidental learning, direct teaching and direct learning, and there should be planned meetings with the words. The time spent on them is well justified by their frequency, coverage and range, and by the relative smallness of the group of words.

Table 14.4 lists some of the teaching and learning possibilities that will be explored in much more detail in other chapters of this book.

In general, high-frequency words are so important that anything that teachers and learners can do to make sure they are learned is worth doing.

It is sometimes argued that teachers should not do much about high-frequency vocabulary because:

1. high-frequency vocabulary occurs frequently and therefore repeated opportunities to meet these words will take care of learning, and
2. high-frequency vocabulary probably contains a lot of concrete words which are much easier to learn and retain than abstract words (see, for example, Sadoski, 2005).

However, high-frequency vocabulary does deserve some deliberate attention because:

1. it covers such a large proportion of connected spoken and written text that such text will be inaccessible until a reasonable amount of high-frequency vocabulary is known (Nation, 2006), so it needs to be learned as quickly as possible;
2. comprehension of text will suffer if learners cannot access high-frequency vocabulary with some degree of fluency (Perfetti and Hart, 2001; Rasinski, 2000); and
3. without knowledge of high-frequency vocabulary, learners will not be able to produce spoken or written text.

It needs to be noted that this argument is about the deliberate learning and teaching of vocabulary, and probably has most to do about the deliberate teaching of vocabulary. In a well-balanced course, just one-quarter of the time should be spent on deliberate study (Nation, 2007), and only a relatively small proportion of that one-quarter should involve deliberate teaching.

Mid-frequency words

There is a large group of generally useful words that occur rather infrequently, but frequently enough to be a sensible learning goal after the high-frequency and specialised vocabulary is known.

Because of the finding (Nation, 2006) that it takes around 6,000–9,000 words plus proper nouns to reach 98% coverage of the text, it is useful to distinguish mid-frequency and low-frequency words. These are largely distinguished on the basis of range, frequency and dispersion. Mid-frequency words consist of 7,000 word families from the third to the ninth 1,000, and low-frequency words are those from the tenth 1,000 onwards.

Let us consider the reasons and evidence for creating the category of mid-frequency words.

1. *Range and frequency*. In my work with the British National Corpus word families, I used the British National Corpus broken into 10 equally sized subcorpora, each of 10 million running words. It was around the tenth 1,000 word families that words with a range of 9 rather than 10 occurred. That is, they did not occur in every subcorpus. This indicates that at around this frequency level, less generally frequent words occur. This may be the point at which individual native speakers' vocabularies start to diverge according to their interests. It may be possible to gain evidence of this from data gained from the expanded version of the *Vocabulary Size Test*. For native speakers from about 13 years old, we should find comprehensive knowledge of the first 9,000 with less shared knowledge between native speakers from the tenth 1,000 onwards. The most frequent 9,000 words are likely to be of roughly equal value for dealing with spoken and written text.
2. *Coverage*. 9,000 word families plus proper nouns provide 98% coverage for novels. 8,000 word families plus proper nouns provide 98% for newspapers. The truly low-frequency words cover less than 2% of the running words.
3. *Familiarity*. Native-speaking teenagers and adults are likely to be largely familiar, at least receptively, with the first 9,000 word families. We would expect even non-literate native speakers to know the first 9,000 words of English reasonably well. High-proficiency non-native speakers, such as those doing doctoral study through the medium of English, are likely to be familiar with most of the first 9,000 word families.
4. *Learning goals*. A vocabulary consisting largely of high-frequency words is insufficient for unassisted reading of unsimplified text. It is important that learners continue to increase their vocabulary size in a systematic way at least until they gain good coverage of text. Around 3,000–4,000 words plus proper nouns provide 95% coverage of novels, newspapers and films (Nation, 2006; Webb and Rodgers, 2009a and 2009b). At least the third 1,000 to the fifth 1,000-word lists should be an explicit vocabulary-learning goal for non-native speakers who know the high-frequency words, and after that the sixth 1,000 to the ninth 1,000 words are the next rational goal. Nation (2009) looks at how much unsimplified text would need to be adapted to produce reading material that would

support these goals. Separating out mid-frequency vocabulary, providing word lists and researching the stability of such lists can raise the profile of such vocabulary and hopefully encourage the deliberate learning (not teaching) of such vocabulary and the production of helpful reading texts that focus on them and bridge the gap between current graded reader series and unsimplified texts (see Paul Nation's website for some free mid-frequency readers at three frequency levels).

What kinds of words are they? Some mid-frequency words are words that did not manage to get into the high-frequency list. It is important to remember that the boundary between high-frequency and mid-frequency vocabulary is an arbitrary one. Any of several hundred mid-frequency words could each be candidates for inclusion within the high-frequency words rather than within the mid-frequency words simply because their position on a ranked frequency list which takes account of range is dependent on the nature of the corpus the list is based on and the way it is divided into subcorpora. A different corpus would lead to a different ranking, particularly among the words on the boundary. This, however, should not be seen as a reason for large amounts of teaching time being spent on mid-frequency words at the 3,000- or 4,000- word level. Here are some words in the British National Corpus that fall just outside the high-frequency boundary: *nod*, *pupil*, *evolution*, *boast*, *glove*, *rod* and *entrepreneur*.

As Table 1.3 shows, each 1,000 level from the third 1,000 onwards provides steeply decreasing coverage of text – third 1,000 4.36%, fourth 1,000 1.77%, fifth 1,000 1.04%, sixth 1,000 0.67% and so on. Clearly, there is more value in learning the third 1,000 than in learning the fourth 1,000, and when the learners deliberately study mid-frequency words, they should largely be guided by frequency lists to make sure that the most useful mid-frequency words are learned first. Note that the text coverage of the third 1,000 words (4.36%) is almost half of the total text coverage of all the 7,000 mid-frequency words (8.84%).

