Ahmad S. Peyawary, *The Core Vocabulary of International English: A Corpus Approach*. Bergen: The Humanities Information Technologies Research Programme. HIT-senterets publikasjonsserie 2/99. 304 pp. ISBN 82-7283-095-7. Reviewed by **Christer Geisler**, University of Uppsala.

This monograph analyzes the scope of the core vocabulary of three major varieties of modern written English: American English in the Brown Corpus, British English in the Lancaster-Oslo/Bergen Corpus and Indian English in the Kolhapur Corpus. The purpose of this University of Manitoba dissertation is to determine the lexical items that are statistically stable across all three varieties. The author argues that there is a basic set of lexical items which occur in all three varieties, and that determining this set will be of importance not only to linguists but also to authors of teaching materials, and to language instructors.

The book comprises 304 pages, over two hundred of which are appendices of various types – there are over forty appendices in the book in the form of lists of statistically significant word groups as well as lists of word groups that did not reach statistical significance. There are nine chapters: five of them serve as introductory reading for the study proper, which begins in Chapter 6.

The study introduces concepts such as corpus design and corpus representativeness in Chapter 2. Unfortunately, one gets the feeling that some time passed between the writing of the manuscipt and its publication, as numerous references to corpus design and corpus linguistics in general are missing. In Chapter 4, the various corpora on the original ICAME CD are introduced, but mention is also made of the Birmingham Corpus (section 4.2), although this is not on the CD-ROM. The British National Corpus (BNC) is not mentioned. Occasionally, the author refers to a work without giving the year of publication, and this is slightly confusing at times.

Section 6.1 introduces the relationship between frequency and word rank in the three corpora and shows how differences in word frequencies between the three corpora gradually increase as word rank decreases. In other words, it is among the low-frequency items that we find most of the lexical differences between the three varieties. There is little discussion of the distribution of vocabulary items in language as a whole, and no general introduction to the particular behaviour of lexemes and their frequency, such as can be found in Adam Kilgariff's work on word frequencies in the BNC. Peyawary presents, however, a detailed discussion of previous word lists that have been claimed to represent a core vocabulary of English, notably Thorndike's (1927) *The Teacher's Word Book* and the more widely known *A General Service List of English Words* by West (1953). The words in the latter list are actually used for the definitions and

language examples in the *Longman Dictionary of Contemporary English* (1984: vii–ix, section 0.3).

I miss a total of word types (as opposed to the three million word tokens) in the three corpora; Johansson and Hofland (1987: Table 9) report that there are 56,166 word types in the LOB corpus alone (the Brown Corpus contains 50,406 types). Providing such data would have been both of relevance to the study as well as informative to the reader.

The study is based on the results of rank order correlations among so-called headwords across the three corpora (cf the use of rank order correlations in Hofland and Johansson 1982: 22–25). A headword is a word with all its inflected and derived forms: the headword *religious* subsumes the adjective use, the noun use, the derived adverb *religiously*, and the derived noun *religiousness*. In the statistical tests, each word contains word-class subcategorization, so that the graphic word *religious* is assigned the label *adj.n.*, indicating that the basic form can be used as an adjective and as a noun. The rank order correlations are then presented according to combinations of word class categories. For instance, pure adjectives are tested separately from multi-group categories such as adjective/noun (as in *religious* above) and adjective/adverbs (as in *only*, *just*, and *daily*).

Peyawary extracts a set of 2,114 lexical items with a frequency above 140 (there are 4,096 headwords with a frequency above 25, and these headwords account for 89 per cent of the vocabulary in the three corpora). The reason for choosing the cut-off frequency of 140 is to get a word list that is similar in size to West's (1953) two thousand words. It is argued that, on the basis of the results of statistical tests, 957 words which do not differ significantly across the three corpora form the core vocabulary. To these 957 items, another 96 words are added, in order to make up a new total of 1,053 lexical items. These 1,053 vocabulary items cover 74 per cent of the vocabulary (cf Johansson and Hofland (1987: Table 9) who find 1,021 words as covering 68 per cent of the LOB corpus). To define the core vocabulary of international English as a cluster of statistically co-occurring headwords across varieties makes perfect sense. But, it is not entirely clear why a fairly small number of mainly culture-specific words (together with some extremely frequent function words, such as the articles *a/an* and *the*) which were originally excluded from the statistical analyses are eventually reintroduced among the core vocabulary items (see Table 1).

The author's final conclusion is that the core vocabulary of international English comprises slightly more than one thousand words, rather than two thousand, as suggested by previous studies. What Peyawary's study shows is the extent of lexical items shared across the computerized corpora of modern

English, and that the frequency rankings of these lexical items do not differ significantly between the varieties. These are important findings and the study opens up new vistas for further research.

Table 1: Summary of word class distribution

Word class / word category	Number of headwords	Coverage in the three corpora
Culture-specific words and function words	96	12%
Function words:		
Adverbs	100	5%
Conjunctions	13	4%
Prepositions	33	14%
Pronouns	37	7%
Total function words	183	31%
Content words:		
Adjectives	281	10%
Nouns	288	9%
Verbs	205	13%
Total content words	774	31%
Total	1053	74%

Note: Due to the rounding-off of the results, percentages of subcategories do not add up to 100 per cent.

References

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

Johansson, Stig and Knut Hofland. 1987. *Frequency analysis of English vocabulary and grammar.* 2 volumes. Oxford: Oxford University Press.

Thorndike, Edward L. 1927. *The teacher's word book*. 2nd edition. New York: Teacher's College, Columbia University.

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