# Compiling a Monolingual Dictionary for Native Speakers*

Patrick Hanks, *Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic (patrick.w.hanks@gmail.com)*

**Abstract:** This article gives a survey of the main issues confronting the compilers of monolingual dictionaries in the age of the Internet. Among others, it discusses the relationship between a lexical database and a monolingual dictionary, the role of corpus evidence, historical principles in lexicography vs. synchronic principles, the instability of word meaning, the need for full vocabulary coverage, principles of definition writing, the role of dictionaries in society, and the need for dictionaries to give guidance on matters of disputed word usage. It concludes with some questions about the future of dictionary publishing.

**Keywords:** MONOLINGUAL DICTIONARIES, LEXICAL DATABASE, DICTIONARY STRUCTURE, WORD MEANING, MEANING CHANGE, USAGE, USAGE NOTES, HISTORICAL PRINCIPLES OF LEXICOGRAPHY, SYNCHRONIC PRINCIPLES OF LEXICOGRAPHY, REGISTER, SLANG, STANDARD ENGLISH, VOCABULARY COVERAGE, CONSISTENCY OF SETS, PHRASEOLOGY, SYNTAGMATIC PATTERNS, PROBLEMS OF COMPOSITIONALITY, LINGUISTIC PRESCRIPTIVISM, LEXICAL EVIDENCE

**Opsomming: Die samestelling van 'n eentalige woordeboek vir moedertaalsprekers.** Hierdie artikel gee 'n oorsig van die hoofkwessies waarmee die samestellers van eentalige woordeboeke in die eeu van die Internet te kampe het. Dit bespreek onder andere die verhouding tussen 'n leksikale databasis en 'n eentalige woordeboek, die rol van korpusgetuienis, historiese beginsels vs sinchroniese beginsels in die leksikografie, die onstabiliteit van woordbetekenis, die noodsaak van 'n volledige woordeskatdekking, beginsels van die skryf van definisies, die rol van woordeboeke in die maatskappy, en die noodsaak vir woordeboeke om leiding te gee oor sake van betwiste woordgebruik. Dit sluit af met 'n aantal vrae oor die toekoms van die publikasie van woordeboeke.

**Sleutelwoorde:** EENTALIGE WOORDEBOEKE, LEKSIKALE DATABASIS, WOORDEBOEKSTRUKTUUR, WOORDBETEKENIS, BETEKENISVERANDERING, GEBRUIK, GEBRUIKSAANTEKENINGE, HISTORIESE BEGINSELS VAN DIE LEKSIKOGRAFIE, SINCHRONIESE BEGINSELS VAN DIE LEKSIKOGRAFIE, REGISTER, SLANG, STANDAARDENGELS, WOORDESKATDEKKING, KONSEKWENSIE VAN VERSAMELINGS, FRASEOLOGIE, SINTAGMATIESE PATRONE, PROBLEME VAN KOMPOSISIONALITEIT, LINGUISTIESE PRESKRIPTIVISME, LEKSIKALE GETUIENIS

* This article is an edited version of a plenary address delivered at the conference on 'Dictionaries, More Than Words', which took place at the Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia, 6 February 2009.

## Introduction: dictionary and database

This article gives an account of the English experience in creating monolingual dictionaries aimed primarily at native speakers rather than foreign learners. It starts by comparing the role of a dictionary with that of a lexical database and saying a few words about the issues of register and correctness. Then, briefly, something will be said about words and word histories, lexicographic research, and coverage — what Dwight Bolinger called 'getting the words in'. I shall also discuss dictionary structure — both macrostructure and microstructure.

A lexical database is a fundamental background resource for use in the creation of many important linguistic artefacts — dictionaries, course books, computer programs for natural language processing among them. A great monolingual dictionary has a different function: it brings together speakers of a language, it has a socially integrative function, making explicit the basis of words and meanings and usage, which all uses of the language rely on. Words have meanings — or rather, strictly speaking, they have the *potential* to make meanings when put into context — and they are associated with particular sets of syntagmatic patterns, which can be discovered through painstaking corpus analysis. But words also have register: that is, not all words are equally appropriate in all circumstances. Some words and some grammatical structures are slang, or only used appropriately in spoken contexts, or characteristic of particular regions or dialects; others are only used in formal legal documents, or in romantic fiction or in poetry; others are meaningful and clear, but should not be used at all in polite society. These aspects are implicit in a database, on the basis of the kinds of texts in which each word and use occurs. But a good dictionary reports all of these aspects explicitly. It is not only an inventory of words, their meanings, and their syntagmatic patterns; it is also a report on matters such as register — social attitudes to 'correct usage'. The dictionary is expected to give rulings on what is correct and what is incorrect in different contexts in matters of usage. It is important that such rulings should be based on empirical analysis of the actual usage of good writers, rather than on the preferences and prejudices of a few journalists, academics, and self-appointed pundits, so there is a need here for interaction between a scientifically constructed lexical database and a dictionary as a social artefact.