How many mid-frequency words are there? The arbitrary figure for the number of mid-frequency words is 7,000 word families, from the third 1,000 to the ninth 1,000 inclusive. With the high-frequency words and proper nouns these provide 98% coverage of most kinds of text. On their own, they cover around 9% of the tokens of texts.

What should teachers and learners do about mid-frequency words? Teachers' and learners' aims differ with mid-frequency vocabulary. The teacher's aim is to train learners in the use of strategies to deal with such vocabulary. These strategies include guessing using context clues, deliberate learning using vocabulary cards or flashcard

Table 1.5 *The differing focuses of teachers' and learners' attention to high- and mid-frequency words*

	High-frequency words	Mid-frequency words
Attention to each word	Teacher and learners	Learners
Attention to strategies	Teacher and learners	Teacher and learners

programmes (Nakata, 2011), using word parts to help remember words and using dictionaries. When teachers spend time on low-frequency words in class, they should be using the low-frequency words as an excuse for working on those strategies. The learners' aim is to continue to increase their vocabulary. The strategies provide a means of doing this.

As Table 1.5 shows, learners should begin training in the strategies for dealing with vocabulary while they are learning the high-frequency words of the language. When learners know the high-frequency vocabulary and move to the study of mid-frequency words, the teacher does not spend substantial amounts of class time explaining and giving practice with vocabulary, but instead concentrates on expanding and refining the learners' control of vocabulary-learning and coping strategies. Learners however should continue to learn new words.

Low-frequency words

There is a very large group of words that occur very infrequently and cover only a small proportion of any text.

What are the low-frequency words and how many low-frequency words are there? Low-frequency words are those beyond the most frequent 9,000 words of English. There are tens of thousands of them.

What kinds of words are they? Many low-frequency words are proper names. Around 3% of the running words in the British National Corpus are words like *Carl*, *Johnson*, *Ohio* (see Table 1.3). The words in the proper nouns list cover 2.6% of the tokens, and about half of the words not in the lists are proper nouns. They make up a very large proportion of the word types in any large corpus (around 50%). In some texts, such as novels and newspapers, proper nouns are like technical words – they are of high frequency in particular texts but not in other texts, their meaning is closely related to the message of the text, and they could not be sensibly pre-taught because their use in the text reveals their meaning. Before you read a novel, you do not need to learn the characters' names.

'One person's technical vocabulary is another person's low-frequency word.' This ancient vocabulary proverb makes the point that, beyond the high- and mid-frequency words of the language, people's vocabulary grows partly as a result of their jobs, interests and specialisations. The technical vocabulary of our personal interests is important to us. To others, however, it is not important and from their point of view is just a collection of low-frequency words.

Some low-frequency words are simply low-frequency words. That is, they are words that almost every language user rarely uses. Here are some examples: *eponymous*, *gibbous*, *bifurcate*, *plummet* and *ploy*. They may represent a rarely expressed idea, they may be similar in meaning to a much more frequent word or phrase, they may be marked as being old-fashioned, very formal, belonging to a particular dialect, or vulgar, or they may be foreign words.

How many low-frequency words do learners need to know? When learners have a vocabulary size of 9,000 words and know the technical vocabulary of the subject areas they are involved in, it is useful for them to keep expanding their vocabulary. The more vocabulary that is known and the better it is known, the more effectively the language can be used. Adult native speakers have receptive vocabulary sizes of around 20,000 word families and learners who already know the mid-frequency words may want to see native speaker vocabulary size as a learning goal.

What should teachers and learners do about low-frequency words? Teachers should teach low-frequency words only when they are essential to the understanding of the text or when they are in a relevant technical vocabulary. Learners may choose to deliberately learn low-frequency words, but it is probably best to learn them largely incidentally through reading and listening. Reading is likely to provide greater opportunities for such learning because written texts typically make use of a larger vocabulary than spoken texts. Dictionary use can help in such learning, particularly where the low-frequency words are adjectives and there are few context clues to their meaning. Research also shows that incidental learning from listening is less than from reading (Brown et al., 2008). Here are some low-frequency words that I have recently looked up on my iPod while reading novels. I found it impossible to guess their meaning from context clues, largely I hope because of a lack of context clues – *adipose*, *afflatus*, *philter*, *cetacean*, *plangent*, *mephitis*, *prelapsarian*. I have yet to find an opportunity to use these in speaking or writing!

Specialised vocabulary

It is possible to make specialised vocabularies which provide good coverage for certain kinds of texts. These are a way of extending the high-frequency words for special purposes.

What special vocabularies are there? Special vocabularies are made by systematically restricting the range of topics or language uses investigated. It is thus possible to have special vocabularies for speaking, for reading academic texts, for reading newspapers, for reading children's stories or for letter writing. Technical vocabularies are also kinds of specialised vocabularies. Some specialised vocabularies are made by doing frequency counts using a specialised corpus. Some are made by experts in the field gathering what they consider to be relevant vocabulary.

There is a very important specialised vocabulary for second language learners intending to do academic study in English. This is the *Academic Word List* (see Appendix 1). It consists of 570 word families that are not in the most frequent 2,000 words of English but which occur reasonably frequently over a very wide range of academic texts. That means that the words in the academic vocabulary are useful for learners studying humanities, law, science or commerce. The list is not restricted to a specific discipline. The academic vocabulary has sometimes been called sub-technical vocabulary because it does not contain technical words but it contains rather formal vocabulary. The *Academic Word List* is drawn from words from the third 1,000 to the seventh 1,000, although in some frequency counts based on formal text, some *Academic Word List* words occur in the first 2,000.

