The treatment of multiword lexemes in some current dictionaries of English

Edward Gates

Conventionalized phrases, clauses and sentences make up a considerable part of the English lexicon and merit more adequate treatment than has been given them in existing dictionaries. In this paper, I examine the treatment given lexemes composed of more than one printed word in six recent large desk dictionaries, three British and three American: THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE, Second College Edition; CHAMBERS 20TH CENTURY DICTIONARY, New Edition; COLLINS ENGLISH DICTIONARY; LONGMAN DICTIONARY OF THE ENGLISH LANGUAGE; WEBSTER'S NINTH NEW COLLEGIATE DICTIONARY; and WEBSTER'S NEW WORLD DICTIONARY, Second College Edition. I refer to these as AHD2, CTCD, CED, LDEL, W9, and WNWD.

1. The inclusion of multiword lexemes

How does the inclusion of multiword lexemes in these desk dictionaries compare with their inclusion of single words? In most of them the ratio of multiword entries is lower than in the large dictionaries of record. In a sample composed of the first 500 entries in the letter R, multiword lexemes represented 25% in the OXFORD ENGLISH DICTIONARY and 35% in WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY. Among the desk dictionaries in this study, CTCD has the highest ratio: 33.5%. CED has 20.5%; LDEL, 19%; WNWD, 17.5%, AHD2, 17%; and W9, 16%. W9 is abridged from WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY and contains 35% of its total entries but only about 18% of its multiword entries. The smaller ratio in the desk dictionaries may reflect an unconscious feeling by the lexicographer that a dictionary is a book that explains words, and that vocabulary items larger than the word are beyond its scope, or at least of marginal importance. The dictionary makers may also have supposed, without really thinking much about it, that people could figure out the meaning of these combinations from their components. I recall my own feeling as an abridger that these familiar expressions presented no problem of understanding and thus were dispensable. Unlike eighteenth century dictionaries, which excluded nearly all multiword lexemes, later twentieth century dictionaries in principle include a selection.

What factors governed the selection of multiword lexemes in the dictionaries of this study? Aside from the usual principles of currency, frequency, and general use, the lexicographers seem to have considered whether a multiword lexeme was an essential part of the vocabulary and whether it could be understood as the sum of its components. Practice seems also to have been affected by less conscious factors.

The dictionaries do not, in general, include merely customary strings of words with no idiomatic features of form or meaning. Although CED enters *take no for an answer,* none enter *in other words* or *make good use of.* However, they do enter certain kinds of transparent collocation. Compounds that are the usual names for things are entered; all six dictionaries have *color scheme, creature comfort, race track,* and *rocking chair.* Similarly, terminological phrases are included; all but CTCD enter the linguistic term *immediate constituent,* the legal term *right of search,* and the medical term *radium therapy.* Familiar hyphenated compounds tend to be entered, but not consistently; *high-pitched, knee-high, long-range, pitch-black,* and *snow-white* are in all six dictionaries, but *low-pitched* is not in W9, *short-range* is not in AHD2, and *waist-high* is not in AHD2, LDEL, or WNWD. *Low-cost* is only in CTCD and AHD2, and *terror-stricken* only in CED.

In principle, the dictionaries do include collocations that pose problems of understanding because of some anomalous or unique feature of form or meaning. Those with grammatical anomaly are not consistently included. The phrases *in the know* and *at random,* with a verb and an adjective as objects of a preposition, are in the dictionaries. However, the subjunctive relic *come what may* and the ungrammatical *as best one can* are in none. Inclusion of lexically redundant phrases also varies; *hem and haw* is in all, but only CED includes *in this day and age*; CED and WNWD have *lo and behold.*

Multiword lexemes composed of words unique to the collocation, such as *spick and span* and *ad hominem* are entered. So are collocations with a phrase meaning that cannot be inferred from its components, like *right away* (in all but CTCD) and *bats in the belfry* (in CTCD, CED, and WNWD).