## Etymology and common sense

What is the role of a dictionary in researching and reporting etymology and word history? Some people think this is the *only* function of a dictionary, but it will be argued that an even more important function is the common sense one of identifying the conventions of word meaning and word use on which members of a language community rely in order to communicate with each other. A distinction must therefore be made between two major kinds of monolingual dictionary for native speakers. On the one hand, traditional major dictionaries

are based on historical principles and report word history and etymology. An example of this kind of dictionary is the multi-volume *Oxford English Dictionary* (OED). The first edition appeared in parts (called 'fascicles') between 1884 and 1928; the second edition was published in 1986; and the whole work is currently being revised at Oxford University Press under the editorship of John Simpson. The revision is available online, so that new research for each word is made available to the scholarly community and the public very soon after it has been completed. The new edition of OED is being made available piecemeal, as the project goes along, without constraints of alphabetical order. In the old days, we had to wait for up to fifty years for publication of lexical research in a big historical dictionary; now we get it within a few weeks of its completion. The OED is a dictionary on historical principles: it places the etymology of the word first and then gives the oldest known meaning of the word after the etymology. Recent developments come last. Word meaning is unstable — it changes quite rapidly — so this means that, in a historical dictionary, the current meaning of many words is placed last or nearly last and is preceded by one or more obsolete, obsolescent, or rare senses. So, for example, a dictionary on historical principles will tell you that a camera is a small vaulted room and next that it is the treasury of the papal curia. Somewhat later on it will tell you that a camera is a darkened room (a camera obscura) at the top of a house, with a hole in the roof, above which is a mirror. The camera obscura reflects images of the surrounding city or countryside on a light table in the room. The importance of this obscure term is that it is the link between our modern word *camera* and the historical meaning, 'small room'. Only right at the end of the entry does a dictionary on historical principles mention that a camera is an apparatus for taking photographs or for making movies.

The other kind of monolingual dictionary is a dictionary on synchronic principles. Basically, a synchronic dictionary reverses the order of senses, placing the modern meaning first. Thus, such a dictionary tells you first that a camera is an apparatus for taking photographs and making movies. It then goes on to explain where the word comes from and how the modern senses developed. The focus in a synchronic dictionary is on reporting conventional meanings and use, rather than on historical and etymological research. Examples of English dictionaries on synchronic principles include the *Encyclopedic World Dictionary* (1971) and *Collins English Dictionary* (1979), both designed and edited by Patrick Hanks, and the *New Oxford Dictionary of English* (1998), which Judy Pearsall and Patrick Hanks designed and edited and whose title was changed (by omitting the word *New*) for the second edition. In America, Houghton Mifflin publishes *The American Heritage Dictionary* (AHD 1969, now in its fourth edition). This owes its origin in the 1960s to the outrage felt by James Parton, publisher of *American Heritage* magazine, at the failure to deal with issues of register and correctness in Merriam Webster's *Third New International Dictionary*, unabridged (1961). Parton decided to commission his own dictionary, and AHD is the result. It is not the prescriptive work that Parton

was expecting, but it is a very good example of a dictionary on synchronic principles. In addition to definitions and examples, AHD contains many short articles on disputed or debatable points of word usage, in which the opinions of over 100 stylistic pundits are compared and collated. The essential principle of AHD is that it is a dictionary on synchronic principles.

The main Australian dictionary, *The Macquarie Dictionary* (1981, now in its fourth edition) is likewise a dictionary on synchronic principles.

What about one-volume dictionaries on historical principles? One might think that, since a one-volume dictionary is a practical tool, there would not be any, but in fact there have been several, and some still survive and thrive. A British example of a dictionary on historical principles was *Chambers Twentieth Century Dictionary,* which lasted for nearly a hundred years until in 1988 the publisher decided, for marketing reasons, to replace it with *Chambers English Dictionary*, and in 1993 by *The Chambers Dictionary*, a work presenting the language on synchronic principles. In 2010, after a long and distinguished history, that publisher went out of business.

In America, rather surprisingly, dictionaries on historical principles are still dominant, even among one-volume dictionaries. America's favourite dictionary — if sales are anything to go by — the *Merriam Webster Collegiate Dictionary*, is a dictionary on historical principles. This work was first published in 1898 and at the time of writing is in its 11th edition. It is very doubtful whether many if any of the purchasers and users of this work are aware of the fundamental difference in principles between this and a more commonsensical synchronic dictionary. It seems possible that some people may even read the dictionary and believe that the 'true meaning' of *camera* is the first one reported — a small room or the Vatican treasury — and that somehow an apparatus for taking photographs is merely an informal late development of low register, to be avoided by careful writers. That would not be an unreasonable inference, given the arrangement of senses, though of course quite wrong. It is not clear how a user of such a dictionary is supposed to divine the modern meaning of a polysemous word, if he/she does not know it already. Current changes in the business model for dictionary publishing raise the question whether there will ever be a 12th Merriam-Webster's Collegiate (or a Fourth Unabridged). Maybe these works will be superseded by continuously updated online versions. I shall return to this question towards the end of this article.

Some words have been stable in meaning since English began; others have changed their meaning more than once. Innumerable examples of lexical meaning change in English could be given. A sock originally meant 'a light shoe'; *dope* was originally a varnish, not a drug; *silly* formerly meant 'happy' and 'of low social class'. Why is the English word *magazine* evidently a cognate of French *magazin*, although they have very different meanings? In English, a magazine is a periodical publication or a part of a gun; in France, the word denotes a department store, where you go shopping. The unifying historical feature is that both go back to an Arabic word meaning 'storehouse'.