Adding the academic vocabulary to the high-frequency words changes the coverage of academic text from 76.1% to 86.1%. Expressed another way, with a vocabulary of 2,000 words, approximately one word in every four will be unknown. With a vocabulary of 2,000 words plus the Academic Word List, approximately one word in every ten will be unknown. This is a very significant change. If, instead of learning the vocabulary of the Academic Word List, the learner had moved on to the third 1,000 most frequent words, instead of an additional 10% coverage there would only have been 4.3% coverage (see Table 1.3).

What kinds of words do they contain? The *Academic Word List* is reprinted in Appendix 1. Hirsh (2004) looked at why the same group of words frequently occur across a very wide range of academic texts. Sometimes a few of them are closely related to the topic and are in effect technical words in that text. Most however occur because they allow academic writers to do the things that academic writers want to do. That is, they allow writers to refer to others' work (*assume*, *establish*, *indicate*, *conclude*, *maintain*). They allow writers to work with data in academic ways (*analyse*, *assess*, *concept*, *definition*, *establish*, *categories*, *seek*). They also add formality and seriousness to what is being said, and in academic text and newspapers they do jobs that would otherwise be done by high-frequency words. We consider this issue again in Chapter 6.

Technical words contain a variety of types which range from words that do not usually occur in other subject areas (*cabotage*, *amortisation*) to those that are formally like high-frequency words but which may have specialised meanings (*chest*, *by-pass*, *arm* as used in anatomy). Chapter 6 on specialised vocabulary looks more fully at technical words.

How large are they? Research on technical vocabularies (Chung and Nation, 2003; Chung and Nation, 2004) shows that technical vocabulary makes up a very large proportion of the running words of a technical text. As we shall see in a later chapter, in Chung's study (Chung and Nation, 2004), around 20% of the running words in an applied linguistics text were technical words, and over 30% of the words in an anatomy text were technical words. Technical words are words that are closely associated with a particular subject area. Some technical words are not likely to be known by people who are not familiar with the subject area. Some technical words are high-frequency words, such as *cost*, *price*, *demand*, *supply* in economics, which still retain most of their generally known meaning. The size of the technical vocabulary will differ from one subject area to another. Subject areas like medicine or botany have very large technical vocabularies, well in excess of 6,000 words. Subject areas like applied linguistics or geography are likely to have smaller technical vocabularies. A rough guess from looking at dictionaries of technical vocabulary is that they are likely to contain between 1,000 and 2,000 words. If multiword units are also counted as technical words, like *gross national product*, this will then increase the size of technical vocabularies.

How can you make a special vocabulary? The *Academic Word List* was made by deciding on the high-frequency words of English and then examining a range of academic texts to find what words were not amongst the high-frequency words (the *General Service List*), but had wide range and reasonable frequency of occurrence. Range was important because the academic vocabulary is intended for general academic purposes.

One way of making a technical vocabulary is to compare the frequency of words in a specialised text with their frequency in a general corpus (Chung, 2003). Words which are proportionally much more frequent in the specialised text, or which occur only in the specialised text, are highly likely to be technical vocabulary.

What should teachers and learners do about specialised vocabulary? Where possible, specialised vocabulary should be treated like high-frequency vocabulary. That is, it should be taught and studied in a variety of complementary ways. The *Academic Word List* should be dealt with across the four strands of a course. Where the technical vocabulary is also high-frequency vocabulary, learners should be helped to see the connections and differences between the high-frequency meanings and the technical uses. For example, what is similar between a *cell wall* and other less specialised uses of *wall*? Where the technical vocabulary requires specialist knowledge of the field, teachers should train learners in strategies which will help them understand and remember the words. Much technical vocabulary will only make sense in the context of learning the specialised subject matter. Learning the meaning of the technical term *morpheme* needs to be done as a part of the study of linguistics, not before the linguistics course begins.

Zipf's law

The psycholinguist George Zipf (1935; 1949) is well known for his work on vocabulary (see Meara and Moller, 2006, for a review of one of his books), and is best known for what is now called **Zipf's law**. Zipf's law says that when we look at a ranked frequency list made from a text or a collection of texts, we can multiply the rank of the item by its frequency and always get the same answer (rank × frequency = a constant figure; see Sorrell, 2012, for a very clear description of Zipf's law). Table 1.6 shows how this works for every tenth word in George Orwell's novel *Animal Farm*.

Table 1.6 *Zipf's law applied to data from* Animal Farm

Word type	Rank	Frequency	Rank × frequency
he	10	324	3,240
farm	20	166	3,320
no	30	102	3,060
work	40	72	2,880
what	50	58	2,900
day	60	51	3,060

Note in Column 4 how the results from multiplying rank with frequency are all roughly the same, around 3,000.

If Zipf's law worked well, we could predict the frequency of any item in a frequency-ranked list if we knew the rank and frequency of a single item in the list. Zipf's law does not work with this degree of accuracy, but when we draw a curve from the application of Zipf's law, it shows us that in a text or collection of texts there will be a small number of words which occur very frequently and a very large number of words that occur infrequently. Zipf's law also allows us to predict how many word types will occur only once, twice, three times and so on in a text. Using the formula '1 divided by frequency times (frequency + 1), – $\frac{1}{f(f+1)}$ –, we can work out that half of the word types in a text will occur only once (frequency). Webb and Macalister (forthcoming) found that 42% of the mid-frequency and low-frequency word families in the New Zealand School Journals (written for children) occurred only once, and 47% of the mid-frequency and low-frequency words in a collection of newspaper and fiction texts occurred only once. If we are interested in words with a frequency of 2, the formula tells us that one-sixth will occur twice. Zipf's law describes a distribution that is not restricted to vocabulary, but applies to many natural occurrences, like the distribution of wealth and the effects of repetitions.

