Wacana, Vol. 12 No. 2 (October 2010): 219—242

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.

data that can be used as evidence. They include the HEADWORD list and entries in existing dictionaries, collections of quotations or citations which throw up new words and provide contexts of use, and electronic corpora which provide very large quantities of primary data for analysis (Jackson 2002: 28-29). In addition to these, the lexicographer's own linguistic intuition, honed by training and experience, may be seen as a source of data. This paper looks at each of these sources of data, considered as a primary or sole source of data for dictionary making, looking at the advantages and disadvantages of each.

If we think about dictionary making as a practical activity that is done in stages, the article focuses on the sequence of events in a dictionary writing project that immediately follow the important decisions about who the users will be, and what type of dictionary is to be created, but precedes the task of creating individual entries within the dictionary's macro and micro-structure.[1] It is assumed that the discussion here refers to the making of larger, general monolingual dictionaries, rather than bilingual ones and small-scale, specialist, micro-lexicography projects, although the principles described here would apply to such smaller-scale works as well.

It is argued that the corpus supersedes traditional sources of data for dictionary writing, namely other dictionaries and collections of citations.

Intuition as a sole source of lexicographic data

Let us imagine that a lexicographer has no other resource for writing a dictionary than the knowledge about words and their use which are found in his or her own mind. The reliance on lexical introspection as the sole source of data for dictionary making has been called "'armchair' lexicography" (Ooi 1998: 48).

Knowledge of language consists of knowledge of WORDS[2] and knowledge about rules of syntax. Our knowledge of words is studied in psycholinguistics as the sub-field of the mental lexicon. Knowledge of words includes knowledge of meaning, both denotational and connotational, of phonology and orthography, of grammatical class and morphology. In fact, our mental

[1] Dictionary making as a sequence of stages – A note on terminology: This paper focuses on the corpus procedures used to provide information for those creating a lexical database. Within the field and practice of lexicography, it is important to distinguish the senses of, for example HEADWORD, LEMMA, LEXEME, and WORD FORM. Definitions for these terms are given in a glossary at the end of this paper. However, when working at the stage of corpus analysis, these distinctions do not yet have to come into play. Rather, it is important to focus on the computational procedures used to produce word frequency lists, and this is captured in the distinction between TYPES and TOKENS. Overlaps and subtle distinctions in meaning occur between pairings of these words, for example the relationship between TOKEN and WORD FORM, between HEADWORD and LEMMA, and between TYPE and either LEMMA or LEXEME. It is important to have a clear idea of the similarities and distinctions between these pairings when working on constructing the lexical database, and when moving to creating the dictionary entries, to keep all of these distinctions clear. However, it is not thought necessary for the purpose of this essay to provide a detailed description of this kind of thing throughout.

[2] WORDS: for definitions of this and other terms, as used in this paper, see the Glossary at the end of the paper.

dictionaries are richly elaborate and contain vastly more information than any book dictionary (Aitchison 2003: 10 and further).

We might start by asking what kinds of mental operations, from a psycholinguistic perspective, would typically be performed by lexicographers when they attempt to make a dictionary, relying solely on their own word knowledge. They would include for example such processes as thinking of a word based on different indexical features, for example spelling (what word comes after "mutiny" alphabetically?) or meaning (what are some words for kinds of furniture?), listing all the inflectional and derivational forms of a LEMMA, thinking if a word, which is known to be one part of speech ("bow" noun), is also used as another part of speech ("bow" verb), finding words with some semantic relation to the head word (what is a near synonym of "explicate"?), thinking of some collocates for a word, making a guess at how frequent or "important" a word is, deciding if a word is used primarily in a particular domain or register, deciding if the headword list isn't "missing" some word, deciding whether a new word should go into the dictionary or not, and thinking of a reasonable sounding context to illustrate the use of a word.

The skilled, professional lexicographer will have a better-informed lexicon than the average person. It will be larger and more up to date on contemporary usage (Ooi 1998: 48). The lexicographer's own store of knowledge about the language is based on their native speaker abilities, enlarged through the knowledge accumulated from their exposure to a variety of types of language genres and registers, and refined through study and professional practice. It is, however, quite likely that, despite the extent to which the mental lexicon of a lexicographer is expanded, honed, and refreshed, it will still suffer from limitations inherent in the mental lexicon of all individuals, and these are likely to have an impact on the lexicographer's work.

We can summarize the disadvantages of the use of linguistic introspection as a sole source of data for lexicography, as follows:

- The relative incompleteness of any individual's mental lexicon: Individuals know less words compared to the sum of all of the words known by all of the people in the language community of which that individual is a member. The lexicon of one individual will be shaped by individual experience and will therefore differ from that of another individual, even if there are common features shared by all. An individual will know words for some domains or registers but not all, making the individual lexicon idiosyncratic and partial (Atkins and Rundell 2008: 47).
- Limitations of access: Even if our individual word store contained everything, which it does not, our ability to search it for information is limited in a number of ways. For example, people are not able to tell you how many words they know or to enumerate all the words they know and sometimes our memories fail so we may not be able to bring to mind a word that we know we actually do know[3] – the so called "tip of the

[3] Know: have in storage in long-term memory.

tongue" phenomenon.[4]

- Limitations on the reliability of judgments about words: Judgements made by individuals about various aspects of word use are prone to subjectivity and unreliability. Types of judgments that are not always reliable include judgments about word frequency, whether a collocation is attested or not, or what the meanings of a polysemous word are. Eliciting the "default" reading of a word's meaning is prone to variation due to the effect of context.

One approach to minimizing the impact of these kinds of limitations has been for dictionary publishers to employ people whose knowledge, taken together, can be seen as representative of an entire community, with knowledge of a range of domains and genres. This is definitely a worthy approach, but although it goes a long way to overcoming the problem of partial knowledge, it still suffers from the problem of subjectivity and it is not open to evaluation. At the end of the day, using the method of lexical introspection, the dictionary will only be "as good as its lexicographer(s)" (Ooi 1998: 48). We can conclude that solely accessing one's own mental lexicon can't be seen as a reliable source of data for a dictionary. As Atkins and Rundell (2008: 47) point out, "Introspection on its own can't form the basis of a reliable dictionary".

Existing dictionaries as evidence

Traditionally, lexicographers have in some cases looked to other, existing dictionaries as the basis for the compilation of another. In particular when a concise version was to be prepared from a larger one or a bilingual from a monolingual one, existing dictionaries have been used as evidence, though the practice is not acknowledged to be common (Zgusta 1971: 239). Existing monolingual dictionaries potentially provide a headword list, an entry structure and existing definitions and examples. This can form a framework for decisions about whether to include new lexical items and remove others. The result is that more recently published dictionaries are found to a greater or lesser extent, to be rewritten versions of older works.

The reader can compare, for example, the *Kamus Umum* (KU) by Poerwadarminta (1954) and its sixth edition, *Kamus Umum Bahasa Indonesia* (KUBI) (Poerwadarminta 1978) with the *Kamus Besar Bahasa Indonesia* (KBBI) (Tim Penyusunan Kamus 1988) and KBBI-3 (Alwi et al. 2001), and then compare the KBBI with the *Kamus Bahasa Indonesia kontemporer* (P. Salim and Y. Salim 1991). When one finds similarities, one can ask if it is due simply to coincidence, homage to tradition, or to the absence of other sources of data.