The arrangement and presentation of information such as this in a dictionary can be a critical and difficult undertaking. In modern English *size* means bigness, dimension, magnitude — a very fundamental concept. One might imagine that the word *size* has always been part of English, but it has not. It is actually a late medieval development — a rather surprising one — due to the cheating habits of medieval bakers. In the 15th century, if you didn't like the loaf your baker gave you — if you thought it was too small, and if the baker consistently gave small measures — he might be taken to the local assizes, a court of law, and punished for unfair trading. So a size loaf was a loaf that was of a dimension or magnitude approved by the court. And from that narrow basis the word broadened outward to mean the dimension or magnitude of anything. The point is that meaning change in words is unpredictable, but it is a common and substantial feature of linguistic development. Today's exploitation of a word sense may become tomorrow's norm.

It was mentioned earlier that word meaning is unstable and may change often. Some people — including bilingual lexicographers — deny the existence of word meaning altogether. And you can see why: if you think of the word *fire*, what does it mean? Is it something burning out there in the field or the forest, out of control; or is it something burning in your house and nicely under control? Has it something to do with guns? Or has it something to do with losing your job or with making pottery? Perhaps it means inspiring enthusiasm? The answer is that *fire* in isolation means all of those things and much more as well — or rather, it has the potential to have these meanings, when used. For reasons such as this, it can be argued that, strictly speaking, a term in isolation has only meaning potential, not meaning. What you get in a monolingual dictionary is a list of meaning potentials, not of meanings. You need context to know which meaning is activated when a word is used; and in order to know what the normal contexts of words are corpora and corpus analysis is needed. As a source of data and a research technique, introspection does not work. We have learned the hard way, through fifty years of generative linguistics, that introspection does not provide reliable data about how words are really employed in everyday usage. One of the most important findings of corpus linguistics is that people, even trained linguists, are not very good at reporting their own linguistic behaviour. Introspection distorts, perhaps because people consulting their intuitions tend to think up unusual examples — illustrating the boundaries of possibility, rather than normal everyday usage. Normal usage seems to be buried so deep in the subconscious that it is hard for people to recall it to the conscious mind and report it accurately.

## Getting the words in

The first duty of a lexicographer is to get the words in. The editor of a dictionary for native speakers must aim at a very wide inclusion policy, for very often it is the rare and unusual words and senses that people will want to look up.

The editor of a dictionary for foreign learners, on the other hand, will aim to be more selective, presenting and explaining just those words that, in his/her judgement, a foreign learner will need to know.

In 1857, Richard Chenevix Trench, one of the founders of the *Oxford English Dictionary*, described lexicographers as 'the inventory clerks of the language'. This seems exactly right. Compiling inventories may seem a lowly occupation, but in fact, as far as the lexicon is concerned, it is a task full of interesting challenges. It may seem that compiling an inventory of all the words in a language and saying what they mean should be easy, but in fact it is hard, it is difficult. Introspection does not yield an inventory of all the words in one's language, partly because of the difficulty of recall, and partly because nobody knows everything. And then there are problem cases, such as deciding what counts as a word. The core vocabulary of the standard language shades outwards in many directions — into technical jargon, regional dialect, slang, archaic vocabulary, poetic coinages, and so on. Should a dictionary include proper names? Where should a lexicographer draw the line? And when all those decisions have been made, there remain questions about word boundaries. English is not an agglutinating language like Turkish or — some would say — German, where the problems of finding word boundaries are much more serious, but nevertheless deciding on word boundaries in English raises a couple of interesting questions. Just two examples will be discussed.

The problem of phrasal verbs in English is well known. It seems obvious that *take off* — what a plane does when it leaves the ground and starts to fly — is not the same word as the base verb *take*. It has been said that *take off* has as about as much to do with *take* as *disease* has to do with *ease*. Clearly, then, these two verbs should be treated as separate lexical items in the dictionary. But then should a dictionary aim to include all phrasal verbs? What about phrasal verbs such as *finish up*? Should that be an entry in the dictionary too? The standard answer is no, because the meaning is compositional: in this expression, *finish* still means 'finish', and the force of *up* is merely completive-intensive. But then there are phrasal verbs such as *break up* which have both idiosyncratic and completive-intensive meanings:

A gathering such as a political demonstration can break up, or can be broken up by the police; a married couple or two people in a relationship can break up (in which case they are no longer in a relationship); in American English, when people break up, they are overwhelmed by emotion and start laughing or crying; and there are several other meanings for which both the verb *break* and the particle *up* must be present. But then there are meanings where the particle is optional: for example, you can break a table, or you can break a table up. Here, the role of the particle is to reinforce the meaning of *break*, suggesting that the table gets broken into not just two but several pieces.

In a dictionary, should all meanings for such verbs be given, or only the idiosyncratic ones?

Even more problematic are multiword expressions such as *fire engine* and

*fire extinguisher*. A fire engine is a truck carrying equipment for putting out fires. The meaning here is clearly not compositional: *engine* does not mean 'truck'. So *fire engine* should be an entry in the dictionary. The American term, *fire truck*, is more compositional. The meaning of *fire extinguisher* is likewise more or less compositional — it is a piece of equipment for extinguishing fires. But because this is a term that denotes a class of artefacts, each of which is a unique object, *fire extinguisher* is usually selected as a dictionary entry. Now, what about other multiword expressions such as *wood fire* and *forest fire*? There are thousands of such multiword expressions in English, and it is a highly productive area of the language: new multiword terms are being coined all the time. Dictionaries do not do a very good job of reporting them. The usual justification for omitting them is that their meaning is compositional: a wood fire is a fire burning wood; a forest fire is a fire burning a forest. But, although this may seem plausible at first sight, ultimately it is not satisfactory. *Wood* and *forest* are synonyms, so if the meaning were truly compositional, *wood fire* and *forest fire* ought to be synonyms too. But they are not. A wood fire is burning wood (a mass noun) under human control in a hearth in the home or in a camp — but a forest fire is burning a forest or forests (a count noun); it is raging out of control in the wild. The two terms have been conventionalized in different ways, employing different meanings of their basic components. Strictly speaking, a dictionary should explain this. But none do.