Even in very controlled texts, such as graded readers, Zipf's law still applies. So, it is not unusual to find lots of words occurring once in coursebooks written for learners of English and in simplified texts. Even well-designed coursebooks and graded readers will contain large numbers of words occurring only once or twice. The major effect of simplification is to remove words which are outside the word lists used to guide the simplification. This may have only a small effect on changing the range of word frequencies in the text. There is another implication of Zipf's law: Any text will contain a large number of words occurring only once or twice, and so if we wish to learn low-frequency words through meeting them in context, very large quantities of input are needed.

Zipf's law is not a rule that language producers follow. It simply describes the nature of vocabulary use. The common sense explanation of Zipf's law is that if we want to say different things we need to use some different words, but these different words will occur with common general-purpose words. It is useful to think of the extremes of this situation. If we wanted to stop Zipf's law working, we should write a text that uses exactly the same sentence which contains no repeated words over and over again. In this way, every word would have exactly the same frequency. At the other extreme, we could stop Zipf's law working by never repeating any word that we have said before. In this way, every word would have a frequency of one. If however we use language normally to speak about different things, then Zipf's law will apply. A part of the explanation of Zipf's law is that some words (function words) are essential no matter what you say, and these make up the bulk of the very high-frequency words. The most frequent 10 word types of English cover around 25% of the tokens. The most frequent 100 word types cover around 50% of the tokens. If we look at Table 1.1, we can see that important topic words (*little*, *pig*, *house*) are also likely to occur among the very frequent words, particularly if the text we are analysing is on a single topic or in a restricted topic area. In Murphey's (1992) frequency count of pop songs, which content word occurred among the most frequent ten words? Love.

So, the implications that we need to draw from Zipf's law are as follows.

1. A small number of high-frequency words will make up a very large proportion of the words in any text. Although some of these words will be function words, many of them will be content words, and it is worth learning these high-frequency words before going on to learn less frequent words.
2. A very large number of different words will make up a relatively small proportion of the tokens in a text. These words eventually need to be learned, but there are so many of them that learning them needs to be the responsibility of the learners rather than the teacher.
3. When analysing the vocabulary in a text, we can expect to see large numbers of words occurring only once or twice. That is the nature of language use. If we want to make texts accessible for learners of English, we should try to replace the low-frequency words that are outside the learners' current learning goals. There will still be many words that occur only once or twice in the text, but if these are known words or words that are currently worth learning, they will not be an overwhelming problem for the learners.
4. When reading texts where one of the goals is to incidentally learn new vocabulary, it is important to do large quantities of reading. By reading large quantities of texts on a variety of topics, learners can have a chance of getting enough repetitions to support the learning of mid-frequency vocabulary.

Zipf is also well known for another law, sometimes called the **law of least effort**. This law states that items which we use frequently tend to be short. Long and complex items tend to be less frequent. We can see this law at work in frequency counts, where most high-frequency words are short single-syllable words. In general, the longer a word is, the less frequent it is likely to be. Both of Zipf's laws also apply to grammar. Simple grammatical constructions are typically more frequent than longer or more complex related grammatical constructions, and the most frequent construction tends to be twice as frequent as the next one in the frequency list and three times as frequent as the third item in the list. A good example of this can be seen with constructions involving the word *too*, for example, *too hot*, *too hot to eat*, *too hot for me* and *too hot for me to eat*. If we draw a graph of the frequencies of these four constructions we get a rough Zipf curve.

In Chapter 5 we will look at vocabulary frequency profiles as a way of assessing productive use of written vocabulary. Edwards and Collins (2010) used Zipf's law and variations of it to evaluate the effectiveness of statistical modelling and lexical frequency profiles as ways of determining vocabulary size. Their plain language discussion of Zipf's law is particularly helpful for non-mathematicians to understand the patterned nature of word frequency distributions. Their findings support the use of lexical frequency profiles as a way of estimating the size of the homogeneous groups of learners, but they caution that lexical frequency profiles are less accurate for individuals and for groups of learners with large vocabulary sizes.

Testing vocabulary knowledge

In this chapter, a very important distinction has been made between high-frequency words, mid-frequency words and low-frequency words. This distinction has been made on the basis of the frequency, coverage and quantity of these words. The distinction is very important because teachers need to deal with these kinds of words in quite different ways, and teachers and learners need to ensure that the high-frequency words of the language are well known to them.

It is therefore important that teachers and learners know whether the high-frequency words have been learned. There are several tests available which will allow teachers and learners to see what is known and what needs to be learned.

The *Vocabulary Size Test* (Beglar, 2010; Nation and Beglar, 2007) is designed to measure a learner's total vocabulary size. There are some bilingual versions of the test available at Paul Nation's website. The learners' score on the test is multiplied by 100 to get their vocabulary size. It is useful to look at their vocabulary size in relation to the text coverage figures in Column 3 of Table 1.3. Someone with a vocabulary size of 3,000 words will have somewhere around 93.6% coverage of text, meaning that around 6% of the words in an unsimplified text will be unknown to them. That works out at about one unknown word in every 17 running words, or about 18 unknown words per 300-word page, quite a heavy vocabulary load.

The 1,000 word family levels in the test are used solely to make sure that there was no frequency bias in the sampling of the items. There are not enough items at any one 1,000-word frequency level (10 items, or 5 items in a reduced version) to give a reliable estimate of a learner's knowledge of that particular level. The test is solely intended to be a measure of total vocabulary size. That is why it is important for learners to sit all levels of the test and not just some of the earlier levels. There is also evidence from data gathered from Myq Larson's website (http://my.vocabularysize.com) that mixing items from different frequency levels results in better sustained attention to the test rather than having the learners go from easy high-frequency items to difficult low-frequency items.

