Data for lexicography

The central role of the corpus

ALLAN F. LAUDER

ABSTRACT
This paper looks at the nature of data for lexicography and in particular on the central role that electronic corpora can play in providing it. Data has traditionally come from existing dictionaries, citations, and from the lexicographer’s own knowledge of words, through introspection. Each of these is examined and evaluated. Then the electronic corpus is considered. Different kinds of corpora are described and key design criteria are explained, in particular the size of corpus needed for lexicography as well as the issue of representativeness and sampling. The advantages and disadvantages of corpora are weighed and compared against the other types of data. While each of these has benefits, it is argued that corpora are a requirement, not an option, as data for dictionary making.

KEYWORDS
Corpus linguistics, lexicography, data, linguistic intuition, citations, reading program, corpora, lexical database, type, token, corpus size, corpus representativeness, corpus balance, headword, lemma, lexeme, word form, word, general corpus, specialized corpus, monitor corpus, reference corpus, synchronic corpus

INTRODUCTION
For a dictionary to be reliable, it should provide generalizations about word behaviour that closely approximate the ways people generally use language for actual communication (Atkins and Rundell 2008: 45-46). These generalizations should be derived from evidence. Lexicographers have a number of sources of

ALLAN F. LAUDER took his PhD at the University of Atma Jaya in Jakarta with a corpus linguistic study of keywords and collocates in an Indonesian paper. He has been a guest lecturer in Postgraduate Linguistics at the Faculty of Humanities at Universitas Indonesia teaching Psycholinguistics but currently holds a full time position in the central administration of the university as Consultant to the Rector. His theoretical interests include the contribution of collocation studies to an understanding of lexical meaning, keywords, semantic fields, and phraseology. His main interests focus on the development of large general purpose Indonesian corpora, and of computer lexicons and lexical databases as well as applications, for example in translation, lexicography, and language planning. Allan F. Lauder can be reached at: allan.lauder@ui.ac.id.

© 2010 Faculty of Humanities, University of Indonesia