Accompanying policy decisions about what counts as a lexical item and where to draw the line is the question, how and where to find the words? During the past 150 years or so, the Oxford Reading Programme has been devoted to reading texts and collecting citations for the words used in them. During its heyday in the late 19th and early 20th century it involved many volunteer readers.

So, before corpora, hundreds of volunteers contributed citations on which the OED was based. Citation collection is a good way of collecting data, especially data for rare and unusual words. But readers have to exercise judgement in deciding what to collect. No one sends in citations for all the uses in a text of ordinary words such as *come* and *go* or *up* and *down*. So, although it was and is a wonderful and admirable enterprise, the Oxford Reading Programme necessarily introduced selective distortion. It did not provide reliable statistical evidence of usage. James Murray, the founding editor of OED, showed that he realized this when he complained in his presidential address to the Philological Society in 1878 (a time when he had only just started on the monumental task of sorting out the citations and writing dictionary entries on the basis of their evidence), 'We have fifty citations for *abusion*, but less than five for *abuse*.'

Citation readers collect citations for unusual words like *triskaidekaphobia* 'irrational fear of the number 13' and for unusual senses. Computers, on the other hand, do not exercise judgement. If asked to find all occurrences of the word *of* in a text or corpus of texts, a computer will do so in a few milliseconds. It can put them in a concordance, otherwise known as a KWIC ('key word in

context') index, which a lexicographer can sample and study and use to compare the patterns associated with each key word. Murray would have seen the point instantly of a large electronic corpus.

Another source of words is existing dictionaries. Lexicography is accretive. One dictionary builds on top of another dictionary. We do not all start from scratch. Dictionary writers are sometimes accused of plagiarizing each other's work, but if you think about it, every dictionary definition can be seen as a small hypothesis. No scientist would publish a hypothesis without consulting the work of his/her predecessors. So it is reasonable to look at other dictionaries and evaluate what they say. What is not reasonable, of course, is mindless copying. That's a danger, for (unlike any other form of research) lexicography requires the lexicographer to say something about everything. Copying is a temptation for the lexicographer. But it is a danger that, in a reputable project, must be resisted. Evaluating definitions in existing dictionaries in the light of new evidence is one thing. Copying out those definitions mindlessly is another.

Thus, there are three main sources of words for a dictionary: citation reading, existing dictionaries, and corpus evidence. Searching corpus data electronically has so far provided only a low yield for new words and new senses, partly because of the difficulty of deciding what counts as a word or sense, and partly because any corpus is only a sample of the language. Searching the Internet may eventually prove more productive, if ways can be found of rigorously defining the existing inventory and defining what counts as a 'lexical item'. As we have seen, identifying uniquely meaningful multiword expressions is a task that poses particular problems. One technique that has been proposed is measuring statistically significant changes in frequency of words and collocations. Thus, twenty years ago, in the infancy of corpus linguistics, Ken Church and Patrick Hanks were able to measure a sudden increase in the frequency of the word *greenhouse* and to note a new pair of associated multiword expressions: *greenhouse effect* and *greenhouse gas*. These terms are now standard entries in monolingual dictionaries.

## Guidance on usage

The needs of users of dictionaries must now be looked at more closely. It has been said that a dictionary has a socially integrating function. This is true up to a point, but only up to a point. There are some good studies of the use of bilingual and foreign learners' dictionaries, but there are no good studies of dictionary use among native speakers, so what follows will necessarily be rather anecdotal, based in part on feedback from marketing departments at Collins in the 1970s and 1980s and at Oxford in the 1990s. Dictionary publishers — who in many cases control lexicographic budgets — often insist that a new dictionary should be 'market driven'. This is rather dangerous, a recipe for extreme conservatism, because until a product has been created, the public — the po-

tential market — has no way of knowing whether it will want the new product or not.

In English, disappointingly for lexicographers who work so hard on definitions and grammar, it seems that people use dictionaries mainly for spelling. This is almost certainly true, at least of dictionaries of the English language, the spelling of which is not phonetic but contains many irregularities and idiosyncrasies. On the other hand, maybe inflections (morphology) and dialect differences are less of a problem in English. Dictionary makers must offer guidance where guidance is needed.

People look to a dictionary for guidance, not only on spelling and inflections, but also on correct usage and word choice. Should we say 'uninterested' or 'disinterested', is there a difference? Nowadays you can hear people saying, 'I totally refute that'. No, you cannot refute a proposition by declaration. You can say, 'I deny that', because *deny* is a performative verb, like *promise*: you can deny or promise something merely by saying so. *Refute*, however, was not a performative verb until recently. To refute a proposition, in the traditional meaning of the word, effective argumentation is needed, not just performance of a speech act. Now, however, it is being used as a strong synonym of *deny*. 'I totally refute that' is politicians' speak for 'I don't want to acknowledge that it's true'. The dictionary should explain that careful writers and speakers still make a distinction in meaning between the two words.