The problem with using dictionaries as data in this way is that existing dictionaries are shaped by the many decisions that lexicographers take. Many of these are never explicit and therefore not necessarily open to evaluation.

[4] In a "tip of the tongue" state, the individual is conscious that the word they are looking for (in their memory) is not presently accessible, while there is also a strong sense that they indeed do know it. It is usually the case that they will recall it at some later time.

The decisions themselves involve varying degrees of subjectivity. When information is taken wholesale without reference to the current status of the language as a whole, then the information in the new dictionary will continue to be shaped by those unspoken assumptions. A dictionary which uses another as its main source of data will represent a limited and probably idiosyncratic view of the lexicon of the language. Tradition can't be relied on to provide evidence of the way the language is used.

Citations as evidence

Citations are the next kind of data that can be used for dictionary making.

What are citations?

A citation is "a short extract" selected from a text by a reader to exemplify the use of a word or phrase in an attested context, and provide evidence for its meaning or usage (Atkins and Rundell 2008: 48). Citations have also served to identify new words and examples of use in context, usually in complete sentences (Jackson 2002: 29). Citations are sometimes also called quotations. When Zgusta (1971: 225) referred to the manual collection of citations as "the excerption[5] of texts" it was still the main way to collect data for dictionary making. Citations have conventionally been collected in a reading program, an "organized data-gathering exercise" set up by a publisher where readers provided citations for a dictionary (Atkins and Rundell 2008: 49; Jackson 2002: 29). We can contrast citation used in this sense with casual citations. Casual citation is the recording by any individual of the lexical behaviour of people they come in contact with, such as family members, friends, or strangers, on a daily basis in unplanned situations. The individual notices something that stands out or seems interesting in the flow of communication. Casual citation also may be used as lexicographic evidence but it has the severe weakness of the small size of the sample and limits on the context (Ooi 1998: 48). The rest of the discussion here refers to citations gathered systematically.

History of use of citations

Citations have been in use for a considerable time, but Samuel Johnson was the first English lexicographer to use them "systematically" (Atkins and Rundell 2008: 48) as the basis for his dictionary (Hartmann 1983: 19). Johnson intended that all his entries in his dictionary be supported by citations taken from the literature of his day (Jackson 2002: 44, 166). Johnson was "unrivalled" in his own day in his use of citations, at least in the sheer number he used. He used citations not only to illustrate but "to drum home his own moral agenda" showing little respect for many of the authors (Green 1996: 230).

The Oxford English Dictionary (OED), the gold standard in historical

[5] The word *excerption*: rarer by the 1990s, it occurred only once in the 100 million words of the British National Corpus, and not in relation to lexicography but in the semantic field of psychoanalysis.

dictionaries (Read 1986: 28), was based on millions of citations (Atkins and Rundell 2008: 49; Green 1996: 316-323; Jackson 2002: 48, 166). When Murray took over the task of editing the OED from Furnivall he inherited a collection of books and source material collected over eighteen years (Green 1996: 316) and about two million slips of paper containing citations. However, due to gaps in the data, Murray initiated a reader program to provide more evidence (Green 1996: 318; Jackson 2002: 49). The entries for a lexical item in the OED were organized in chronological order (Hartmann 1983: 123) so as to show semantic change (Read 1986: 28). Murray intended that the citations should not come solely from eminent authors and prestigious literary works, but be drawn from all types of publications, even "low and trivial sources" (Read 1986: 28-29).

The reading program for Webster's Third New International Dictionary amassed about four and a half million citations to add to those collected by their team of 200 subject specialist consultants (Jackson 2002: 65).

For about two centuries, citations have played a major role in providing evidence about word use in a way that lexical intuition or reliance on existing dictionaries could not.

Advantages and disadvantages of citations

Atkins and Rundell (2008: 51-52) point out a number of advantages and disadvantages of gathering citations as data for lexicography.

The advantages of gathering citations are:

- Monitoring language change. Human readers can be extremely good at noticing new things such as the detection of newly created terms or new senses being found for existing words. They have up to now been more effective than computers at doing this. However, the use of diachronic corpora not only makes new word detection possible, but it can also identify and measure changing trends in the complementary use of near equivalent forms, for example modal verb *will* versus periphrastic modal *going to*. The evidence that corpora can provide in both these cases can be much more convincing than that provided by individual human experience.
- Gathering subject- or variety-specific terminology. Experts in a field will be good at identifying the technical terms in their field. However, subject experts who are very familiar with the terminology in their field are not likely to function simultaneously as lexicographers. In the field of terminology, the increasing power of computers is making automatic term extraction viable in a number of areas, such as legal terms in court and policy documents for languages in the European Union.
- Training lexicographers. The collection of citations helps trainee lexicographers to focus on "what counts" for items in a dictionary. Here also, in the area of training, corpora are beginning to be used to provide examples of language use in context for trainees.

The disadvantages of citations as lexicographic evidence are:

- It is a labour-intensive activity. Unless sufficient funding is available, it will be possible to gather only relatively small quantities of data. In addition, in most publishers' reading programs, a large proportion of the citations provided are not usable because the data covers what is already known (Atkins and Rundell 2008: 49). This has an important impact on the quality of the entries that can be written. The smaller the number of examples of use of a word that you have, the less likely that you can properly identify and describe its full range of uses.
- The selection of citations is subjective. This is the case even though the citations themselves are examples of attested use. The more subjective the selection, the less likely that the dictionary will serve the needs of its intended users well.
- There is a bias towards novel, idiosyncratic, rare or unconventional uses. The human mind tends to notice what is unusual rather than what is typical or conventional. This leads to skewing, something that Murray complained about more than once.[6]

Citations continue to be a useful way of providing data for dictionaries. We can enhance their use with new technology. For example, manually collected citations can now be checked against evidence on the internet. However, corpora have largely superseded citations as the central source of data for lexicography.

Corpora as evidence

What is a corpus?

A corpus[7] in linguistics is a collection (body) of texts stored in electronic format in a database. It is usually, although not always, assembled purposefully through sampling, so as to be as representative as possible of a language variety, and so that conclusions may be drawn about this variety using the corpus as data for research. A corpus is thus a sample of language use, of naturally occurring speech acts and texts, which have been chosen to characterise a state or variety of language. Corpora may be annotated, for example with each word having a part of speech (POS) tag (or label) attached, to facilitate linguistic processing. The individual texts in a corpus may also come with descriptive information such as author, date of publication, text type and so on, so that sets of texts which share particular features can be formed and so analyses can be carried out on these sub-sections of the whole corpus. See for example (Baker et al. 2006; Centre of Computational Linguistics 2006; McEnery and Wilson 2001; Sinclair 1991).

[6] Eighth Annual Address of the President to the Philological Society, Transactions of the Philological Society (1877-79), 561-586, reported in Atkins and Rundell 2008.

[7] The word *corpus* (plural *corpora*) is from Latin for "body".

