Foreword

A glance through the pages of this book will show that it is an unusual type of publication. It consists largely of lists of words and of numbers, and looks like a cross between a dictionary and a telephone directory. These two analogies are not too wide of the mark, since this is a reference book: a book to refer to and to browse through, not a book to read through.

To be more precise: this is a **word frequency book**, a book which lists words of the English language and gives information about their frequency in actual use. Although quite a number of word frequency books have been published before (see below), a likely reaction of the present-day reader will be to ask: Why do we need to know about word frequency? What is the point of such a book?

There are a number of purposes for which knowledge about word frequency is needed, and probably the most important of these are educational.

(a) Educational needs

For the teaching of languages, whether as a mother tongue or as a foreign or second language, information about the frequencies of words is important for vocabulary grading and selection. Here frequency has applications to language learning in such areas as: syllabus design, materials writing, grading and simplification of readers, language testing and perhaps even at the 'chalkface' of classroom teaching.

Historically, the pioneering impetus[1] for frequency listings (for example, E. L. Thorndike's *Teacher's Wordbook*) in the early decades of the twentieth century was decidedly educational—see Thorndike (1921), (1932), Thorndike and Lorge (1944), Lorge (1949). It focused on the counting of word occurrences in texts used in the education of American children. Later counts were based also on magazines and general reading matter. A more modern and systematic project to obtain frequency counts from children's reading materials resulted in the *American Heritage Word Frequency Book* (Carroll *et al.* 1971). An improved kind of count (taking account of meaning but with a smaller wordlist), primarily for foreign learners of English, led to the publication of the *General Service List of English Words* by Michael West (1953—based on work begun in the 1930s).

Although these books, old as they are, have still not been entirely superseded, the lists of texts on which the frequency counts were founded strike the modern reader as decidedly dated. In fact, even when the first counts were made, they incorporated frequencies derived from books written many years before the twentieth century. These included such nineteenth century classics as Lamb's *Tales from Shakespeare*, Austen's *Pride and*

[1] Our concern here is with the English language. Word frequency lists have also been produced for other languages, such as Dutch, Italian, Japanese, Latin, Russian and Spanish (see Kennedy 1998: 16). For German, Kaeding's monumental work, which is claimed to have employed over 5,000 assistants, dates from the 1890s (*ibid*). For Spanish and French, the work of Juilland is particularly significant (see Juilland and Chang-Rodriguez 1964, Juilland et al 1970).

Prejudice, Hawthorne's *Tanglewood Tales*, and even older texts such as the United States *Declaration of Independence*, Gibbon's *Decline and Fall of the Roman Empire* and Defoe's *Robinson Crusoe* (see Thorndike and Lorge 1944: 249–55). Such works, however excellent for an education in English literature, cannot be said to represent the frequency of vocabulary in the present-day English language in any sense. Consequently, there has been a growing need for adequate frequency lists derived from more up-to-date sources.

(b) Other needs

Apart from educational applications, word frequency information can be used for **natural language processing by computer** (also known as **language technology**). In building modern language-processing software, from speech recognizers to machine-aided translation packages, it can be important to be able to determine which word, from a range of competing items, is more *likely* to occur. Yet other applications are to linguistic research—for example, in the study of style and register—and to psychological research, where frequency of vocabulary use is valuable evidence for understanding the human processing of language, whether in speaking, listening, writing or reading.

Finally, word frequency information can appeal to the curiosity of the general reader. Why, for example, in the British National Corpus on which this book is based, is *man* more than twice as common as *woman*, while the plural *women* is more common than *men*? Such observed facts of usage are worth pondering over, and may even spark off a small research project. In this book, we break up the monotony of wordlists by inserting 'interest boxes' focusing on the relative frequencies of a group of related words, such as colour words or words dealing with human kinship.

The advantages of a computer corpus

Since the early days of Thorndike and Lorge, a big transformation has taken place through the development of computers and modern computer technology. Nowadays, a very large collection of texts (normally called a **corpus**) can be stored and searched on a computer, and the frequencies of words in that corpus can be determined and listed by a fairly trivial computer program. The first to take advantage of this change were Francis and Kučera (1967), the compilers of the so-called Brown Corpus, consisting of 500 texts of varied kinds of written American English, and amounting in all to about a million words. A matching corpus of British English (the Lancaster-Oslo/Bergen Corpus) was compiled a few years later, and an equivalent frequency list for that corpus was produced by Hofland and Johansson (1982). This book also contained a comparison of the differences in frequency between the corresponding American and British corpora, thus introducing the idea that frequency lists could be **comparative**. Although these corpora were restricted to written English and were (by present standards) relatively small, they showed how a computer corpus could bring the advantages of accurate, automatic production of frequency lists, as well as providing additional statistical information on the dispersion of vocabulary through a corpus, and the distinctiveness of the vocabularies of two corpora or text collections.

A further important step forward was achieved when the Brown and Lancaster–Oslo/Bergen corpora were grammatically tagged: that is, each word in each text in each corpus was labelled with a grammatical tag (e.g. as a noun, an adverb or a preposition). After this had been done, it became possible to produce word frequency lists recognizing the **grammatical** identity of words, not merely their **orthographic** status as written sequences of letters. For example, in the Hofland and Johansson book just mentioned, the string *bear* has just one entry in each frequency list—there is no means of distinguishing between the verb *bear* (= 'carry, endure') and the noun *bear* (= 'thick-furred plantigrade quadruped'). Similarly, there is just one indiscriminate entry for *like*, a word which may be a verb, a preposition, an adjective, a noun or a conjunction, with a range of different meanings. By using a tagged corpus and these part-of-speech distinctions, later word frequency books based on these two corpora (for example, Johansson and Hofland 1989) were able to use the grammatical notion of 'word' which is found in a dictionary, and which is the basis for describing meaning.