A kind of multiword lexeme not consistently included in these dictionaries is the polite form. Perhaps, in spite of their frequency, it seem superfluous to enter expressions so familiar to native speakers. Although *How do you do?* is an entry in all but AHD2 and W9, *thank you* is only in CTCD and WNWD, though covered by a note in LDEL and W9.

Signals interjected into writing and speech are also not fully included, again perhaps because their familiarity obscures their semantic anomaly; *that is to say* is not in AHD2 or CTCD and *you know* is not in CED or LDEL.

Idiomatic phrases composed wholly of familiar function words tend to be overlooked. I found *and how* missing from AHD2 and WNWD, *in on* missing from W9, and *as is* from CTCD; but *and all, as for,* and *in for* were in all six.

If there has been an unconscious resistance to including phrases and clauses, the resistance has been even stronger to including sentences and other indepen-

dent utterances. Though these are not a large proportion of the multiword items in the inventory of English, there are a surprising number of them. I have compiled a list of more than a hundred, and this is by no means complete. None of the dictionaries cover them adequately. *No dice!* and *You bet!* are in all but AHD2 and W9, though covered in the latter by a note. *The fat's in the fire* is only in CTCD, CED, and WNWD. *One's eyes are bigger than one's stomach* is only in CED. *Go fly a kite* and *Has the cat got your tongue?* are in none.

A subclass of sentences is the proverb. Since there are special reference works to explain these, the general dictionary maker might reasonably exclude them. However, some non-transparent proverbs are in the dictionaries under investigation. WNWD enters the most; W9 has none.

Many conventionalized phrases and clauses are made up of combinations in which one word or meaning of a word occurs uniquely or usually in a particular collocation. In principle a lexicographer could treat these as contextually bound uses of a single word. CED states as policy that words and senses which occur only or usually in fixed collocations, such as *kith* in the phrase *kith and kin,* are entered and defined as words and the constraint noted (p. xvii). However, the distinction between word anomaly and phrase anomaly is often not clearcut. Moreover, in compiling reference works, the convenience of users may override theoretical considerations. In deciding whether to treat these as multiword or single-word entries, the dictionaries of this study had few if any systematic guidelines, judging from their lack of consistency. Decisions seem to have been made by individual definers, item by item. However, on some kinds of phrase they generally agree.

Some kinds are usually treated as multiword entries. Collocations forming names, as in the case of transparent compounds, are entered; e.g. *runcible spoon,* containing the unique form *runcible.* Other conventional collocations containing unique forms are often multiword entries, e.g. *taken aback* and *in cahoots.*

The dictionaries usually agree to treat as contextually conditioned uses, rather than as multiword entries, some other kinds of phrases. The use of a word with a unique meaning in the phrase rather than a unique form is defined along with other meanings of the word. The constraint on usage in the phrase may be indicated in a note or merely in an example; e.g., the use of *naked* in the familiar expression *naked eye* is treated as a sense of *naked,* "unaided by any optical instrument," in all the dictionaries except CTCD. The same is true when one meaning of a word, although not unique to a particular collocation, is most often found in it; e.g., *high noon.*

Words used in pairs (e.g. *as . . . as, either . . . or, the . . . the*) to introduce parts of correlative constructions might be considered for phrase treatment, which would be easier for users to find, but except for *as . . . as* in CTCD, all are treated as a sense of the word in the dictionaries of this study. Also given single-word treatment are such miscellaneously anomalous collocations as *a few, a lot, a good* (or *great*) *deal, about to* (used with the infinitive), *many a* (with noun), and *up and* (as in "He up and did it.").

Some of the decisions to treat a collocation as a single sense were clearly wrong, because a meaning of the whole collocation, not of a single word, was involved. W9 in particular displays a tendency to this. The idiom *one's neck of the woods* is covered in W9 by a sense of *neck*: "REGION, PART;" and the idiom *grist for one's mill* by a sense of *grist*: "something turned to advantage."