Another example of the sort of guidance that a dictionary should give concerns grammatical complementation. Should one say 'bored with' or 'bored of'? More and more people say 'I'm bored of that'. Is it right or wrong? The usage of an increasingly large number of educated speakers of English cannot be ignored. There is no logical argument against *bored of*. Prepositional choice represents a set of arbitrary conventions. This example contrasts with other common usages, which can be objected to on logical grounds: for example, 'He could of done it' and 'He should of done it', which, though common, are errors. Here, the auxiliary verb *have* is clearly required; it has been replaced by the preposition *of*, which in rapid speech is a homophone of *have*, as a result of grammatical ignorance.

Likewise, 'between you and I', which is even more common, is objectionable on the logical grounds that English prepositions govern object-case pronouns (in no variety of standard English does anyone say, 'He gave it to I' or 'She came home with I'); there is a perfectly good object-case pronoun, *me*. This error is a result of the death of grammatical case in English (except for a few pronouns) coupled with hypercorrection. What has happened is this: schoolchildren are taught that it is wrong to say 'me and' in subject position (as in 'Me and my friends are going on holiday'), but they do not fully understand the nature of the grammatical error, so, with hypercorrection, they use 'and I' in place of 'me and' on all occasions, regardless of case. Thus, 'Between you and I' has become established as a formula among people who have no sensitivity to grammatical case. The question is, should a dictionary acknowledge such errors as part of English, and if so, what should it say?

Another example concerns the so-called split infinitive. In English, for at least three hundred years, there has been a lively debate — especially among people who believe that English is really Latin in disguise — about whether it is acceptable to 'split' an infinitive by putting an adverb between the infinitive marker and the verb. Can you say 'to boldly go [where no man has been before]'? There is no logical objection to this, but conservative self-appointed pundits object to it. The dictionary should give a ruling.

## Lexical and paralexical content

It is fashionable in some academic circles to make a distinction between 'lexical semantics' and 'encyclopedic information'. But ordinary monolingual dictionary users do not make the same careful distinction. People want instant cultural reference information and do not care whether it is classified as 'semantic' or 'encyclopedic'. Here are examples of the sort of questions in English that people ask and expect to have answered by a dictionary:

— 'What's the scientific name for a thrush?'
— 'Is your scapula your shoulder blade, your backbone, or your collarbone?'
— 'What's the capital of Chile?'
— 'Why is a madrigal called a madrigal?'
— 'What does *nook-shotten* mean?' (We find Shakespeare talking about England as a nook-shotten island. What does it mean?)
— 'What is a predator?' people might ask. 'Is a penguin a predator?' [Well, they catch fish, don't they?]
— 'What are chinos?'
— 'What's an ohm? What's a joule, and why is it called a joule?'
— 'Is *aa* an English word?'

People use their dictionaries for Scrabble and for crossword puzzles. As a matter of fact, *aa is* an English word: it is a kind of volcanic lava, a word of Hawaiian origin — not very common in everyday reading and conversation, but remarkably useful for Scrabble.

In addition, people want to have a dictionary as an authoritative inventory of their language, even if the dictionary sits on their bookshelves and they never look inside it. They want it there on the bookshelf just in case they might one day want to look something up.

If they do look inside a dictionary, sometimes people just want to browse. So they also want fun words. For example, here are some words denoting criminals of various kinds from different periods and different sources: a cutpurse (a street thief; used by Shakespeare), a mosstrooper (mosstroopers were criminal raiders on the borders between England and Scotland in the 16th and 17th centuries). What's a yegg? What's a snakehead? What's a tsotsi? What's a

rudeboy? What's a grifter? These are all different kinds of criminals in different parts of the English-speaking world. And above all, the marketing department will tell you, 'We want *new* words, because then the journalists will write about our dictionary.'

## Corpus evidence and examples

A corpus shows how each word is used. It does not show directly what each word means, but it provides evidence on the basis of which meanings can be inferred. The first editions of the two large monolingual dictionaries of British English (Hamlyn 1971 and Collins 1979) were designed and edited before corpus evidence became available. Then, in 1983, in the earliest days of corpus lexicography, John Sinclair and others started working on the first edition of the COBUILD dictionary. They discovered that many of the generalizations made in pre-corpus dictionaries, though plausible, were not quite right. That is, they did not stand up well to comparison with corpus evidence. The *New Oxford Dictionary of English* (1998), mentioned earlier, was the first (and, so far, the only) dictionary for native speakers of English to be based on corpus evidence and well as citations from a reading programme.

Corpus evidence provides an essential source of information for collocations and syntagmatics, which need to be studied statistically in order to understand the relationship between word use and word meaning. This provides a structure or framework of a dictionary. Patterns of word use can be detected in corpora, but these patterns provide hints, associations, and probabilities about meaning and usage, rather than certainties. They point, in fact, beyond lexicography to a need for new lexically based approaches to linguistic theory.

An essential design feature of a natural language is that it is full of uncertainty. This is because the categories found in natural languages are built around prototypical 'best examples' and have boundaries that are fuzzy, rather than being sharply defined. This, of course, presents a big problem for lexicographers. To take a very simple example, corpus evidence shows that the prototypical use of the verb *hazard* in English is with the noun *guess* as a direct object. The phraseology is 'hazard a guess' in more than 50% of the uses of this verb in all the corpora consulted. The meaning of the phrase as a whole is, 'to say something without much confidence that it is true'. The prototypical direct object, *guess*, is a noun denoting a speech act or a concept. On closer inspection of the corpus (for example, the British National Corpus), all sort of other speech-act and concept nouns are found in the same slot — not only the near synonym *conjecture*, but also *inference, opinion*, and even *definition*. There are even some examples of the verb governing indirect and direct speech, as in 1 and 2. These are boundary cases, ungrammatical in most varieties of American English.