The *Vocabulary Levels Test* (Nation, 1983; Schmitt et al., 2001) can be used to measure whether the high-frequency words have been learned, and where the learner is in the learning of academic and low-frequency vocabulary. There are also productive versions of the original form of the test (Laufer and Nation, 1995; Laufer and Nation, 1999; see also the freely available *Vocabulary Resource Booklet* on Paul Nation's website). See Read (1988) and Schmitt et al. (2001) for some research on this test. The test is designed to be quick to take, to be easy to mark and to be easy to interpret. It gives credit for partial knowledge of words. Its main purpose is to let teachers quickly find out whether learners need to be working on high-frequency or mid-frequency words, and roughly how much work needs to be done on these frequency bands. Before using the test, it is important to understand how it is designed and how to interpret the results. It differs from the *Vocabulary Size Test* not only in its format, but also in that it is a diagnostic test. It does not measure how many words someone knows but indicates whether learners need to be focusing on high-, academic or mid-frequency words. There are 1,000- and 2,000-level bilingual versions of the *Vocabulary Levels Test* in several languages (look in the Vocabulary Resource Booklet on Paul Nation's website). These are very useful for measuring how many of the high-frequency words are known.

In a very interesting and detailed analysis of the use of the *Vocabulary Levels Test* with his French-speaking learners, Cobb (2000) found that their performance on the *Vocabulary Levels Test* was largely a result of the ease with which they answered the items involving Graeco-Latin words (either the word itself and/or in the definition). The test was thus not measuring just learning of English, but was also measuring the learners' skill at making use of cognate relationships between French and English. Cobb then developed an L1-specific test which deliberately excluded cognates, based on the Productive *Vocabulary Levels Test* format. The results of this test correlated very highly (.9) with other proficiency measures, compared to a correlation of (.59) between the *Vocabulary Levels Test* and reading comprehension. Cobb suggests the new test was so effective because it measured actual learning, not guessing from cognates. Boyle (2009) found an appropriate balance of Germanic (33%) and Graeco-Latin words (66%) in the old *Vocabulary Levels Test*. However, unlike Cobb, Boyle found his Emirati students gained much higher scores on the Germanic words than on the Graeco-Latin words, and that some of the Graeco-Latin words that were known in the test were loanwords in the local Arab dialect of the United Arab Emirates. This lack of Graeco-Latin words could act as a barrier to successful academic reading.

Vocabulary tests like the *Vocabulary Size Test* and the *Vocabulary Levels Test*, which sample from frequency levels without concern for the L1 of the learners, will always involve a guessing from cognates effect (Nguyen and Nation, 2011). One solution is to do what Cobb did and remove such items from the tests. The problem is that then tests like the *Vocabulary Size Test* are no longer a measure of vocabulary size because significant portions of the vocabulary of the language are left out. As Cobb points out, when such tests are used, we have to realise that they are not just measuring learning but are also measuring learning burden in that cognates and loans are being answered correctly from L1–L2 parallels, not from learning. These parallels however do reflect ease of learning.

There is much more to vocabulary testing than simply testing if a learner can choose an appropriate meaning for a given word form, and we will look closely at testing in Chapter 13. However, for the purpose of helping a teacher decide what kind of vocabulary work learners need to do, the *Vocabulary Size Test* and the *Vocabulary Levels Test* are well proven, reliable and very practical tests.

Training learners in choosing which words to learn

Measuring vocabulary size is a useful step in deciding which words to learn. Barker (2007) makes a good case for training learners to take a systematic and principled approach to choosing the vocabulary they learn. He provides a very practical checklist that learners can use, noting that they are likely to feel a sense of empowerment when they find that the information they need is available through their own searching. This training should cover the following points.

1. *Sources of information about word frequency and lists of useful words*. These sources should include how to access the BNC/COCA lists, the *General Service List* (West, 1953) the Academic Word List (Coxhead, 2000) and the Academic Vocabulary List, how to use Tom Cobb's web-based version of the lexical frequency profiler (www.lextutor.ca), what dictionaries provide frequency information and how to interpret it, where word frequency lists can be found, and for the more adventurous and computer-literate learners how to make your own word frequency lists using, for example, the Frequency or Range programs available from Paul Nation's website.
2. *An understanding of the nature of word frequency.* This should relate particularly to Zipf's law which shows that from a word frequency perspective not all words are created equal, and that a relatively small number of words occur very frequently, and a very large number of words occur very infrequently. Using the Frequency program which comes with the Range program is a very effective way of bringing this message home (see also Tom Cobb's website, www.lextutor.ca). In relation to point 1 above, it is also useful for learners to realise that lists like the Academic Word List assume previous knowledge of the *General Service List*, and that usually it is best to know *General Service List* words before Academic Word List words.
3. *Practice in considering personal language needs*. The frequency level of words is a useful guide to the likely value they will give as a result of learning them. However, we all have special interests and what would be a low-frequency word for one person may be an essential word for another person. If you love a particular sport then the vocabulary of that activity is of great value to you. Barker (2007) also notes that some words are very attractive for a variety of reasons, and this attractiveness can make learning them a pleasant task. Learners may also feel gaps in their knowledge that they need to fill. When learning Japanese while living in Japan, I felt the need to be able to ask if it was all right to go into a certain part of a shrine or not. When our son started school as the only non-native speaker in the school, we taught him how to say *I want to go to the toilet* as that seemed to us as likely to be his most pressing language need on his first day at school.