A class of lexical units that can perplex a dictionary maker is made up of construction patterns rather than conventionally fixed wording; e.g., the set expressing distribution in small amounts: *bit by bit, inch by inch, two by two,* etc. It does not seem feasible for a dictionary to attempt to cover most of them, since they have minimal fixed lexical content. Nevertheless, the dictionaries examined do inconsistently include a few members of a few of the sets. All the dictionaries have *again and again*; all but CED have *more and more*; but only LDEL and W9 have *less and less. Day after day* is an entry in all but CTCD. LDEL and WNWD also enter the related *week after week* and *year after year*, but only WNWD has *month after month.*

How do the dictionaries compare in their coverage? For coverage of multiword lexemes other than compounds and phrasal verbs, CTCD is the most useful and W9 the least. According to my samples, idiomatic expressions comprise 5% of the entries in CTCD; in W9, only 1%.

2. Place of entry

Should a multiword lexeme be a main entry or a subentry? Five of the six dictionaries include some kinds of multiword lexemes as main entries, and the others as subentries. The exception is CTCD, which runs on all lexemes except basic forms. In the other dictionaries main entry is given to compound nouns and adjectives, noun phrases like *rule of thumb*, and hyphenated verbs like *rubber-stamp.* Also main entries are foreign phrases like *ad hoc* and *raison d'être,* regardless of their grammatical function, perhaps because there is seldom a single-word entry at which they could be run on.

Beyond this, policies differ. In all but CTCD and LDEL, compound conjunctions like *inasmuch as* are main entries. LDEL and W9 enter phrases like *rank and file* and *out-and-out*; AHD2, WNWD also does if these are hyphenated. In CED, phrasal verbs are main entries. In W9, phrasal verbs with adverbs are main entries, but those with prepositions are run-ons – a distinction that surely eludes most users. W9 also enters at their own alphabetical place compound conjunctions of the type *as far as* and compound prepositions and adverbs like *as to* and *at all.*

At which word in the multiword lexeme should a subentry appear? All the dictionaries of this study subenter lexemes having variable wording at the first major invariable word; e.g., *go* (or *run*) *to seed* is run on at *seed.* Otherwise their policies differ. W9 places run-on entries under what it calls the "major element,"

interpreted as a noun or verb, e.g. *in spite of* at *spite*. When there is no noun or verb, the phrase is run on at the first word, e.g. *and so forth* at *and*. The policy of LDEL is detailed. Invariant expressions are entered at the first noun if there is one; if not, at the first adjective; similarly at the first adverb or verb; if none of these occur, at the first word. The policy of WNWD is to enter idiomatic phrases "wherever possible under the key word" (p. xiv). This rather subjective criterion produces no predictable choice for users. The expressions *chew the fat, curry favor,* and *scratch the surface* are found at the verb, but *bite the dust, break one's heart,* and eight others that I checked are at the noun. However, this policy does allow treatment of closely related sets like *bring to pass* and *come to pass* at one entry. AHD2, CTCD, and CED also enter run-ons at the most significant word and display a similar variety of location; some phrases are entered in two places, particularly in CTCD.

The location of run-on phrases containing only function words is not consistent, even in the same dictionary. Some but not all phrases beginning with the conjunctions *and* and *as,* and with the prepositions *in* and *of* are run on at those entries in all the dictionaries, but they are not the same phrases in different dictionaries.

Location is also a problem when the word at which the multiword lexeme would be entered is unique to the lexeme. CTCD runs on *Achilles heel* at the related word *Achillean.* LDEL enters *inasmuch* in order to run on *inasmuch as.*

Some dictionaries subenter one multiword lexeme at a main entry for another. CED and W9 enter *run away with* at *run away.* AHD2, CED, and WNWD enter *penny-pinching* after *penny-pincher,* while W9 and LDEL do the reverse. None of the dictionaries subenter these derivatives at the base form *pinch pennies,* or even provide a cross reference.

