

Alenka Vrbinc

University of Ljubljana Faculty of Economics, Slovenia

Macrostructural Treatment of Multi-word Lexical Items

Summary

The paper discusses the macrostructural treatment of multi-word lexical items in mono- and bilingual dictionaries. First, the classification of multi-word lexical items is presented, and special attention is paid to the discussion of compounds – a specific group of multi-word lexical items that is most commonly afforded headword status but whose inclusion in the headword list may also depend on spelling. Then the inclusion of multi-word lexical items in monolingual dictionaries is dealt with in greater detail, while the results of a short survey on the inclusion of five randomly chosen multi-word lexical items in seven English monolingual dictionaries are presented. The proposals as to how to treat these five multi-word lexical items in bilingual dictionaries are presented in the section about the inclusion of multi-word lexical items in bilingual dictionaries. The conclusion is that it is most important to take the users' needs into consideration and to make any dictionary as user friendly as possible.

Key words: macrostructure, multi-word lexical items, compounds, monolingual dictionaries, bilingual dictionaries

Makrostrukturna obravnava večbesednih leksikalnih enot

Povzetek

V prispevku avtorica govori o makrostrukturni obravnavi večbesednih leksikalnih enot v enoin dvojezičnih slovarjih. Najprej predstavi klasifikacijo večbesednih leksikalnih enot in nameni
posebno pozornost zloženkam – posebni skupini večbesednih leksikalnih enot, ki jim je
najpogosteje dodeljen status iztočnice, vendar je njihovo vključevanje v geslovnik odvisno tudi
od njihovega zapisa. Nato avtorica podrobneje obravnava vključevanje večbesednih leksikalnih
enot in predstavi rezultate kratke raziskave o vključevanju petih naključno izbranih večbesednih
leksikalnih enot v sedem angleških enojezičnih slovarjev. V poglavju o vključevanju večbesednih
leksikalnih enot v dvojezične slovarje predlaga načine obravnave istih petih večbesednih
leksikalnih enot v dvojezičnih slovarjih. V zaključku povzame, da je najpomembneje upoštevati
potrebe uporabnikov in sestaviti slovar na način, ki je uporabnikom kar najbolj prijazen.

Ključne besede: makrostruktura, večbesedne leksikalne enote, zloženke, enojezični slovarji, dvojezični slovarji

Macrostructural Treatment of Multi-word Lexical Items

1. Introduction

"They that take a dictionary into their hands, have been accustomed to expect from it a solution of almost every difficulty." These are the words of Samuel Johnson, which still hold true today, since dictionary users have high expectations from dictionaries. We all know how bothersome it is to find that the word we are looking up is not in the dictionary. But is it necessarily so that the word we do not find is actually not in the dictionary? Is there perhaps a gap between the users' expectations and the actual inclusion and treatment of various pieces of information in a given dictionary?

One of the dilemmas the lexicographer is faced with at the very beginning of work on the dictionary is what to include in the dictionary macrostructure (cf. Cowie 1999; Béjoint 2000; Hartmann 2001; Landau 2001; Jackson 2002; Svensén 2009). Traditionally, the wordlist consisted of single-word lemmas, but modern dictionaries include an increasingly large number of multi-word lexical items featuring as lemmas. How does this affect the dictionary user? Does he/she recognize a string of words as belonging together? If he/she does, does he/she know where to look it up? Is it to be found in the headword list, or within individual entries?

The headword list is a list of words. Consequently, we first have to define what a word is. The most basic definition of a word is a group of characters placed together with spaces or punctuation marks before or after (Svensén 2009, 102). But, then, how should we deal with expressions, such as the compound *airdrop* that can be spelt either solid (*airdrop*), with a hyphen (*air-drop*) or as two words (*air drop*)? Should we treat *airdrop* and *air-drop* as one word and *air drop* as two words? Should we then include *airdrop* and *air-drop* in the headword list because they fit the above definition of a word, and treat *air drop* either in the entry for *air* and/or in the entry for *drop*? It is because of these difficulties in defining the word 'word' that I will avoid using it and rather use the term 'lexical item' to refer to any word, abbreviation, partial word, or phrase which can figure as the lemma in a dictionary.

The issues raised in the previous paragraphs will be addressed in this article and solutions will be proposed as to which lexical items should be given headword status in the dictionary.

2. Multi-word Lexical Items

A study conducted into dictionary use of Slovene learners of English (Vrbinc and Vrbinc 2004) comprising 70 test subjects tested among other things students' expectations of where in the dictionary they can find different multi-word lexical items (e.g. idioms, phrasal verbs, compounds). The results of the survey clearly show that students do not consider a multi-word lexical item as a lemma, since less than 22 % of the respondents would look up a multi-word lexical item in the headword list.