Corpus types and purpose

A number of different kinds of corpus can be distinguished either according to their internal properties or their intended purpose(s). Commonly recognized types of monolingual corpora are GENERAL, SPECIALIZED, LEARNER, DIACHRONIC, and MONITOR. However, these types do not represent mutually exclusive categories. Rather, we sometimes find that pairs of corpus types, such as general and specialized, represent different ends of a spectrum, where internal properties are shaped to varying degrees by slightly different purposes.

For example, an important factor with general and specialized corpora is the purpose or purposes for which they are intended. A GENERAL CORPUS is designed to be used for a wide variety of uses, such as lexicography, natural language processing (NLP), linguistics and stylistics. The general corpus will generally be broad in scope and use internationally agreed standards for encoding. Consequently, a general corpus will contain a wide variety of text types, balanced according to genre and domain in order to represent as wide a range of the language as possible. The corpus may contain written or spoken language or both. An example of a general corpus is *the British National Corpus* (BNC) (McEnery et al. 2006: 59-60).

Meanwhile, a SPECIALIZED CORPUS is designed with a more specific purpose than a general corpus. It is intended to represent a specific variety of language, genre or domain which will be the object of study of the particular research endeavour. Examples of specialized corpora are for written English from the petroleum domain, or computer science. Another is *The Michigan Corpus of Academic Spoken English* (MICASE), a corpus of spoken academic English[8] (Simpson et al. 2002). Specialized corpora vary in size and composition according to their purpose(s). They can be relatively small, as in the case of a corpus of the works of one author, for example Shakespeare, or for frequency-based studies of grammatical behaviour. They may also be larger, for example to study particular specialist genres of language such as child language, or the language used by learners of English (McEnery et al. 2006: 60-61)

A general corpus may also be seen as a "standard reference" for the language variety which it represents and may thus be referred to as a REFERENCE CORPUS. This is because it is composed on the basis of relevant parameters agreed upon by the linguistic community and will usually include spoken and written, formal and informal language representing various social and situational strata. These corpora provide "a yardstick" for comparing successive studies with and as "a benchmark" for lexicons and for the performance of generic tools and specific language technology applications (McEnery and Wilson 2001: 32).

Another dimension that underlies corpus design is the time period during which the texts were produced. This may be a relatively short period, with

[8] MICASE is a collection of nearly 1.8 million words of transcribed speech (almost 200 hours of recordings) from the University of Michigan. It contains data from a wide range of speech events (including lectures, classroom discussions, lab sections, seminars, and advising sessions) and locations across the university.

the corpus being seen as a slice of time. It may also stretch over considerable periods of time and allow the study of language change. A DIACHRONIC (or historical) CORPUS contains examples of language from different time periods. It is used to allow tracking of language change over decades or centuries (McEnery et al. 2006: 65). Meanwhile, the term SYNCHRONIC CORPORA is used for those which are designed to represent a particular national or regional variety of English (World Englishes) and which can be used to compare varieties. There are few synchronic corpora which allow the comparison of geographical variation (dialects) (McEnery et al. 2006: 64).

A corpus which covers language from a particular, usually relatively short, period and to which no new data is added once it is complete, is seen as "static". Many corpora are like this. It will usually be designed to achieve a balance among its components by using a sampling frame. This is often useful as a reference for studying a particular time period of a variety or a language. Howevever, languages change over time. Eventually, a static corpus will no longer reflect the contemporary state of the language and will become out of date. There is a type of corpus which can keep track of the present state of the language because material is constantly or regularly being added to it annually, monthly, or daily (McEnery et al. 2006: 67). This is a MONITOR CORPUS which was conceptualized and developed by Sinclair (1991: 24-26). The monitor corpus is "dynamic", always increasing in size. This helps it keep current but makes it difficult to keep it balanced. The goal of achieving balance is largely achieved by sheer size.

In corpus design, many factors come into play and each may affect the others. A clear picture of all foreseeable purposes is essential before deciding on what the content will be and how it will be selected.

Corpus design criteria

A number of factors related to design need to be taken into consideration when planning a lexicographic corpus. Two important factors are the size of the corpus, and the issues of representativeness and balance.

Corpus size

In the design of corpora, size is an important issue. Discussion of how large corpora should be is found in a number of publications (Atkins and Rundell 2008; Biber et al. 1998; Kennedy 1998; Krishnamurthy 2002; McEnery et al. 2006; Ooi 1998). The size of a corpus is normally given in terms of the total number of words (TOKENS) in it. During the 1960s and 1970s, corpus size was constrained to about a million words by practical considerations and this led to problems such as "data sparseness" (Atkins and Rundell 2008: 57). Today, technical constraints no longer place a limit on corpus size which means that many of the criticisms levelled at the use of corpora in the 60s and 70s, for example by Chomsky, no longer are valid. Corpora have been increasing by a factor of one order of magnitude per decade since the 1970s. *The Oxford English Corpus* (OEC) broke the one

billion[9] word barrier in the 2000s and is continuing to grow (Atkins and Rundell 2008: 58). The question, "What is the maximum size for a corpus?" doesn't have a definitive answer. On the one hand there are those who suggest that there is no upper limit on the size a corpus can be. Sinclair (1991: 18), for example, states that "a corpus should be as large as possible and keep on growing". Not everyone agrees with this view. As Leech (1991: 8-29) observes, "size is not all important".

A more important question is really "What is the minimum size needed?". This will depend on the purpose for which it is intended and also some practical considerations (McEnery et al. 2006: 71). Thus, the problem of size comes back to a matter of "descriptive adequacy" (Kennedy 1998: 67). Corpus size depends on "the frequency and distribution of the linguistic features under consideration" for a particular purpose (McEnery et al. 2006: 72). Two distinct purposes are the study of lexis and the study of grammar. In general, the study of lexis in general require much larger corpora than the study of grammatical behaviour. The size of corpora required to perform quantitative studies of grammatical features can be relatively small because "the syntactic freezing point is fairly low" (McEnery et al. 2006: 72). On the other hand, in contrastive lexical studies, to model the frequency distribution of a word, it is necessary to be able to contrast it with enough occurrences of others of the same category and this will require a much larger corpus. Therefore, we can predict that a corpus designed for writing a good dictionary would need to be larger than one to be used for grammatical studies.

The reason for the difference between required corpus size for grammatical and lexical studies lies in the regularities of the frequency distribution of words in language, something usually referred to as Zipf's Law. George K. Zipf[10] (1902-1950) was a Harvard professor of philology. In the 1930s, Zipf studied the word-frequencies of texts in English, German, Chinese and Latin. He noted "the orderliness of the distribution of words[11] (Zipf 1935)" and found that "a few words occur with very high frequency while many words occur but rarely" (Zipf 1935: 40), quoted in (Atkins and Rundell 2008: 59).

[9] Billion: 1,000,000,000, a thousand million. This is the usage followed in the United States. In the UK, a billion is a million million, 1 followed by 12 zeros.

[10] Zipf: prounounced /'z?f/. The p is silent.