The lack of a clear policy on the location of multiword lexemes compounds the problem that users have in deciding where to look in a dictionary for help in understanding an obscure sentence. Not only may they not know which, if any, of the words is being used in an unfamiliar way, but even if they identify an anomalous collocation, it may not be explained at the first word they look up. To help the user, cross references can be provided from the other major words to the word where the entry is found. None of the American dictionaries had this index feature, but all the British dictionaries did. LDEL has a thorough system of indexing major components. At *bite*, there is a note "see also *bite the DUST*" in which the word *dust* is capitalized, indicating to the user where to look.

Dictionaries differ not only on the place in the dictionary where a multiword lexeme is explained but also on the place within the entry. In CED the lexeme may follow a related sense of the word, as the next numbered definition. Otherwise, CED follows the same plan as AHD2, LDEL, and W9, in which run-ons follow the entry or part of the entry containing the senses that belong to the same part of speech. In CTCD and WNWD, all run-ons are found after all the senses for different parts of speech. Some dictionaries have subsections for different kinds of multiword lexeme.

3. Form of the lemma

Problems arise from variation in the wording of multiword lexemes. When only one word varies, five of the dictionaries show one single word variant after the other; e.g., *chew the rag* (or *fat*). Their format for separating the two differs. W9, however, prints out the entire phrase twice: *chew the rag* or *chew the fat.* WNWD sometimes puts variants at the end of the entry, after the senses; e.g. at *bend over backward* is the note "also *lean over backward.*" When variant words are numerous, the definer may resort to *etc.*, as WNWD does at *up to the ears* (where LDEL has *up to one's armpits/ears/eyes/eyebrows/neck*).

Some variations involve additional words. Additions that do not come at the beginning, where they affect alphabetization, are often shown in parentheses, like alternatives. They can also be indicated by a note; e.g. CED, at the subentry *over and over,* adds "often followed by *again.*"

Another problem is what form, if any, to put in the lemma for a variable possessor. W9 and WNWD substitute *one's*, as in *break one's heart.* The other dictionaries observe a distinction between a possessor referring to the subject of the sentence, for which *one's* is substituted, and a possessor not the subject, for which *someone's* or *somebody's* is substituted. Thus they enter *make up one's mind* but *break somebody's heart.*

AHD2 and WNWD seem occasionally to observe the same distinction for variable personal objects; e.g., *give (someone) the eye.*

4. Kinds of information

What kinds of information need to be given about multiword lexemes? Pronunciations are needed only for words in the entry that are not given a pronunciation elsewhere, such as *raree* in *raree show* and foreign phrases. However, pronunciations are given by the dictionaries of this study for some other multiword lexemes, on the basis of their written form. Although the components of a compound are the same whether they are written as a single printed word, joined by a hyphen, or separated by a space, these graphic differences determine whether a pronunciation is given. None of the dictionaries in this study give pronunciations for compounds when they are main entries written without hyphenation. AHD2, W9, and WNWD given hyphenated compounds pronunciation on the same basis as single words. LDEL and CTCD (where they are subentries) indicate only stress. Because compounds have different stress patterns, it would be useful for dictionaries to indicate stress on all compounds, but none in this study do. The user cannot tell that only *white* has primary stress in *white sauce*, but both words do in *white dwarf.*

Though nearly all multiword lexemes have grammatical functions corresponding to those of single words, dictionaries do not provide all with part-of-speech

labels. AHD2, CED, and LDEL label all main entries. W9 labels main entry compounds but not phrases like *bed of roses*. WNWD labels hyphenated main entries, which by policy are compound adjectives and verbs. CTCD enters multiword lexemes only as subentries and gives part-of-speech labels only sporadically. AHD2 uses as a subentry section heading the label "Phrasal verbs." CED sometimes labels subentries; e.g., *all-out* and *of course* are labeled as adverbial, and *like hell* is labeled "(adv.) (intensifier)."