Taking these results into consideration, the question should be raised of what actually constitutes a legitimate dictionary entry and what it is that makes a multi-word lexical item worth including in the macrostructure of any mono- or bilingual dictionary. Before going deeper into discussion



about the lexical items that should be given headword status, we should first take a closer look at a very complex group of multi-word lexical items.

Multi-word lexical items are very frequent in a language. According to the XMELLT project, they comprise about 30 % of the lexical stock, which means that no dictionary can ignore this common phenomenon. The inclusion of multi-word lexical items causes problems for compilers and users of mono- and bilingual dictionaries because the question arises of which of the possible entries such lexical items should be placed and found. If we study the principles applied in existing dictionaries, we can see that they vary, which means that the user may have difficulties in finding multi-word lexical items. It is of the utmost importance to find a consistent procedure and then live up to it in order for the user to know where he/she should expect a certain type of information to be placed in the dictionary (cf. Martin and Al 1990).

As multi-word lexical items often pose real problems of identification, it is necessary to first determine types of multi-word lexical items. In dictionaries written in the Anglo-American tradition, multi-word lexical items are classified and treated as 'phrases' or 'idioms' (depending on the metalinguistic terminology of a particular dictionary). Such items include pure idioms, proverbs, similes, institutionalized metaphors, formulae, sayings, catch phrases, quotations and various other kinds of institutionalized collocation (cf. Moon 1998a, 2–3; Moon 1998b, 79; Atkins and Rundell 2008, 166–72). Apart from these items, phrasal verbs can also be regarded as multi-word lexical items and the same holds true of (transparent) collocations, compound nouns, adjectives and verbs.

It has to be stressed that not all multi-word lexical items can be given headword status in a general mono- or bilingual dictionary. Multi-word lexical items that are usually not afforded headword status are:

- (transparent) collocations (e.g. *suffer a worse fate*); they are commonly included as (parts of) examples illustrating use, sometimes given in bold;
- idioms, proverbs, similes, institutionalized metaphors, formulae, sayings, catch phrases, quotations (e.g. *a hard/tough nut (to crack)*); they are commonly included in a special idioms section.

Multi-word lexical items that may be given full headword status, but more commonly appear as secondary headwords are phrasal verbs (Atkins and Rundell 2008, 182). This mainly depends on the policy of each individual dictionary. If at all, full headword status was given to phrasal verbs in previous editions of monolingual dictionaries for native speakers (this policy is still adhered to in *Collins English Dictionary*, 9th edition), but the majority of monolingual dictionaries for native speakers now handle phrasal verbs as secondary headwords appended to the entry for the verb itself (the same as monolingual learner's dictionaries).

Of all the various types of multi-word lexical items, compounds are most commonly afforded headword status. Since compounds are not always easy to identify and since they represent a complex group, a few more words should be dedicated to this specific group of multi-word lexical items.

2.1 Compounds

Compounds of interest to lexicographers belong mainly to three word classes: nouns (e.g. number plate), adjectives (e.g. blood-red) and verbs (e.g. deep-fry). They may be idiomatic and non-idiomatic. Non-idiomatic compounds (Atkins and Rundell 2008, 169) are semantically transparent, they are spontaneously produced and are found in the corpus data with a high frequency rating. These are the reasons why they pose few problems to lexicographers and dictionary users. They are often included as lemmas in English dictionaries primarily due to their heavily institutionalized character (e.g. animal rights, travel agency, tourist office). On the other hand, if we take, for example, table leg (209 hits in the ukWaC), we can see that it does not have full headword status. It is included as a separate sense of the noun leg (= one of the long thin parts on the bottom of a table, chair, etc. that support it).

Idiomatic compounds, on the other hand, are more problematic to identify. They share a few properties (ibid., 170–1), one of them being frozenness of form. The only change such compounds can undergo is that they can take inflections: e.g. mother figures, letters of credit. Compounds of this type are mostly included as headwords.