[11] Words: Zipf discusses the statistical properties of words in Chapter II of *The psycho-biology of language*, and in particular the relationship between word length and word frequency. He turns to the question of defining the term "word" on page 39 and further. He notes the ambiguity of the term "word" in that *child* and *children* may be seen as either one word or two, and that *give*, *gives*, *given* as one or three. It should be noted that his terminology differs from that used today. He refers to the different inflected forms (those are word forms) as "words" and the form used by lexicographers to indicate a family of inflected forms as a "lexical item". He also notes that in statistical studies of words, a failure to distinguish these two senses of word could lead to very "different quantitative results". Zipf states (1935: 40): "In the present investigation the term word will always designate a word in its fully inflected form as it occurs in the stream of speech. In the rare instances where the sense of lexical unit is intended (not until Chapter V) this unusual sense of the term will be plainly stated. The Kaeding, Chinese, Plautine investigations and the Eldridge investigation of English are each an analysis of the frequency distribution of words in fully inflected form, i.e. words and not lexical units".

In corpus linguistics, when counting the frequency of words in a corpus, it is necessary to distinguish between TOKENS, single instances of a graphic word (word form) occurring in a corpus and TYPES, word forms, seen as distinct from other word forms. In a corpus with a million tokens there may be only around 25 thousand different TYPES (Hartmann and James 1998). The term word TYPE is not specific about whether the word is to be treated separately or whether it is a member of a family of word tokens (Scott and Tribble 2006: 13; Scott 2007). This distinction only becomes important at a later stage in the dictionary making project, for example during the construction of the lexical database, or the writing of dictionary entries.

Of all the TOKENS in a corpus, a small number of TYPES account for a large proportion of total tokens while a large number of types account for a very small proportion of all tokens. For example, in a corpus of 100 million words, with approximately 160,000 types, 8,000 types will likely occur 1,000 times or more each and account for 95% of all tokens in the corpus. Meanwhile, the remaining 152,000 types will only account for 5% of all tokens (Kennedy 1998: 68).

Scott and Tribble (2006: 23) state that the 100 or 200 most frequent words in a corpus word list are mostly closed-set words, prepositions, determiners, pronouns, conjunctions. Medium frequency items come from frequency levels of around 5,000, 4,000 or 3,000 per 100 million words. At these frequencies, the words are all content words, nouns, verbs, adjectives (Scott and Tribble 2006: 25). Meanwhile, 40% of the items in the frequency list will be HAPAX LEGOMENA, appearing only once (Scott and Tribble 2006: 26). See also Kennedy (1998: 67) and Scott and Tribble (2006: 27, 29) who give the figure of around 50%.

Based on the general observation that the less the frequency, the greater the number of words, Zipf formalized this relationship in a mathematical formula (Oakes 1998: 54-55; Pustet 2004: 8; Scott and Tribble 2006: 26; Zipf 1965: 24). Zipf's Law, as it has become known, shows that there is a constant relationship between the rank of a word in a frequency list, and the frequency with which it is used in a text. When all of the words (tokens) in a corpus are placed in rank order by frequency descending, and each rank is given a number, then the rank number (r), multiplied by the frequency (f) for each token will be approximately constant (C). This is expressed as ($r \times f = C$). The relationship, one of inverse proportionality, holds for most words except those of the highest and lowest frequencies (Crystal 1997b: 87).

Zipf's findings still hold good today. This means that in any corpus, about 40 percent of all of the word types will occur only once. In lexicography, a single occurrence of a word (hapax legomena) is not enough for the lexicographer to describe how that word is used. The implication for lexicography is that in any corpus, the data for a large proportion of words will be inadequate to create an entry for them and only those words which occur often enough can find their way into the dictionary. The questions are, "How big would a corpus need to be to produce different types of dictionaries?" and "How many tokens would be enough 'as a basis for description'?" (Kennedy 1998: 67). Krishnamurthy (2002),

drawing on his experience of working on Cobuild at Birmingham University, has set out a way of working out how many headwords can be derived from corpora of different sizes. He bases his calculation on the assumption that ten tokens would be a bare minimum for adequate description for each word type. It should be noted that Krishnamurthy does not explicitly state what the relationship is between TYPE and LEMMA or between token and word form.[12] Krishnamurthy (2002) states that the number of headwords derivable can then be matched with different dictionary sizes, for example, pocket, collegiate, unabridged and English for Foreign Learners (EFL). His calculation is worked out as follows. About half of all word types in corpora appear only once. One occurrence is not enough for description, but ten might be, although this is a very "modest" figure. Many types which occur more than ten times will not appear in the dictionary. He states that there are approximately 2.2 types per lemma on average in English. The calculation involves identifying the number of word types which occur ten or more times, and dividing this by 2.2. Using this calculation to work out the maximum headwords derivable from a number of different corpora, the results are as follows:

- Cobuild corpus (1986), 18 million words: 19,800 lemmas with 10 or more tokens and therefore potential dictionary headwords.
- Cobuild Bank of English corpus (1993), 120 million words: 45,150 headwords.
- Bank of English corpus (2001), 450 million words: 93,000 headwords.

Krishnamurthy (2002) provides some examples of dictionary sizes with data taken from the Web. Pocket dictionaries are around 37,000 entries; Collegiate dictionaries range from around 100 to 200 thousand entries; Unabridged dictionaries are between 300 and 500 thousand; and EFL dictionaries are between 75 and 100 thousand.

This means, he suggests, that a corpus of 100 million words, the size of the British National Corpus, would be good enough for producing a pocket dictionary, but would "struggle to meet Collegiate requirements". He concludes that if one wishes to produce an unabridged dictionary, then a billion word corpus would be an entry level requirement, and that bigger would be better.

The calculations used by Krishnamurthy (2002) are based on the assumption that ten occurrences (TOKENS) would be adequate. However, Atkins and Rundell (2008: 60-61) demonstrate that even a hundred tokens might not be adequate for achieving "descriptive certainty" in some cases. They give the example of the verb *adjudicate* which occurs 121 times in the BNC, about 1.2 times per million. With 121 concordance lines, it is possible to create a reliable description of the word's use that includes about seven different

[12] Although he is not explicit about this, we can still take his explanation seriously. Firstly, he is providing a relatively simple formula that is to be seen as a rough guide. It is a model that leaves out some of the details. Secondly, his writing carries considerable credibility because of his years of experience as a professional lexicographer, for example with Collins on COBUILD and with Helicon, Macmillan, Routledge, and CUP.

points. They suggest, however, that less than a hundred occurrences might not be enough to achieve this degree of certainty of description. Words like *temerity* (73 occurrences), *exasperating* (45), *inattentive* (31), and *barnstorming* (20) are not so rare that they "fall outside the scope of a standard learners' dictionary", but it is doubtful that 20 occurrences, in the case of *barnstorming*, would be enough.

Furthermore, the number of occurrences required for describing single lexical items would be a bare minimum. If you wish to describe a word's phraseology, you would need a great deal more data than even a hundred lines.

Atkins and Rundell (2008: 60-61) give the example of the verb *break*. It occurs nearly 19,000 times in the BNC, which might be thought to be more than adequate for lexicographic description. However, high frequency words tend to be polysemous. At least twenty distinct senses can be distinguished for *break* along with a dozen phrasal verbs, some of which are also polysemous. The word is also found in combination in many collocations, phrases, and grammatical patterns. Among these different phrases, patterns, and collocations, some are frequent and some rare. This means that some uses of *break* are important in a particular domain but rare in a large general corpus. For example, there are only eight occurrences of the phrase "break someone's serve/service" (in the field of tennis) (Atkins and Rundell 2008: 61).