Other frequency lists have been compiled for particular varieties of English: for example, James *et al.* (1994) is a frequency book for the vocabulary of computer science; Dahl (1979) is a frequency book for the English of psychiatric interviews. The latter is one of two existing frequency lists for spoken English (the other being an early list based on a limited corpus of 135,000 words, and published by Jones and Sinclair 1974). Although useful and interesting in their different ways, these publications cannot be said to fulfil the need for adequate and up-to-date frequency listings for the present-day English language.

The advantages and disadvantages of this book

In this book, we have aimed to satisfy the above-mentioned need, by making use of a grammatically tagged corpus—the British National Corpus—which is both large (100 times larger than the Brown Corpus) and representative of many varieties of both written and spoken English. Our claim is that this word frequency book goes far beyond any previously published word frequency book in

(*a*) using a corpus which is large enough and varied enough (100 million words) to represent an adequate cross-section of written and spoken language
(*b*) using a corpus which is far more up-to-date (dating mainly from the period 1985–94) than that used in any other comparable project
(*c*) providing frequency lists for spoken as well as written English
(*d*) providing frequency comparisons between different varieties of both spoken and written English.

This last provision is particularly important: for the various uses of frequency information mentioned earlier, particularly in the educational arena, we need to reckon on different frequency profiles for different varieties of the language. The idea that one monolithic frequency list for the whole language can satisfy all needs is, of course, unrealistic. As one small illustration of this, it is worth noting that of the top 50 words of the written part of the British National Corpus, there is an overlapping subset of only 33 words shared with the top 50 words of the spoken part of the corpus.

Having mentioned the advantages of this book over previous ones, we should end by admitting two significant drawbacks. Firstly, for reasons explained in Section 4.2.3 of the Introduction (page 14), there is a built-in element of approximation in the frequency data. On average, the margin of error is estimated to be 1.21 per cent—an amount too small to affect the interpretation of the majority of frequency figures in this book, given that they are normed to frequency per million words, but capable of affecting the figures for the most frequent words. Second, the book is far from exhaustive. For each list, we have had to recognize a frequency threshold below which a word does not qualify for listing. In all lists, this is 10 occurrences per million words or higher. If the book had been expanded to make each list complete, the result would have been a book of many thousands of pages. However, it is possible for the reader to consult exhaustive versions of each list, by visiting the Pearson Educational website—see www.booksites.net/leech—or the Lancaster (UCREL) website—www.comp.lancs.ac.uk/ucrel/bncfreq/—where such complete lists can be consulted and searched on-line.

Acknowledgements

As authors of this book, we owe a debt of gratitude to research staff and students of UCREL, who helped to develop the software, 'textware' and documentation which made this a feasible project. Roger Garside's contribution, as director of UCREL, has been indispensable. We are especially grateful to Nick Smith for his help on the grammatical tagging of the British National Corpus (BNC) and on the use of the BNC Sampler Corpus. We are also very grateful to David Lee, for hunting down errors of text classification and compiling a text information database of the existing version of the BNC: this enabled our lists to be more accurate in their reflection of text varieties. Peter Kahrel deserves our heartfelt thanks, too, for his care and expertise in the design and preparation of the book.

The British National Corpus is a collaborative initiative carried out by Oxford University Press, Longman, Chambers Harrap, Oxford University Computing Services, Lancaster University's UCREL, and the British Library. The project received funding from the UK Department of Trade and Industry and the Science and Engineering Research Council (1991–94) and was supported by additional research grants from the British Academy and the British Library. Later, in 1995–96, an additional grant from the Engineering and Physical Sciences Research Council (Research Grant GR/F99847) enabled the BNC to be retagged to a greater degree of accuracy.

Geoffrey Leech
Paul Rayson
Andrew Wilson

UCREL, Lancaster University, November, 2000

References

Carroll J B, Davies P and Richman B 1971 *The American Heritage word frequency book.* Boston, Houghton Mifflin.

Kučera H and Francis W N 1967 *Computational analysis of present-day American English.* Providence, RI, Brown University Press.

Dahl H 1979 *Word frequencies of spoken American English.* Michigan, Verbatim.

Hofland K and Johansson S 1982 *Word frequencies in British and American English.* Bergen, The Norwegian Computing Centre for the Humanities.

James G, Davison R, Cheung A H Y and Deerwester S (eds) 1994 *English in computer science: A corpus-based lexical analysis.* Hong Kong, Longman, for the Language Centre, Hong Kong University of Science and Technology.

Johansson S and Hofland K 1989 *Frequency analysis of English vocabulary and grammar.* 2 vols. Oxford, Clarendon Press.

Jones S and Sinclair J McH 1974 English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24: 15–61.

Juilland A and Chang-Rodriguez E 1964 *Frequency dictionary of Spanish words.* The Hague, Mouton.

Juilland A, Brodin D and Davidovitch C 1970 *Frequency dictionary of French words.* Paris, Mouton.

Kennedy, G 1998 *An introduction to corpus linguistics.* London, Longman.

Lorge I 1949 *Semantic count of the 570 commonest English words.* New York, Columbia University Press.

Thorndike E L 1921 *Teacher's word book.* New York, Columbia Teachers College.

Thorndike E L 1932 *A teacher's word book of 20,000 words.* New York, Columbia Teachers College.

Thorndike E L and Lorge I 1944 *The teacher's word book of 30,000 words.* New York, Columbia University Press.

West M 1953 *A general service list of English words.* London, Longman.