Another problem connected with compounds is their spelling. They can be spelt in three ways: solid, with a hyphen or as two words. If a compound is spelt solid, i.e. as a single word and not hyphenated, it is not problematic at all because it can only appear as a headword. The same goes for hyphenated compounds. Compounds spelt as two words may be the most difficult for the user to find, since he/she may look them up under the first element, the second element or as a unit included as the headword. The look-up operation mainly depends on the user's recognition of two words as belonging together, thus forming a compound. If we go back to our example given in the introduction (airdrop, air-drop, air drop), we can see that the three expressions have been formed in an exactly parallel way and the graphic form cannot be held to justify treating them in different ways (Svensén 2009, 102). The conclusion is that items of this kind, whether written separately, hyphenated or solid, should be accorded the same lemma status. In connection with this, lexicographers have to decide right at the beginning what form of a certain multi-word lexical item to put in; here, the corpus is indispensable. It has to be stressed that among the many advantages of using a corpus in lexicography, perhaps frequency counts are the most important (cf. Landau 2001, 302–3). If an item has a frequency below a certain value in a large, representative corpus, one can conclude that the item is relatively uncommon and omit it with some degree of confidence. The relative frequency of variants in the spelling of a word can lead one to a decision about what to regard as the lemma or preferred spelling.

There are, however, two more criteria (Landau 2001, 358) that have to be taken into consideration when deciding which word to classify as a compound. First, a multi-word lexical item must function like a unit so that its meaning inheres in the whole expression (e.g. *guinea pig*) rather than in its separate elements. No part of it can be replaced without the loss of its original meaning. The existence of semantically comparable one-word units (e.g. *rat*, *rabbit*) is further evidence that *guinea pig* is a unit. Second, the stress pattern of compounds is usually distinctive, with primary stress on the first element and very little pause, if any, between the two elements (e.g. *blackbird*). But the stress is not always a reliable criterion as the stress test does not work with every multiple lexical unit (e.g. *safety glass*).



3. The Inclusion of Multi-word Lexical Items in Monolingual Dictionaries

If we closely examine the inclusion of multi-word entries in several monolingual English dictionaries, we can establish that they adopt a different policy. Every dictionary includes many phrasal entries that are not lexical items. As Landau (2001, 358) states, encyclopaedic terms, i.e. biographical (e.g. *Julius Caesar, Alexander the Great*) and geographical (e.g. *Julian Alps, United Kingdom*) entries, need no elaboration. Less obvious are entries such as *Copernican system, listed building* or *Riemannian geometry* which are included principally because the user expects to find them in a dictionary.

In order to study how multi-word lexical items are included, we have randomly chosen five multi-word lexical items (i.e. *old wives' tale, black and white, New Age traveller, act of God, walk of life*) and checked their inclusion in five leading British learner's dictionaries (OALD7, LDOCE5, COBUILD5, CALD3 and MED2) and two British dictionaries for native speakers (CED9, ODE2). Here are the results of our survey:

	old wives' tale
OALD7	included in idioms section under the headword <i>old</i> (adjective)
LDOCE5	included in idioms section under the headword <i>old</i> (adjective)
COBUILD5	headword status
CALD3	headword status
MED2	headword status
CED9	headword status
ODE2	headword status

Table 1. The inclusion of old wives' tale in English monolingual dictionaries.

The multi-word lexical item *old wives' tale* is given full headword status in the majority of dictionaries under scrutiny and is treated as an idiomatic expression in only two monolingual learner's dictionaries.

	black and white
OALD7	included in idioms section under the headword <i>black</i> (noun) in the form of three different idioms: <i>black and white</i> , <i>in black and white</i> , <i>(in) black and white</i>
LDOCE5	headword status
COBUILD5	headword status
CALD3	included as a 'phrase' under <i>black</i> (noun), sense 2 included in idioms section under the headword <i>black</i> (noun) in the form of three different idioms: <i>be</i> (<i>down</i>) <i>in black and white</i> , <i>black-and-white</i> , <i>see things in black and white</i>
MED2	headword status
CED9	headword status, hyphenated spelling included as an idiom under the headword black-and-white (noun): in black and white

ODE2	headword status

Table 2. The inclusion of black and white in English monolingual dictionaries.

The treatment of the multi-word lexical item *black and white* is similar to that of *old wives' tale* in that it is included as a headword in five out of seven dictionaries. In CED9, the spelling of the headword differs in comparison to the spelling in other dictionaries where it also appears as the headword, since it is spelt as a hyphenated compound. In OALD7 and CALD3, *black and white* can be found in the idioms section under the headword *black* (noun) in the form of different idioms. Apart from including *black and white* in the idioms section, CALD3 also treats this item as a separate sense of the noun *black* (as a kind of 'phrase' describing photography that has no colours except black, white and grey).

	New Age traveller
OALD7	an example of use in the entry for the adjective <i>New Age</i> (with an explanation in brackets)
LDOCE5	headword status
COBUILD5	headword status
CALD3	headword status
MED2	headword status
CED9	not included
ODE2	headword status with a cross reference to traveller (noun)

Table 3. The inclusion of New Age traveller in English monolingual dictionaries.