In addition, they note that their own lexical profiling software, Word Sketch Engine,[13] only works well for lemmas with at least 500 occurrences (TOKENS). In my own experience of collocation analysis, between one and five thousand occurrences of a word (TOKEN) might be needed to obtain a hundred or so collocates.

It therefore appears simplistic to set a minimum requirement for a word's frequency in a corpus at ten or a hundred even five hundred. We need to take into account whether the word is polysemous, is involved in complex lexico-grammatical patterning, or in a rich phraseology.

Atkins and Rundell (2008: 61) state that "there is no definitive minimum size" for a corpus to be used for lexicography. However, the arguments presented here demonstrate that in a corpus of any particular size, due to Zipf's Law, a large proportion of words will be very rare and the scarcity of data for them would mean it would be difficult to reliably describe them. If we wished to describe these rare words, we would need to increase the size of that corpus manyfold to get to a point where the corpus could provide enough examples of those words' use for lexicographic description. In the case of monitor corpora, it is possible to reach a point of "saturation" where the addition of new material will not substantially affect the composition or proportion of some of the data. However, it is also possible that no corpus,

[13] The Sketch Engine (SkE, also known as Word Sketch Engine) is a Corpus Query System incorporating word sketches, grammatical relations, and a distributional thesaurus. Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. The SkE is in use for lexicography at Oxford University Press, FrameNet, Collins, Chambers Harrap, Macmillan and elsewhere. Its capabilities are described in Kilgarriff et al. 2004.

no matter how big it gets, would no longer have HAPAX LEGOMENA. If that was the case, there will always be a good number of words in the language which cannot be reliably described.

If we take into consideration the continued growth trend of corpora, we can conclude that Krishnamurthy's (2002) figure of a billion words being suitable for producing an unabridged dictionary probably no longer applies. Rather, a ten billion word corpus would be a more reliable starting point for such a dictionary and that even a corpus of this size should be seen as a starting point, as something that could be added to. In this way, it should be possible to keep up with changes in the language and to provide opportunities for expanding the range of lexis which can be described and improving the degree of detail for those already covered.

Corpus content: representativeness and balance

Representativeness and sampling

The lexicographic corpus has to provide data from which generalizations can be drawn about some variety of language or other. When the corpus is designed to study very large varieties or "language" as a whole, then it would either have to contain everything or be a sample of that totality. Language, however, in the sense of a form of communication used by a community of speakers, would have to mean everything, all of the communicative events occurring, the totality of texts produced, spoken or written. While the population of actual individual texts that constitute a language must be finite, in practical terms it must be seen as unlimited because there is no way that all texts produced can be inventoried. Any attempt to create a corpus of "language" would obviously be a gargantuan, utopian, or even Sisyphean task. If we wish to study large varieties or even "language" itself, then sampling is the only option (McEnery and Wilson 2001: 29). We must attempt to make the corpus as representative of that totality as possible.

Representativeness has been recognized as a fundamental issue in corpus construction (Biber 1993: 243) "ever since linguists started using corpora" (Teubert and Cermáková 2004: 113), and "underlies the whole question of corpus design" (Barnbrook 1996: 24).

The sample should be "maximally representative" of the variety under investigation (McEnery and Wilson 2001: 30). A sample is assumed to be representative if what we find for the sample also holds for the general population (Manning and Schütze 1999: 119).

The concept of representativeness and sampling is well-established in the social sciences, the idea being that a sample can be used to make generalizations about the whole population (Kennedy 1998: 62). This is possible because the population can usually be well-defined and is limited in extent, for example, all heads of households in the capital with incomes of between two and four thousand dollars a month. However, natural languages do not lend themselves to analysis using this model and there are problems with using this approach when studying language. Most importantly, it is

difficult to define what language is because this involves defining the language community and who the speakers and writers are. Do we include immigrants and foreign language learners or exclude them; what are the boundaries of the geographical area of the language's use? What about dialects and so on? (Teubert and Cermáková 2004: 116). In addition, it is difficult to define what the population of all communicative events or texts in the language would be (Atkins and Rundell 2008: 63). In order to obtain a representative sample from a population, it is necessary to define both the sampling unit and the boundaries of the population (McEnery et al. 2006: 19). A sampling unit can be any unit of language, a book, a periodical, an article, a newspaper. "The population is the assembly of all the sampling units while the list of sampling units is referred to as the sampling frame" (McEnery et al. 2006: 19). This is a kind of stratified sampling. In this approach, the whole population is divided up into sub-categories or types and random samples are taken from each group.

Two approaches have been taken for building corpora of written texts. The first is based on a comprehensive bibliographical index. The sampling units for the LOB corpus was written British English text published in the United Kingdom in 1961. The sampling frame was taken from the *British National Bibliography Cumulated Subject Index 1960-1964* for books and *Willing's Press Guide* for periodicals (McEnery et al. 2006: 19). The second is to define the sampling frame as the contents of a particular library which belong to the variety and period in which the researcher is interested in. For the above case, this might be defined as all the German-language books and periodicals in the Lancaster University Library which were published in 1993. This approach was taken for the Brown corpus (McEnery and Wilson 2001: 78-79). In Indonesia, it would be problematic because of the absence of these kinds of catalogues or indexes. It would also be harder to do with informal language such as conversations or private correspondence because these are not published and so not found in libraries.

Further, there are also problems in deciding how to define the subcategories and assign texts to them, or what proportion of each category to include, or how to deal with overlapping categories (Atkins and Rundell 2008: 64).

Should the corpus designer take into account such things as topical domains, genres and registers? Problems arise with complex categories like domain and genre because these can be structured. Newspapers represent a genre of text, but contain multiple sub-genres, like news article, leader, editorial, or opinion piece.

The decision about categories for the sampling frame also involve considerations about their relative importance so that we can decide on what proportion of the total corpus they should be. However, this is not easy to determine because the factors that may determine such a decision are not easy to determine. Considerations may include such things as the frequency of their use and also differences in production and reception (Kennedy 1998: 63). A common problem is in deciding whether the corpus should include

more speech than writing. However, in comparing speech and writing, if the quantity of output (production) is the deciding criteria, no one actually knows what proportion of words produced in a particular time period, such as a day, are spoken and what are written. It is possible that the majority of daily communication is face to face "conversation", spoken rather than "written". Another way to look at it is that speech is "ephemeral", while written texts are "enduring". Does the longevity of written language make it more valuable? Then there is the criteria of reception. If "audience size" is a factor, then conversation, which only involves pairs or very small groups, is less important than news media texts which are read by thousands or millions. A conversation in a store involving a sales transaction might be heard by only the buyer and the seller. Meanwhile a similar dialogue, as part of a television drama, might be heard by extremely large numbers of people. Should the "influence" of a text be a factor? A pulp fiction novel may be read by millions, but the work of an award winning author may be read by less people but be studied over many years in schools and universities. In the case of newspaper texts, would "audience size" be more of a deciding factor than "quality" meaning the tabloid *News of the World* would be seen as more important than the broadsheet *The Times*? What about regional papers with small circulation figures? In general, contemporary or current works might be favoured. However, a "classic" work, one which continues to exert influence, might still be considered. Finally, in regard to the balance between speech and writing in a corpus, there are pragmatic constraints. Speech may be found to be the most important, yet, this is the most difficult and expensive to obtain.