1. I would **hazard** that the ratio of real balances to total private sector net worth is less than 1% ...
2. "My uncle," said Wendy, expanding further on her family, "was Provost of Dumfries; he had a rather odd name — 'Chicken'." "Not Hen Chicken?" I **hazarded**, as this humorous diminutive was part of my family mythology.

The dilemma for the lexicographer in such cases is whether to represent and gloss the prototypical example (in this case, *hazard a guess*) or whether to set such a broad scope (*hazard something*) that the normal phraseology and its meaning are in danger of being lost sight of. There is no single correct solution to this dilemma. It is a matter of judgement and choice, taking account of the likely needs of the intended users. A useful compromise involves using the word 'typically', for example by defining the first sense of *hazard* as 'to state a proposition, typically a guess, without any great confidence that it is true'.

If a broader scope is chosen for the definition of this sense, typical phraseology can be highlighted by means of an example. For this and other reasons, the dictionary maker should resist the temptation to choose weird, inventive, creative, unusual boundary-case examples, and instead choose examples that represent central and typical, normal usage, even though such usage may seem slightly boring. The objective in selecting examples should be to illustrate normal usage, not to illustrate the boundaries of all imaginable possibilities. Unfortunately, for some reason there is a strong human urge to focus on boundary cases and unusual usage, so young lexicographers have to be *trained* to select examples that are normal, even boring.

## Interpreting the evidence

The example of *hazard*, verb, shows how corpus evidence can augment and even supplant intuitions. The first thought of many English speakers, consulting their intuitions without the benefit of objective evidence, is that this verb means 'to put at risk'. It certainly does have this meaning, but only in about 20% of all uses, if the corpus evidence is to be believed. The corpus nudges the lexicographer into recognizing facts of the language that are not intuitively obvious.

On the other hand, an example of how *not* to use evidence and examples in a dictionary may be given from Wiktionary; in the monolingual English version, the verb *hazard* is defined as:

1. To expose to chance; to take a risk: *I'll hazard a guess.*
2. To incur or venture.

These definitions appear to have been copied, with minor alterations, from another dictionary, with no thought as to how the word is actually used. An example has been tacked on to sense 1, although it actually illustrates sense 2. This is not an isolated error; indeed, it is fairly typical of the monolingual English Wiktionary.

There is a dramatic contrast between Wikipedia and Wiktionary. Wikipedia has been a great success. It is a vast anthology of encyclopedic articles written by people claiming to have knowledge about a particular subject. If that claim turns out to be ill-founded, i.e. if an article turns out to be erroneous, then, if it is of any public interest at all, it is pounced on by genuinely knowledgeable people and improved or replaced. It seems that the model of an encyclopedia as a collectively written anthology is a good one. This model, however, cannot be extended to a dictionary. A dictionary is not an anthology. Wiktionary is full of second-hand derivative entries, often wrongly defined or with erroneous examples. For reliable information about the words and meanings of a language, the evidence of actual usage — corpus or citations — must be interpreted by knowledgeable and trained people following a set of consistent principles.

## Writing definitions

As indicated in the preceding section, the first priority for a monolingual lexicographer is to give shape to each dictionary entry, writing definitions that reflect the evidence by selecting a middle course somewhere between accounting only for prototypical uses of a word and accounting for all imaginable uses.

The next priority is to achieve technical accuracy in definitions. Writing definitions of technical terms is a particular problem for monolingual lexicographers. In order to understand and explain a term — and definitions should aim to *explain*, not merely to define — the definer needs to be a user of the terms being explained. This applies not only to scientific definitions but also to other domains such as sports. Anybody who has ever played cricket knows how badly worded American dictionary definitions of cricketing terms can be. For example, in one recent American dictionary, we are told that the bowler in cricket is 'the player who throws the ball to the batsman' — there is no mention of the obligatory straight arm that distinguishes bowling from throwing by fielders and from pitching in baseball. The distinction is important because the sporting pages of English-language newspapers outside North America quite often contain sentences such as:

> Brett Lee, Australia's answer to Shoaib Akhtar, is the latest fast bowler to be accused of throwing. — Simon Briggs, *Daily Telegraph,* London, 22 February, 2009.

A reader trying to interpret this sentence and consulting the American definition of *bowler* just mentioned, would be puzzled rather than enlightened.

No doubt Americans find the same kind of bad wording in British definitions of baseball terms, or American football, which they call football, or hockey, which we call ice hockey. Definers need to be users of the terms being defined to appreciate the importance of technically accurate components of

meaning. But expert users of terms who are not trained lexicographers are often particularly bad at defining them and explaining them, so there needs to be interaction between the technical adviser — the scientist or the sportsman, as the case may be — and the lexicographer, who has the skill of defining and succinctly explaining.

In some words there is a clash between the meanings used by the scientific community and the meanings of ordinary people. Ordinary people, when they say 'wait a second', do not mean 'wait the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom'. This, however, is the definition of *second* as the basic scientific unit of time, agreed by the General Conference on Weights and Measures, which meets from time to time in Paris, as part of the Système International d'Unités (SI units). A serious dictionary must give both the scientific definition and explain the ordinary language usage.