In the case of *New Age traveller*, the dictionaries under discussion appear to have reached a consensus on its status, since five of them include it as a headword. ODE2 does not provide a definition but only a cross reference to the noun *traveller*, where *New Age traveller* is treated as a subsense of the headword. Interestingly, OALD7 includes this multi-word lexical item neither as a headword nor as an idiom, but rather as an example used to illustrate the use of the adjective *New Age*. It seems that the compilers of this dictionary considered it necessary to explain the meaning of *New Age travellers*, since an explanation (= people in Britain who reject the values of modern society and travel from place to place, living in their vehicles) is provided in brackets.

	act of God
OALD7	included in idioms section under the headword act (noun)
LDOCE5	included in idioms section under the headword act (noun)
COBUILD5	headword status
CALD3	an example of use in the entry for act (noun)
MED2	headword status

The term 'phrase' is used in the front matter of CALD3 to refer to a string of words that is not regarded as an idiom.



CED9	headword status
ODE2	included in idioms section under the headword act (noun)

Table 4. The inclusion of act of God in English monolingual dictionaries.

Obviously, the status of *act of God* is more problematic, since three dictionaries give it full headword status, three include it in the idioms section and one treats it as an example of use under the noun *act*.

	walk of life
OALD7	included in idioms section under the headword walk (noun): a walk of life
LDOCE5	headword status
COBUILD5	headword status
CALD3	included in idioms section under the headword walk (noun)
MED2	included in idioms section under the headword walk (noun): from all walks of life
CED9	sense 23 of the headword <i>walk</i> (noun): <i>walk of life</i> included as an additional piece of information provided in brackets
ODE2	included in idioms section under the headword walk (verb, noun)

Table 5. The inclusion of walk of life in English monolingual dictionaries.

The majority of the dictionaries tested include *walk of life* in the idioms section and only two give it the status of a headword. Interestingly, *walk of life* can be found in CED9 under the headword *walk* (noun), sense 23, where the definition (i.e. a chosen profession or sphere of activity) is followed by the information in brackets (i.e. esp. in the phrase *walk of life*).

As is evident from the results of our short survey, the inclusion of multi-word lexical items differs if we compare different dictionaries. Full headword status seems to be preferred in *old wives'* tale, black and white and New Age traveller (in five out of seven dictionaries). A greater degree of uncertainty as to its status can be observed in act of God, since three dictionaries treat it as a headword and three as an idiom, while walk of life is included as an idiom in four dictionaries and as a headword in two.

From the point of view of user-friendliness, the treatment of *black and white* in CALD3 is not the best option, since it makes a distinction between idioms and 'phrases', which means that one and the same multi-word lexical item is dealt with in different places of the dictionary entry. Consequently, the user is expected to know that multi-word lexical items can have a different status in one particular dictionary. The look-up process is more demanding in such cases, since the user must refer to various parts of the dictionary entry.

If we compare full headword status and the inclusion of multi-word lexical items in idioms sections, it can be concluded that OALD7 prefers to include and treat them in the idioms section (none of the above-mentioned multi-word lexical items is given headword status). On

the contrary, COBUILD5 lists all five multi-word lexical items as headwords and also MED2 seems to be in favour of the headword status (four out of five multi-word lexical items). All other dictionaries do not show such great differences between the headword status and the inclusion as idioms. However, it is difficult to draw any definitive conclusions on the basis of such a small-scale study. Therefore, a further investigation into this matter would be needed to test the validity of the above results.

4. The Inclusion of Multi-Word Lexical Items in Bilingual Dictionaries

It can be seen that the inclusion of multi-word lexical items in monolingual English dictionaries differs, and a question can thus be posed regarding how to include them in a bilingual dictionary. Should the bilingual lexicographer follow the same principles as the monolingual one? In many cases, it is the compiler's decision where and how to include multi-word lexical items, but this decision has to be based on a careful study of existing monolingual sources and electronic corpora. If we take the multi-word lexical items whose inclusion in monolingual English dictionaries has been discussed in section 3, we can see that in a bilingual English-Slovene dictionary they can be treated in the following way (the examples below are taken from an ongoing project aimed at the compilation of a general English-Slovene dictionary):

old wives' tale sam. stare vraže, babje čenče

(plus as an idiom in the entries for other constituent elements with a cross reference to **old wives' tale**)

black and white prid. 1. črno-bel 2. jasen, očiten IDIOMI (in) black and white črno-belo; in black and white črno na belem

(plus as an idiom in the entries for other constituent elements with a cross reference to **black** and white)

New Age traveller sam. (v Veliki Britaniji) kdor zavrača vrednote sodobne družbe in potuje iz kraja v kraj ter živi v vozilu