Table 1.7 *A staged set of vocabulary-learning goals*

Language use	Number of words	Source of words
Survival vocabulary for foreign travel	120 words and phrases	Nation and Crabbe (1991)
Reading the easiest graded readers	100–400 word families	
Reading intermediate-level graded readers	1,000 word families	
Basic speaking skills	1,200 word families	West (1960: 38–40, 95–134: ‘A minimum adequate vocabulary for speech’)
Basic listening skills	3,000 word families	
Reading graded readers and using monolingual dictionaries	3,000 word families	*A General Service List of English Words* (West, 1953); BNC/COCA word family lists
Reading mid-frequency readers	4,000/6,000/8,000 word families	BNC/COCA word family lists
Reading unsimplified text with the help of a dictionary, and watching TV	3,000 words	BNC/COCA word family lists
Unassisted reading of unsimplified text	6,000–9,000 words	BNC/COCA word family lists

4. *The importance of knowing roughly how many words you know and what a reasonable learning goal should be in terms of number of words.* The my.vocabularysize.com website provides an easily used measure. Tom Cobb’s website provides several computerised vocabulary tests that can provide quick results. Table 1.7 suggests several useful staged vocabulary goals that the results of these tests can be related to.
5. *Options for dealing with vocabulary.* When learners meet an unknown word, they can choose what to do about it. Learners should get some guided practice in applying these options. One way of doing this is to provide the learners with a list of actions and possible reasons for those actions. They then try to justify each of the actions by matching reasons to them. The same reason can be used to justify several actions.

Actions

1. Deal with the word quickly by ignoring it or guessing it from context.
2. Find the meaning and mark the word in the dictionary so that you know you have met it before if you look it up again.
3. Find the meaning and put the word on a word card to learn later.
4. Find a meaning and work on the word now.

Reasons

1. It is a high-frequency word.
2. It is a low-frequency word.
3. It is a useful technical term for me.
4. I think I have seen this word before.
5. I have never seen this word before.
6. I can easily guess the meaning of the word.
7. I can see how this word is related to an L1 or L2 word that I already know.
8. I need to use this word receptively or productively now.
9. This seems like a word I could use often.
10. This word is one that I feel like learning.

So, the first option, of dealing with the word quickly, could be justified using reasons 2, 5, 6, 7.

6. *Ease or difficulty in learning a particular word.* Sometimes a new word will be easy to learn because it contains word parts that the learner already knows. If the learning burden is light, then for only a little effort a new word can be learned. The word may also be easy to learn because it is a loan word or cognate in the learner's L1. Words that are easy to spell and easy to pronounce may also be easy to learn. Occasionally a word may be easy to learn because of the striking and memorable situation in which it was met. Learners can be given practice in recognising word parts and should be encouraged to deliberately learn the most frequent prefixes and suffixes (see Chapter 9). Looking at a list of words and deliberately considering which have known parts and which are loan words in the L1 may also be a useful consciousness-raising activity.

The six points that we have just covered involve deciding whether to learn a particular word or not. How this deliberate learning can be done is the subject of several chapters of this book, particularly learning from word cards, using mnemonic techniques like the keyword technique and word part analysis, and using a dictionary as a learning tool. Vocabulary learning also involves knowing what to learn about a word, and it also involves making sure that there will be repeated spaced opportunities to meet, use and learn more about the word. It also involves knowing the importance of learning from input and making use of what has been learned through output.

Most of the questions looked at in this chapter will be looked at again in later chapters.

References

Barker, D. (2007). A personalized approach to analyzing 'cost' and 'benefit' in vocabulary selection. *System*, **35**, 523–33.

Bauer, L. and Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, **6**, 4, 253–79.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, **27**, 1, 101–18.

Biemiller, A. and Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, **93**, 3, 498–520.

Boyle, R. (2009). The legacy of diglossia in English vocabulary: What learners need to know. *Language Awareness*, **18**, 1, 19–30.

Brown, R., Waring, R. and Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, **20**, 2, 136–63.

Carroll, J. B., Davies, P. and Richman, B. (1971). *The American Heritage Word Frequency Book*. New York: Houghton Mifflin, Boston American Heritage.

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology*, **9**, 2, 221–45.

Chung, T. M. and Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, **15**, 2, 103–16.

Chung, T. M., and Nation, P. (2004). Identifying technical vocabulary. *System*, **32**, 2, 251–63.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review*, **57**, 2, 295–324.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, **34**, 2, 213–38.

Daulton, F. E. (2008). *Japan's Built-in Lexicon of English-based Loanwords*. Clevedon: Multilingual Matters.

Diller, K. C. (1978). *The Language Teaching Controversy*. Rowley, MA: Newbury House.

Edwards, R., and Collins, L. (2010). Lexical frequency profiles and Zipf's law. *Language Learning*, **61**, 1, 1–30.

Francis, W. N. and Kučera, H. (1982). *Frequency Analysis of English Usage*. Boston: Houghton Mifflin Company.

Goulden, R., Nation, P. and Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, **11**, 4, 341–63.

Hirsh, D. (2004). *A functional representation of academic vocabulary.* Victoria University of Wellington, Wellington.

Hu, M. and Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, **13**, 1, 403–30.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, **16**, 3, 307–22.

Laufer, B. and Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, **16**, 1, 36–55.

Leech, G., Rayson, P. and Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Longman.

Meara, P. and Moller, A. (2006). Review of The Psycho-biology of Language by G. K. Zipf. *System*, **34**, 455–7.

Mulder, K. and Hulstijn, J. (2011). Linguistic skills of adult native speakers as a function of age and level of education. *Applied Linguistics*, **32**, 5, 475–94.

Murphey, T. (1992). The discourse of pop songs. *TESOL Quarterly*, **26**, 4, 770–74.

Nagy, W. E. and Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, **19**, 3, 304–30.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, **24**, 1, 17–38.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, **5**, 1, 12–25.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing*, **10**, 1, 27–40.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, **63**, 1, 59–82.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, **1**, 1, 1–12.

Nation, I. S. P. (2009) New roles for L2 vocabulary? In Li Wei and Cook, V. (eds.), *Contemporary Applied Linguistics Volume 1: Language Teaching and Learning* Continuum, Chapter 5, pp. 99–116.

Nation, P. and Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, **31**, 7, 9–13.

Nation, P. and Crabbe, D. (1991). A survival language learning syllabus for foreign travel. *System*, **19**, 3, 191–201.

Nation, I. S. P. and Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, **23**, 1, 35–41.