A similar problem arises in defining many everyday creatures and other objects, e.g. *spider*. Consider the following extract from the *Oxford Dictionary of English*:

> an eight-legged predatory arachnid with an unsegmented body consisting of a fused head and thorax and a rounded abdomen. Spiders have fangs which inject poison into their prey, and most kinds spin webs in which to capture insects. Order Araneae, class Arachnida.

The first part of this entry aims to define — to set boundaries around a classification — rather than to explain. Why mention 'an unsegmented body' and 'a fused head and thorax'? These features are mentioned in order to distinguish spiders from insects, which form a completely different zoological class. It is the second sentence in this entry — which is not part of the formal definition — that goes some way towards explaining. Now contrast this with the COBUILD entry:

> A spider is a small creature with eight legs that looks like an insect. Most types of spider make webs in which they catch insects for food.

This entry is clearly more concerned with explaining matters to foreign learners than with scientific definition. Definitions in monolingual dictionaries aim to do both.

Using a corpus such as the British National Corpus and with the help of a corpus analysis tool such as the Sketch Engine, it is possible to compile a corpus-based linguistic profile of terms such as *spider*.

— Many thousands of species of spiders are known (*funnel-web, web-building, orb-weaving, bird-eating, ground-dwelling, giant, huge, large, tiny, poisonous, black widow, camel, redback, trapdoor, wolf, whitetail, crab, tarantula*, etc.).
— Some species of spiders *hunt* prey.

— Spiders *bite*.
— Some species of spiders are *poisonous*.
— Many species of spiders *spin webs*, with threads of *strong silk*.
— Spiders *lurk* in the centre of their *webs*.
— Spiders *control* what is going on in their *webs*.
— Spiders have eight *legs*.
— Their legs are *thin*, hairy, and long in proportion to body size.
— Spiders have *eight* eyes.
— Spiders spend a lot of time being *motionless*.
— Spiders' *movement* is *sudden*.
— Spiders *crawl*.
— Spiders *scuttle*.
— Spiders are *swift* and *agile*.
— Spiders can *run up walls*.
— Many people have a *dread* (*hate, fear*) of spiders.
— People *kill* spiders.
— English people are much concerned with trying to get spiders out of the *bath*.

Such a profile summarizes the beliefs that most people, at least in England, have about spiders. (With regard to the last point in this list, it should be added that, although there is a mildly significant association between the lexical items *spider* and *bath* in some British corpora, it is not suggested that this is a serious fact requiring scientific investigation.)

## Consistency of sets

Once the framework for each ordinary word has been created using corpus evidence, other kinds of information must be slotted in. The principle of coverage of terms in all fields of human activity, including sports, leads to another principle of monolingual lexicography, namely consistency of sets. All the terms in a set — the chemical elements, for example, or the organs of the human body — should be defined in a similar and consistent style, regardless of frequency, i.e. even though some members of the set may be so rare that they do not show up at all in a corpus. The same principle holds good for the terminology of activities such as snooker, curling, and Australian rules football — all of which have sprung into prominence only in recent years, disseminated mainly by television. If people watching snooker on TV hear the commentator say, 'There is a possible plant here', they may well turn to the dictionary for a relevant definition of *plant*. *The Oxford Dictionary of English* defines this as 'a shot in which the cue ball is made to strike one of two touching or nearly touching balls with the result that the second is potted'. This in turn implies that there must be an adequate definition of the snooker sense of *cue ball* and the verb *to pot*.

Thus, when a lexicographer reads a newspaper or watches TV, it is often the case that the lexicographer is less interested in the content of what is being said than in how it is being said. What are the words being used? Do they need to be dictionary entries?

The editor of a monolingual dictionary has to decide how far to go in technical fields. Should *strobilus*, *strobila*, *strobilation* be entries? Native speakers who do not know these words may expect to find them in a dictionary. On the other hand, a dictionary is not a termbank. The terminology of the sciences in particular but also of technological activities is so vast and specialized that much of it does not belong in a dictionary, but rather in a project like IATE (Inter-Active Terminology for Europe), the 23-language terminology database of the European Union, which collects and stores technical terminology, much of it of an extremely abstruse variety, with stipulative definitions.

What all this adds up to is that native-speaker dictionary users expect the inventory of words in a dictionary to be complete, and the lexicographer must find ways of satisfying that expectation, despite the fact that the goal is impossible: the lexicon of a living language is dynamic and the boundaries of its vocabulary are fuzzy, so that new words and expressions are being coined — invented or borrowed — all the time.

## New words

Most dictionary publishers issue, with each new edition of a major dictionary, a booklet of new words to excite the journalists. The *Macmillan English Dictionary for Advanced Learners* (MEDAL) is no exception. It is actually a dictionary for foreign learners, but the publishers well understood the virtues of publicity, so in 2008 they issued a small, free booklet that included such 'new words' as *blogosphere*, *chav*, *air kiss*, *career gapper* (somebody who's taken time out from their career), *Chelsea tractor* (one of those tank-like vehicles supposedly driven especially by people who live in Chelsea, a somewhat rich district of London). *Chick lit* is a genre of literature written for young women to read, typically while sunning themselves on the beach; *civil partnership* usually denotes a homosexual marriage. A *designer baby* — it is amazing what you can do with genetics these days.