Should a literal definition be given for multiword lexemes with figurative meanings? Occasionally this may be useful to some users, as for *go* (or *run*) *to seed,* and this is given in all six dictionaries. It seems unnecessary, however, to say that *get back* means 'return' or 'recover'; yet the four dictionaries that enter the phrase give these meanings as well as 'retaliate'.

Etymologies are not given for multiword lexemes when their linguistic origin is obvious. Etymologies are given for foreign phrases and for other words not entered elsewhere, like *runcible* in *runcible spoon.* Sometimes what seems obvious is not; *upside down* is shown by all the dictionaries to be an alteration of an earlier phrase, *up so down,* meaning 'up as if down'. An etymology for the odd phrase *as it were* (a grammatical relic) would be useful, but none of the dictionaries give it. Only W9 indicates the origin of main-entry compounds; e.g., *rake-off* derives from *rake* (the verb) plus *off.* For expressions like *rake-off,* users are usually more interested in historical than in linguistic origins, and here W9 adds that *rake-off* comes "from the use of a rake by a croupier to collect the operator's profits in a gambling casino." This explanation is also given in AHD2, LDEL and WNWD. However, such explanations are not plentiful in any of the dictionaries, perhaps because the information is not readily available to the lexicographer.

Instead of giving an etymology, dictionaries sometimes indicate the origin in the definition. AHD2 includes the literal meaning of *on the beam* as the first sense: "following a radio beam, as an aircraft." And CTCD slips in the origin of *know the ropes* by defining the phrase "to understand the detail or procedure, as a sailor does his rigging."

In summary, one can say that the treatment of multiword lexemes in desk-size dictionaries of English can be improved in several ways.

1. Formulate policies of collection and selection that will include more multiword lexemes of value to dictionary users. Fewer transparent collocations and more idiomatic sentences are in order.
2. Weigh the relative advantages of covering single-word anomalies at single-word senses or at multiword entries and formulate a consistent policy.
3. Deal consistently with quasi-lexemes like *day after day.*

4. Formulate simple policies regarding place of entry and explain them in the front matter of the dictionary.
5. When indicating a variable possessor, distinguish between one referring to the subject of the sentence and one referring to someone else.
6. Eliminate pronunciations readily found at other entries and include stress marking for all compounds.
7. Consider providing grammatical labels for all lexemes.
8. Eliminate unnecessary literal senses.
9. Consider providing historical etymologies when these are appropriate and available.

Today's dictionaries provide users with much better treatment of multiword lexemes than those of the eighteenth century. Let us continue to advance.

Cited dictionaries

AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE (AHD2)
W. Morris/M. Berube. Boston: Houghton-Mifflin (1969/82).

CHAMBERS TWENTIETH CENTURY DICTIONARY (CTCD)
T. Davidson/A.M. Macdonald. Edinburgh: Chambers (1901/83).

COLLINS ENGLISH DICTIONARY (CED)
P. Hanks et al. London & Glasgow: Collins (1979/80).

LONGMAN DICTIONARY OF THE ENGLISH LANGUAGE (LDEL)
H. Gay et al. Harlow & London: Longman (1984).

THE OXFORD ENGLISH DICTIONARY (OED) A NEW ENGLISH DICTIONARY ON HISTORICAL PRINCIPLES
J. Murray et al. Oxford: Oxford University Press (1928).

WEBSTER'S NEW WORLD DICTIONARY OF THE AMERICAN LANGUAGE (WNWD)
D. Guralnik. Cleveland: World/New York: Simon & Schuster (1953/80).

WEBSTER'S NINTH NEW COLLEGIATE DICTIONARY (W9)
F. Mish. Springfield, Massachusetts: Merriam (1983).

WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY (W3)
P. Gove. Springfield, Massachusetts: Merriam (1961).