Nation, I. S. P. and Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.

Neubacher, K. and Clahsen, H. (2009). Decomposition of inflected words in a second language. *Studies in Second Language Acquisition*, **31**, 403–35.

Nguyen, L. T. C. and Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, **42**, 1, 86–99.

Perfetti, C. and Hart, L. (2001). The lexical basis of comprehension skill. In Gorfien, D. S. (ed.), *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*. Washington, DC: American Psychological Association, pp. 67–86.

Rasinski, T. V. (2000). Speed does matter in reading. *The Reading Teacher*, **54**, 2, 146–51.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, **19**, 2, 12–25.

Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading & Writing Quarterly*, **21**, 221–38.

Schmitt, N. (2008). Teaching vocabulary. *Pearson Education handout*.

Schmitt, N., Jiang, X. and Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, **95**, 1, 26–43.

Schmitt, N. and Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, doi:10.1017/S0261444812000018.

Schmitt, N., Schmitt, D. and Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, **18**, 1, 55–88.

Seashore, R. H. and Eckerson, L. D. (1940). The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, **31**, 14–38.

Sinclair, J. M. (1991). *Corpus*, *Concordance*, *Collocation*. Oxford: Oxford University Press.

Sorrell, C. J. (2012). Zipf's law and vocabulary. In Chapelle, C. A. (ed.), *Encyclopaedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Swenson, E. and West, M. P. (1934). On the counting of new words in textbooks for teaching foreign languages. *Bulletin of the Department of Educational Research*, *University of Toronto*, **1**.

Thorndike, E. L. and Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College Columbia University.

van Zeeland, H. and Schmitt, N. (2012). Lexical coverage and L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, doi:10.1093/applin/ams074.

Ward, J. and Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, **37**, 461–9.

Webb, S. and Macalister, J. (forthcoming). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*.

Webb, S. and Rodgers, M. P. H. (2009a). The lexical coverage of movies. *Applied Linguistics*, **30**, 3, 407–27.

Webb, S. and Rodgers, M. P. H. (2009b). The vocabulary demands of television programs. *Language Learning*, **59**, 2, 335–66.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green and Co.

West, M. (1960). *Teaching English in Difficult Circumstances*. London: Longman.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A. and Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, **27**, 2, 201–12.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Hafner.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Cambridge, MA: MIT Press.

2 *Knowing a word*

Words are not isolated units of the language, but fit into many related systems and levels. Because of this, there are many things to know about any particular word and there are many degrees of knowing. One of the major ideas explored in this chapter is the relationship and boundaries between learning individual items and learning systems of knowledge. For example, it is possible to learn to recognise the form of a word simply by memorising its form. It is also possible to learn to recognise the form of a regularly spelled word by learning the systematic sound–spelling correspondences involved in the language. Recognition of the word then involves the application of some of the spelling rules. The relationship between item knowledge and system knowledge is complex and there has been enormous debate about certain aspects of it, for example, as it affects young native speakers of English learning to read. For each of the aspects of what it means to know a word, we will look at the item–system possibilities. A second major idea explored in this chapter is what some see as the receptive–productive scale of knowledge and how it applies to each aspect of vocabulary knowledge.

The aims of this chapter are to examine what could be known about a word, to evaluate the relative importance of the various kinds of knowledge, to see how they are related to each other, and to broadly suggest how learners might gain this knowledge. The chapter also looks at the learning burden of words, that is, what needs to be learned for each word and what is predictable from previous knowledge.

Learning burden

The learning burden of a word is the amount of effort required to learn it. Different words have different learning burdens for learners with different language backgrounds. Each of the aspects of what it means to know a word can contribute to the learning burden of a word. The general principle of learning burden (Nation, 1990) is that the more a word represents patterns and knowledge that the learners are already familiar with, the lighter its learning burden. These patterns and knowledge can come from the first language, from knowledge of other languages, and from previous knowledge of the second language. So, if a word uses sounds that are in the first language, follows regular spelling patterns, is a loanword in the first language with roughly the same meaning and fits into roughly similar grammatical patterns as in the first language with similar collocations and constraints, then the learning burden will be very light. The word will not be difficult to learn. For learners whose first language is closely related to the second language, the learning burden of most words will be light. For learners whose first language is not related to the second language, the learning burden will be heavy. De Groot (2006) presents evidence which shows that learning burden affects learning. L2 words that most closely resembled L1 spelling patterns were easier to learn and were less likely to be forgotten. Learning L2 word forms is strongly affected by the orthographic nature of the learners' L1. From an L2 English perspective, learners within L1 using the same letters have an easier job than learners with a different alphabetic system (such as Korean) who have an easier job than learners whose L1 uses characters (Chinese) (Hamada and Koda, 2008).

Teachers can help reduce the learning burden of words by drawing attention to systematic patterns and analogies within the second language, and by pointing out connections between the second language and the first language.

Teachers should be able to quickly estimate the learning burden of words for each of the aspects involved in knowing a word, so that they can direct their teaching towards aspects that will need attention and towards aspects that will reveal underlying patterns so that later learning is easier.

Do L1 and L2 words share the same lexical store?

Research shows that particularly at low proficiency levels, L2 words are directly connected to their L1 equivalents (Jiang, 2002; Kroll et al. 2002; Kroll and Stewart, 1994). Whether words are learned with L1 translations or pictures does not affect connection to the L1, it happens regardless (Altarriba and Knickerbocker, 2011; Lotto and De Groot, 1998). However, even newly learned words can also access meaning directly without going through the L1 (Finkbeiner and Nicol, 2003).