Other criteria that have been suggested for making decisions about representativeness include "influentialness" in the literary or academic world, demography, typicality based on the subjective judgements of native speakers, and availability (Kennedy 1998: 63-64; Summers 1991).

It might seem that one way around the problem of the definition of criteria used as a sampling frame for corpus construction was so problematical that it could be avoided altogether if corpora were constructed not on the basis of such socially determined categories but on the basis of the internal, or linguistic characteristics of the texts (Otlogetswe 2004). However, a number of authors maintain that the text attributes or parameters used to do this should be extra-linguistic (external) and independent of linguistic criteria (Atkins et al. 1992: 5-6; Biber 1993: 256; McEnery et al. 2006: 14; Sinclair and Ball 1995). The reason for this, pointed out by Sinclair, is that any conclusions about text types based on word frequency distribution in such a corpus would be circular and invalid.

Considerations such as these have been discussed in great detail since the 1990s. Recently, the problem has been compounded by the emergence of new text types that are the result of the growth of new media such as the internet, social networking websites like FaceBook and smart phones with short messaging and chatting. Some of these new forms of communication don't fit neatly into the traditional written versus spoken mode distinction

(Atkins and Rundell 2008: 65).

It turns out that it is practically impossible to define the whole population which the corpus is supposed to represent, and consequently, it would therefore be "logically impossible to establish what the 'correct' proportions of each component" should be (Atkins and Rundell 2008: 66).

Balance

While the goal of building a representative corpus is "unachievable", it remains a worthy "aspiration". Some principles can guide the process of corpus construction. One such principle would be to avoid sampling from only one text type, a "monolithic" corpus such as one constructed solely from news media texts. It is tempting to use news media corpora. They are easily available and often extremely large because of the availability of news in electronic form. However, this is to be avoided for lexicographic work because no matter how big a corpus of news texts is, it will never contain the kind of lexis that would be found in other genres, such as literary or academic. This means that it is advisable to sample from as wide a range of text types as possible. However, practical considerations constrain our ability to define and sample from all known text types to create a truly "representative" corpus. A modest, practical compromise between these two extremes would be to try to create a "balanced" corpus (Atkins and Rundell 2008: 66).

A "balanced" corpus is a rational compromise but because it involves many subjective decisions and is also shaped by practical considerations like budgets and time frames, it can never be considered the result of a "scientific process" (Atkins and Rundell 2008: 66).

A "balanced" corpus, however, has a number of advantages. The use of good criteria allows the set up of a useful typology of text types. If stratified sampling is used to identify candidate texts for each text category, the result will be systematic and will reflect the actual types. If each text is labelled with information about its key features, such as genre, authorship, date of publication and so on, then users will be able to query subsets of the corpus to research how these things influence language use (Atkins and Rundell 2008: 66).

Balance, the range of text categories in a corpus is a significant factor in representativeness, and what would be an "acceptable balance" is "determined by its intended uses" (McEnery et al. 2006: 16). A general corpus should contain a wide range of text categories which need to be sampled proportionally to some rational and explicit estimate of the population so that "it offers a manageably small scale model of the linguistic material which the corpus builders wish to study" (Atkins et al. 1992: 6). However, achieving balance is more "an act of faith" than a statement of fact or the result of some scientific measure. While there is no overall agreement on how to achieve balance, work in text typology, the classification and characterizing of text categories, is relevant to such attempts.

Another acceptable approach is to emulate existing corpora which are

generally acknowledged to be balanced. The designers of the British National Corpus have made a number of subjective decisions about balance but they have done so in a way that is as reasonable as any (McEnery et al. 2006: 17).

Even though achieving representativeness and balance may be highly problematical, this can not be seen as a reason for avoiding corpus linguistic work or for dismissing the results of corpus analysis as unreliable or irrelevant (McEnery et al. 2006: 19). Corpora have been for some time and are likely to remain the central type of data for the creation of dictionaries.

Advantages and disadvantages of corpora

Advantages of corpora

Corpora have the following advantages as data for lexicography as compared with other types of data:

- The corpus has a central role to play in lexicography because it can successfully provide data in the large quantities needed for drawing conclusions about "normal" language events. The corpus can do this in a way that citation banks cannot (Atkins and Rundell 2008: 54).
- The procedures involved in corpus creation, querying and lexical database construction can be described as empirical. They thus stand in contrast to rational approaches such as the use of linguistic intuition.
- Corpus data is stable, and can be revisited many times until the lexicographer is satisfied that the decision taken is correct. Also, the software allows different ways to look at a word and its contexts and collocates. It is possible to see these contexts together in systematic ways that would be impossible in any other way. This all makes the data exceptionally valuable for semantic analysis. Thus, the problem of the "slippery" nature of meaning and of variation due to context, can be managed.
- Corpora favour everyday language over 'high quality' language. Corpora are designed to be inclusive of many text types, rather than being based on a priori assumptions about what kinds of writing exemplify "good" usage.
- Corpora can be constructed so that their contents represent the speech acts of a discourse or linguistic community. For example, a corpus consisting of news articles published by a newspaper is synonymous with the community of people who produced them. Thus, conclusions derived from a corpus, unlike the discourse analysis of a single text, can be seen as applying to that discourse community. This is a step towards a socially contextualized linguistics.
- Corpora allow the discovery of previously unnoticed patterns in language. For example, the collocate patterns of phraseology reveal the semantics of stretches of language that had previously escaped definition. An example is the phrase, "I'm no boy scout". While boy scout has the sense of "a member of the scouting movement", in the context of the phrase "I'm no …" it means someone with the admirable qualities a scout is supposed to

have. However, further study of the other collocates in the noun position reveal that they fall into three main classes, admirable people (saint, gentleman), disreputable people (fool, idiot, villain), or type of person by occupation (judge, historian). This reveals that the phrase "I'm no + noun" itself has a definable sense and function. It seems to be self-description by comparison. It is used as an excuse for not having desirable qualities in the admirable and occupation classes and as a defence for the disreputable class.

Disadvantages of corpora

It is realistic for corpus lexicographers to acknowledge that the use of corpora involves a number of real world constraints. Atkins and Rundell (2008: 55-57) describe a number of these:

- The corpus is a sample. A corpus is a finite entity. For some types of text, it may be possible to include all examples in the corpus. All the articles published by a newspaper, or all the novels written by an author are cases. However, general corpora are supposed to represent the language of entire communities of language users. They thus will inevitably be only very small samples.
- The concept of representative sample is difficult to achieve. For a corpus to be representative, it has to be based on clear criteria. While this task may be achievable for small, specialist corpora, it would be far more challenging to do this for larger, general purpose corpora.
- Corpora are the result of pragmatism and compromise. Even for corpora that are the result of careful planning and clear, explicit decisions about their design, they will still be the product of subjective decisions made by one group. They will also suffer from real-world constraints such as certain kinds of texts not existing, or the refusal of publishers to waive copyright restrictions, or limitations on the size of the spoken part of the corpus due to the expense of data collection. The degree of detail in coding is also a matter where subjectivity comes in.