## The need to get on with it

Compiling an authoritative monolingual dictionary for native speakers is a daunting task. A natural language consists of thousands of lexical items, ranging from extremely common items such as function words and light verbs to rare technical terminology and compounds. The editor-in-chief must develop clear policies for all of the issues mentioned in this article, together with many others, and ensure that they are followed consistently by other contributors

working together, not merely as a team, but as a 'single collective author'. The whole policy cannot be established before actual entry writing begins. Rather, broad outlines are established at the outset, which are then extended and modified as the project goes along, in response to particular issue that come up.

Conscientious definition writers tend to agonize over capturing the precise meaning of each word. But agonizing is counter-productive. It sometimes happens that the first form of words that a lexicographer jots down is a perfectly good one; they then agonize and gradually make the entry less and less satisfactory and less and less comprehensible, typically by trying to cover all eventualities. Fear of making a mistake is another factor that slows lexicographers down without bringing any noticeable benefit. The sad fact is that slow writers who agonize make just as many mistakes as quick writers working according to a good plan. The plan needs to be outlined very broadly at first, then developed as the project starts up. The details cannot be developed satisfactorily in advance, in an abstract theoretical vacuum.

This problem is best dealt with by setting up a system for a dictionary project where each compiler is free to do his/her honest best and move quickly on. The system says, 'Don't worry about making a mistake; somebody else will check what you have written.' The editor-in-chief will check and ensure that obvious errors and accidental infelicities are corrected, and give feedback. Lexicography is a team game; the team should have a structure; lexicographers should read and check each other's work in a co-operative environment.

## The medium

In the modern world, the question arises, in what medium will any new monolingual dictionary be published? The traditional medium of a bound book for reference information is being superseded by the Internet. This raises major questions for the future of lexicography, including the following:

— Is the Internet as secure and durable a medium as the printed page? Will future readers, in five hundred years time, be able to consult an electronic dictionary on the Internet in the same way that a present-day reader can consult an old book in a library?

— Can a new dictionary any longer be a product created within the capitalist system as an investment by a publisher or risk capitalist? Or must all new dictionaries be funded by central, government-controlled funding agencies, as they were in the former totalitarian states of Eastern Europe?

— If funding is to come from a commercial investor such as a publisher, what is the business model? People have got into the habit of expecting information to be freely available and contributed by volunteers, on the model of Wikipedia, but, as discussed above, this is not a satisfactory

model for a dictionary. Is it realistic to expect sufficient revenue to accrue from advertising to justify the huge investment required to fund the creation of a good new dictionary, or can the habits and expectations of online users be changed, so that they will pay for the information they obtain?

## Conclusion: evidence and interpretation

A large modern monolingual dictionary of any language has an important role to play in the community, and language communities, large and small, need their dictionaries. An essential requirement is a lexical database, constructed by analysis of corpus evidence, but also reflecting social attitudes to language. A corpus shows patterns of word usage. Supplementary research is also needed for terminology, unusual words, names, word histories, attitudes to correctness, and other matters — but such research is all designed to find evidence, not to promote the opinions of self-appointed pundits. At the core of lexicography, therefore, lies the corpus. The basic task is to report all normal uses and meanings of all normal words. But a dictionary must also reflect social attitudes to language and give guidance on meanings and etymology, but to be authoritative, all pronouncements must be based on evidence of usage — corpus evidence. Public attitudes to points of 'correct' usage should be reported and evaluated. If a dictionary tries to cover all imaginable possibilities of use of any content word in the language, there is a danger of lapsing into incoherence. The language will defeat the over-ambitious lexicographer. This is because word meaning and use is infinitely flexible. What a dictionary is reaching for is the central norm that speakers of a language rely on when they speak to each other, not the wild imaginings of linguists dreaming up remote possibilities.

## Acknowledgement

This work was funded in part by the Academy of Sciences of the Czech Republic (project T100300419) and the Czech Ministry of Education (National Research Program II project 2C06009) as part of a corpus-driven investigation of lexical issues at the Institute of Formal and Applied Linguistics of the Charles University in Prague.

## References

### Dictionaries

**Delbridge, A. et al. (Eds.).** 1981. *The Macquarie Dictionary*. Sydney: The Macquarie Library.

**Gove, P.B. (Ed.).** 1961. *Webster's Third New International Dictionary of the English Language*. Unabridged. Springfield, Mass.: Merriam Webster.

**Hanks, P. (Ed.).** 1971. *Encyclopedic World Dictionary*. London: Hamlyn.

**Hanks, P. (Ed.).** 1979. *Collins Dictionary of the English Language*. London: Collins.

**Morris, W. (Ed.).** 1969. *The American Heritage Dictionary of the English Language*. Boston: Hougton Mifflin.

**Mish, F. (Ed.).** 2004. *Webster's Eleventh Collegiate Dictionary*. Springfield, Mass.: Merriam Webster.

**Pearsall, J. and P. Hanks (Eds.).** 1998. *New Oxford Dictionary of English*. Oxford: Oxford University Press.

**Rundell, M. and G. Fox (Eds.).** 2002. *Macmillan English Dictionary for Advanced Learners*. London: Macmillan.

**Sinclair, J. (Ed.-in-Chief).** 1987. *Collins COBUILD English Language Dictionary*. London/Glasgow: Collins.

## Other Literature

**Bolinger, Dwight.** 1970. Getting the "Words" In. *American Speech* 45(1–2): 78-84.

**Trench, Richard Chenevix.** 1857. On Some Deficiencies in Our English Dictionaries. Hartmann, R.R.K. (Ed.). 2003. *Lexicography: Critical Concepts*: 171-216. London/New York: Routledge.