In a fascinating series of experiments, Williams and Cheung (2011) show that when learning words from another language (L2 or L3) between-language connections are made for aspects of the meaning that are context independent, that is, they are part of the core concept of the word. However, aspects of meaning that are context dependent, such as collocates, are not transferred from the L1 but need to be learned through experience with the L2. They note that newly learned words rapidly access meaning, but do not necessarily inherit all of the semantic information related to the translations with which they were paired during learning. This finding does not agree with Webb's (2009) findings which found transfer from L1. The differences may have been a result of the different ways of testing.

Williams and Cheung's findings underline the importance of learning through the four strands which involve a balance between deliberate and incidental learning, but also a balance between concept-focused and associative learning. That is, between 'learning the words' and learning through meeting and using the word. Deliberately learning an L2→L1 connection is fine, but it is only one step towards knowing the word.

The Williams and Cheung (2011) studies provide strong cross-language support for Elgort's (2011) finding that deliberate learning directly results in implicit knowledge. Newly learned L2 or L3 words can act as primes for L1 words, showing that deliberately learned L2 or L3 words can be fluently accessed subconsciously and are integrated into the semantic system.

Wolter (2001) suggests that the L1 and L2 lexicons are basically structurally similar, and that differences are caused by differences in depth of knowledge of particular words and also in the number of words known (see also Zareva, 2007). Wolter (2006) has a very interesting discussion of L1→L2 lexical and conceptual relationships, suggesting that paradigmatic relationships may require little if any modification as a result of mismatches between L2 and L1, while syntagmatic relationships like collocations are more likely to require modification, although not necessarily if there are L1→L2 parallels. Webb's research (Webb, 2008) provides some support for this idea.

The receptive / productive distinction

This section looks at what is involved in making the receptive / productive distinction in order to examine some of the issues involved in the distinction.

The validity of the receptive / productive distinction in most cases depends on its resemblance to the distinction between the receptive skills of listening and reading, and the productive skills of speaking and writing (Crow, 1986; Palmer, 1921: 118). **Receptive** carries the idea that we receive language input from others through listening or reading and try to comprehend it. **Productive** carries the idea that we produce language forms by speaking and writing to convey messages to others. Like most terminology, the terms receptive and productive are not completely suitable because there are productive features in the receptive skills – when listening and reading we produce meaning. The terms **passive** (for listening and reading) and **active** (for speaking and writing) are sometimes used as synonyms for receptive and productive (Corson, 1995; Laufer, 1998; Meara, 1990) but some object to these terms as they do not see listening and reading as having some of the other characteristics which can be attached to the term passive. I will use the terms receptive and productive and, following Schmitt (2010: 86), will use the terms **meaning recognition** and **meaning recall** for receptive knowledge, and **form recognition** and **form recall** for productive knowledge where this makes things clearer.

Essentially, receptive vocabulary use involves perceiving the form of a word while listening or reading and retrieving its meaning. Productive vocabulary use involves wanting to express a meaning through speaking or writing and retrieving and producing the appropriate spoken or written word form. Melka Teichroew (1982) shows the inconsistent use of the terms receptive and productive in relation to test items and degrees of knowing a word, and considers that the distinction is arbitrary and would be more usefully treated as a scale of knowledge.

Although reception and production can be seen as being on a continuum, this is by no means the only way of viewing the distinction between receptive and productive. Meara (1990) sees the distinction between productive and receptive vocabulary as being the result of different types of association between words. Productive vocabulary can be activated by other words, because it has many incoming and outgoing links with other words. Receptive vocabulary consists of items which can only be activated by external stimuli. That is, they are activated by hearing or seeing their forms, but not through associational links to other words. Meara thus sees productive and receptive as not being on a cline but representing different kinds of associational knowledge. One criticism of this view might be that language use is not only associationally driven, but, more basically, is meaning driven. Being able to actively name an object using an L2 word can be externally stimulated by seeing the object without necessarily arousing links to other L2 words.

According to Corson (1995: 44–5) receptive vocabulary includes the productive vocabulary and three other kinds of vocabulary – words that are only partly known, low-frequency words not readily available for use, and words that are avoided in productive use. These three kinds of vocabulary overlap to some degree. Corson's description of productive and receptive vocabulary is strongly based on the idea of use and not solely on degrees of knowledge. Some receptive vocabulary may be very well known but never used and therefore never productive. Some people may be able to curse and swear but never do. Thus Corson occasionally uses the term **unmotivated** to refer to some of the receptive vocabulary.

Corson (1995: 179–80) argues that for some people the Graeco-Latin vocabulary of English may be receptive for several reasons. Firstly, Graeco-Latin words are generally low-frequency words and thus require more mental activation for use. Secondly, the morphological structure of Graeco-Latin words may be opaque for some learners, thus reducing the number of nodes or points of activation for each of these words. Thirdly, some learners because of their social background get little opportunity to become familiar with the rules of use of the words. Corson's (1995) idea of the lexical bar (barrier) is thus important for the receptive/productive distinction.

> What the lexical bar represents is a gulf between the everyday meaning systems and the high status meaning systems created by the introduction of an academic culture of literacy. This is a barrier that everyone has to cross at some stage in their lives, if they are to become 'successful candidates' in conventional forms of education. (Corson, 1995: 180–81)

In short, the barrier is the result of lack of access to the academic meaning systems strongly reinforced by the morphological strangeness of Graeco-Latin words. For some learners much vocabulary remains at best receptive because of the lexical bar.

The scope of the receptive/productive distinction

The terms receptive and productive apply to a variety of kinds of language knowledge and use. When they are applied to vocabulary, these terms cover all the aspects of what is involved in knowing a word. Table 2.1 lists these aspects using a model which emphasises the parts. It is also possible to show the aspects of what is involved in knowing a word using a process model, which emphasises the relations between the parts. At the most general level, knowing a word involves form, meaning and use.

From the point of view of receptive knowledge and use, knowing the word *underdeveloped* involves:

- being able to recognise the word form when it is heard;
- being familiar with its written form so that it is recognised when it is met in reading;