Conclusions

By thinking about the nature of the data used to make dictionaries, we can see that it makes sense for lexicographers to ensure that the data they use is the most reliable and representative and also presented in such a way that reduces the probability of subjectivity or other problems listed above.

Studies of collocation in corpora demonstrate very clearly that word meaning does not reside in single lexical items but in the way words pattern together and in the way words appear in different contexts, and also in the way our minds respond to words in contexts. This view is supported in the contextual approach to meaning in lexical semantics, a view of meaning in which a word's semantic properties are fully reflected in its relations with its contexts (Cruse 1986: 1, 16). Words have been described as "slippery customers" (Aitchison 2003: 41 and further), and their meaning subject to

change in different contexts. The principles which govern the workings of the mental lexicon are not as far as I am aware subject to major modification by conscious decision. Therefore, it makes sense to provide the conscious mind with the kind of data that will best lend itself to study and will allow it to come up with the least subjective and most representative judgments about words and their use. That kind of data comes from corpora, rather than from existing dictionaries, or citations, or even linguistic intuition.

Just because corpora can and should play a central role in providing data for lexicography does not mean that other sources of information need be totally ignored. For example, Schryver and Prinsloo (2001) state that while corpora are preferable for lemma-sign list creation, this does not mean that alternative methods for the creation of the dictionary's macrostructure such as intuition-based compilations have no merit. They provide examples from the compilation of a dictionary for Sepedi, also known as Northern Sotho, which belongs to the Bantu language family (S32 in Guthrie's classification) and is one of South Africa's eleven official languages. They say that when corpora are absent, despite the shortcomings of individually compiled word lists, if a number of lists are combined, then the result may approximate the words obtained from a corpus of at least two million words.

Because the mind is subject to so many limitations, we should perform psycholinguistic studies to identify the particular sub-processes that are involved when lexicographers are making judgments about lexical behavior in the course of making a dictionary. The methods of psycholinguistics would be suitable for answering the question of whether the kind of limitations that apply in general to the workings of the mental lexicon also apply at least to some extent to the minds of trained professional lexicographers.

Glossary

HAPAX LEGOMENON: A word or phrase which occurs only once in a corpus and which may therefore not be included in a dictionary; a nonce word (Hartmann and James 1998).

HEADWORD: "The form of a word or phrase which is chosen for the lemma, the position in the dictionary structure where the entry starts" (Hartmann and James 1998); the vocabulary and other items that the editors of a dictionary have chosen for inclusion in a dictionary (Jackson 2002: 25). The headwords in a general-purpose dictionary are not just lexemes, but include "derivational affixes and combining forms", or in some cases names (Jackson 2002: 25).

LEMMA: the position at which an entry can be located in a reference work; a synonym for headword or even the whole entry (Hartmann and James 1998). The lemma is "an abstract representation, subsuming all the formal lexical variations which may apply: the verb WALK, for example, subsumes *walking*, *walks* and *walked*" (Crystal 1997a). The lemma is "the base form", typically the "stem" or the simplest form (singular noun, present/infinite of the verb) under which the word is assigned its place (Halliday 2004: 6).

See also Atkins and Rundell (2008: 162) and Hunston (2002: 18).

LEXEME: "a word in the vocabulary of a language" (Jackson 2002: 2); "A basic unit in the linguistic study of vocabulary" (Hartmann and James 1998); the smallest unit in the meaning system of the language that can be distinguished from other similar units (Stubbs 1996: xv); "the minimal distinctive unit in the semantic system of a language" (Crystal 1997a); the smallest unit in the meaning system of a language that can be distinguished from other similar units (Richards and Schmidt 2002). Lexemes may be seen as a combination of form with a meaning in a particular grammatical context. Lexemes may occur as simple words, morphologically complex words, phrasal and compound words, "multi-word expressions", and shortened forms (prefabs) (Hartmann and James 1998), and also idiomatic phrases such as kick the bucket (= die) (Crystal 1997a). Lexemes are abstract units in the sense that a single lexeme may be found used with a number of different orthographical or phonological forms, for example the inflected forms of a verb like walk (Crystal 1997a). See also Stubbs (1996: xv), Jackson and Zé Amvela (2000: 63), Richards and Schmidt (2002). The term lexeme may also refer either to the units in dictionaries listed as separate entries (Crystal 1997a). "[A lexeme] may occur as a headword in a dictionary" (Jackson 2002: 2). Lexeme is also used to refer to the sub-entry (Richards and Schmidt 2002). When lexemes are used in dictionaries as headwords, they are cited in their canonical forms (Hartmann and James 1998).

TOKEN: (in corpus linguistics) a single instance of a graphic word (word form) occurring in a corpus. Frequency counts in corpora are usually made as a number of tokens (Hartmann and James 1998). See also TYPE.

TYPE: (also word type) (in corpus linguistics) a word form, seen as distinct from other word forms, in a corpus. "The individual examples of different words or combinations of words occurring in a given CORPUS". A count of the number of different graphic words in a corpus is a reference to types. In a corpus with a million tokens there may be only around 25 thousand different types (Hartmann and James 1998). See also TOKEN.

WORD FORM: (in corpus linguistics) a string of characters between spaces in a text, a single member of a word family; "an inflectional variant of a lexeme" (Jackson 2002: 4). The term is also used in contrast with lemma, which is a collection of systematically related word forms that are thought to share the same meaning (Sinclair 1991: 176g, 2003). In corpus linguistics, word is often used in the same sense as word form to refer to a sequence of valid characters with a word separator at each end. It is seen as the simplest item for a computer to search for. This definition conceals the need to make decisions about what the valid characters and punctuation are and whether word types will be treated separately or as a member of a family of word tokens (Scott and Tribble 2006: 13; Scott 2007).

WORD: The term 'WORD' is ambiguous and therefore a degree of ambiguity is inevitable in any definition of what a word is, no matter how careful we are (Jackson and Zé Amvela 2000: 52). Word may be used to refer to tokens,

types or lemmas (Atkins and Rundell 2008: 162). In this article, it is mostly used to refer either to tokens or to word forms.

References

Aitchison, Jean. 2003. *Words in the mind; An introduction to the mental lexicon*. Third Edition. Malden, MA.: Blackwell. [First Edition 1987.]

Alwi, Hasan, Dendy Sugono, and Tim Penyusunan Kamus (eds). 2001. *Kamus Besar Bahasa Indonesia*. Third Edition. Jakarta: Pusat Bahasa, Departemen Pendidikan Nasional. [First Edition: Tim Penyusunan Kamus 1988, Jakarta: Balai Pustaka.]

Atkins, Sue, Jeremy Clear, and Nicholas Ostler.1992. "Corpus design criteria", *Literary and Linguistic Computing* 7(1): 1-16.

Atkins, Sue, and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Baker, Paul, Andrew Hardie, and Tony McEnery. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Barnbrook, Geoff. 1996. *Language and computers; A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press. [Edinburgh Textbooks in Empirical Linguistics Series.]

Biber, Douglas. 1993. "Representativeness in corpus design", *Literary and Linguistic Computing* 8(4): 243-257.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics; Investigating language structure and use*. Cambridge: Cambridge University Press. [Cambridge Approaches to Linguistics, Jean Aitchison (ed.).]

Centre of Computational Linguistics. 2006. *Systematic dictionary of corpus linguistics*. Kaunas, Lithuania: Centre of Computational Linguistics, Vytautas Magnus University. [Online address: http://donelaitis.vdu.lt/publikacijos/SDoCL.htm.]

Cruse, Alan. 1986. *Lexical semantics*. Cambridge and New York: Cambridge University Press. [Cambridge Textbooks in Linguistics Series.]

Crystal, David. 1997a. *A dictionary of linguistics and phonetics*. Fourth Edition. Oxford, UK and Cambridge, Mass.: Blackwell. [First Edition 1980.]

Crystal, David. 1997b. *The Cambridge encyclopedia of language*. Second Edition. Cambridge: Cambridge University Press. [First Edition 1987.]

Green, Jonathon. 1996. *Chasing the sun; Dictionary-makers and the dictionaries they made*. London: Pimlico.

Halliday, M. A. K. 2004. "Lexicology", in: M. A. K. Halliday, W. Teubert, C. Yallop, and A. Cermáková (eds), *Lexicology and corpus linguistics; An introduction*, pp. 1-22. London and New York: Continuum.

Hartmann, R. R. K. (ed.). 1983. *Lexicography; Principles and practice*. London: Academic Press.

Hartmann, R. R. K. and G. James. 1998. *Dictionary of lexicography*. London: Routledge.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press. [Cambridge Applied Linguistics Series, Michael H. Long

and Jack C. Richards (eds).]
Jackson, Howard and Etienne Zé Amvela. 2000. *Words, meaning, and vocabulary; An introduction to modern English lexicology*. London and New York: Cassell. [Open Linguistics Series.]
Jackson, Howard. 2002. *Lexicography; An introduction*. London and New York: Routledge.
Kennedy, Graeme D. 1998. *An introduction to corpus linguistics*. London and New York: Longman.
Kilgarriff, Adam, Pavel Rychly, Pavel Smrž, and David Tugwell. 2004. "The sketch engine", *Euralex* (July): 105-116. [Online address: http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf.]
Krishnamurthy, R. 2002. *Corpus size for lexicography*. [Corpora-list; Online address: http://www.hit.uib.no/corpora/2002-3/0254.html, accessed: 25 July 2010.]
Leech, Geoffrey N. 1991. "The state of the art in corpus linguistics", in: K. Aijmer and B. Altenberg (eds), *English corpus linguistics; Studies in honor of Jan Svartvik*, pp. 8-29. London: Longman.
Manning, C. D. and H. Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA.: MIT Press.
McEnery, Tony and Andrew Wilson (eds). 2001. *Corpus linguistics; An introduction*. Second Edition. Edinburgh: Edinburgh University Press. [First Edition 1996; Edinburgh Textbooks in Empirical Linguistics Series.]
McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies; An advanced resource book*. London and New York: Routledge.
Oakes, Michael P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press. [Edinburgh Textbooks in Empirical Linguistics Series.]
Ooi, Vincent B. Y. 1998. *Computer corpus lexicography*. Edinburgh: Edinburgh University Press. [Edinburgh Textbooks in Empirical Linguistics Series.]
Otlogetswe, T. 2004. "The BNC design as a Model for a Setswana language corpus", *Proceedings of CLUK* '04: 193-198.
Poerwadarminta, W. J. S. 1954. *Kamus Umum Bahasa Indonesia*. Second Edition. Jakarta: Balai Pustaka. [First Edition 1952.]
Poerwadarminta, W. J. S. ed. 1976. *Kamus Umum Bahasa Indonesia*. Fifth Edition. Jakarta: Balai Pustaka. [First Edition 1952.]
Pustet, R. 2004. "Zipf and his heirs", *Language Sciences* 26(1), 1-25.
Read, Allen Walker. 1986. "The history of lexicography", in: Robert Ilson (ed.), *Lexicography; An emerging international profession*, pp. 28-50. Manchester, UK and Dover N.H.: Manchester University Press in association with the Fulbright Commission, London.
Richards, Jack C., and Richard Schmidt. 2002. *Longman dictionary of language teaching and applied linguistics*. Third Edition. London: Longman, Pearson Education. [First Edition 1985.]
Salim, Peter and Yenny Salim. 1991. *Kamus Bahasa Indonesia kontemporer*. Jakarta: Modern English Press.

Schryver, Gilles-Maurice de and D.J. Prinsloo. 2001. "Corpus-based activities versus intuition-based compilations by lexicographers, the Sepedi Lemma-Sign List as a case in point", *Nordic Journal of African Studies* 10(3): 374-398.

Scott, Mike and Christopher Tribble. 2006. *Textual patterns; Key words and corpus analysis in language education*. Philadelphia: John Benjamin. [Studies in Corpus Linguistics 22, Elena Tonigni-Bonelli (ed.).]

Scott, Mike. 2007. *WordSmith Tools 5.0 Help File*. Oxford: Oxford University Press.

Simpson, R. C., S.L. Briggs, J. Ovens, and J.M. Swales. 2002. *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, J. M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press. [Describing English Language Series, John Sinclair and Ronald Carter (eds).]

Sinclair, J. M. and J. Ball. 1995. *Text typology (external criteria); Draft version*. Pisa EAGLES ftp server, Birmingham. [Online address: www.ilc.cnr.it/EAGLES/pub/eagles/corpora/texttyp.ps.gz; accessed 25 Sept 2010.]

Sinclair, J. M. 2003. *Reading concordances; An introduction*. Harlow, Essex and London: Pearson Longman.

Stubbs, Michael. 1996. *Text and corpus analysis; Computer-assisted studies of language and culture*. Oxford UK and Cambridge, Mass.: Blackwell.

Summers, Della. 1991. *Longman/Lancaster English language corpus; Criteria and design. technical report*. Harlow, Essex: Longman.

Teubert, Wolfgang and Anna Cermáková. 2004. "Directions in corpus linguistics", in: M. A. K. Halliday, W. Teubert, C. Yallop and A. Cermáková (eds), *Lexicology and corpus linguistics; An introduction*, pp. 113-165. London and New York: Continuum.

Tim Penyusunan Kamus. 1988. *Kamus Besar Bahasa Indonesia*. Jakarta: Balai Pustaka.

Zgusta, Ladislav. 1971. *Manual of lexicography*. The Hague: Mouton. [Janua Linguarum; Series Maior.]

Zipf, G. K. 1935. *The psycho-biology of language; An introduction to dynamic philology*. Boston: Houghton Mifflin Company.

Zipf, G. K. 1965. *Human behavior and the principle of least effort; An introduction to human ecology*. New York: Hafner. [Facsimile of 1949 Edition.]