(plus as an idiom in the entries for other constituent elements with a cross reference to **New Age traveller**)

act of God sam. PRAVO višja sila

(plus as an idiom in the entries for other constituent elements with a cross reference to act of God)

walk of life sam. družbena plast

(plus as an idiom in the entries for other constituent elements with a cross reference to walk of life)

All of these can be included as headwords, but it is next to impossible to predict whether users will look up a multi-word lexical item as a headword or will simply look up one of the constituent elements of such a lexical item (but which one?). This depends mainly on the user's ability to recognize a multi-word lexical item as a unit. It is therefore recommendable to approach this problem in a more user-friendly way, i.e. to include such items in two ways: as headwords



and as units in the idioms section of the entries for all constituent elements (e.g. *old wives' tale* should also be included in the entries for the adjective *old* and the nouns *wife* and *tale* and the appropriate cross references should be provided to guide the user to the entry where such a multiword lexical item is treated).

Including multi-word lexical items as headwords and at the same time as idioms in the idioms section with the cross reference is one possibility, but there are cases where a multi-word lexical item can be given either full headword status or be treated in the entry for one of its constituent elements as a separate sense. For example:

- The multi-word lexical item off day is included as a headword:
 off day sam. POG. slab dan
- One of the senses of the adjective *off* is 'below the usual standard or rate' and 'dan, teden' (= day, week) can only function as an element of equivalent differentiation in the form of a collocator (sense 4 in the example below):
 - **off** prid. **1.** (hrana) pokvarjen: *go off* pokvariti se **2.** BRIT., POG. nevljuden, neprijazen, nesramen **3.** BRIT., POG. nesprejemljiv **4.** (dan, teden) slab **5.** (sezona) mrtev

Since a lexicographer cannot presuppose where in the dictionary a user will perform a lookup operation, it is sensible to consider the option of including a multi-word lexical item in the idioms section although it cannot be classified as an idiom according to the phraseological criteria. For example:

day sam. ... IDIOMI ... off day POG. slab dan ...

If we closely observe the inclusion of hyphenated and non-hyphenated items in the monolingual English learner's dictionaries, we can see that the treatment varies according to spelling. The hyphenated item appears in the macrostructure as an entry, whereas the same item that is not hyphenated is included in the idioms section. It seems sensible to adopt the same policy when compiling a bilingual English-Slovene dictionary. For example:

off-the-cuff prid. iz rokava

(when hyphenated, it is included in the macrostructure as an entry)

cuff sam. ... IDIOMI off the cuff iz rokava

(when it is not hyphenated, it is included as an idiom in the entry for cuff, noun)

Such a treatment of compounds is recommendable for the sake of user-friendliness as the user may come across different spellings of the same expression which also dictate his/her look-up operation.

5. Conclusion

The lexicographers' task is the selection and classification of multi-word lexical items, which should be done in such a way as to ensure that users will have as few problems as possible finding such items in a dictionary. Users may have difficulties with identifying such items already in texts and if they fail to identify them in a text, they cannot successfully look them up in a dictionary. Before starting to compile a dictionary, whether a monolingual or a bilingual one, a decision should be reached as to what multi-word lexical items should be included and how they should be included – in the macrostructure (i.e. as entries in their own right) or in the microstructure (i.e. as idioms in the idioms section or as examples of use) or both, so that subjective judgements of dictionary compilers are minimized entirely and so that such items are treated in a way that is as consistent as possible. For the sake of user friendliness, it may be recommendable to include one and the same multi-word lexical item in two places in a dictionary, although this is a space-consuming policy. The front matter should provide clear instructions as to where these items are included and how they are treated so that users become familiar with the principles of inclusion and consequently, the number of times they look up a multi-word lexical item in a dictionary in a wrong place is reduced to a minimum.

Bibliography

A. Dictionaries

Anderson, S., et al., eds. 2007. Collins English Dictionary. 9th edn. Glasgow: HarperCollins Publishers. (CED9)

Mayor, M., ed. 2009. Longman Dictionary of Contemporary English. 5th edn. Harlow, Essex: Pearson Education Limited. (LD0CE5)

Rundell, M., ed. 2007. Macmillan English Dictionary for Advanced Learners. 2nd edn. Oxford: Macmillan Education. (MED2)

Sinclair, J., and M. Clari, eds. 2006. *Collins COBUILD Advanced Learner's English Dictionary.* 5th edn. London: HarperCollins Publishers. (COBUILD5)

Soanes, C., and A. Stevenson, eds. 2003. Oxford Dictionary of English. 2nd edn. Oxford: Oxford University Press. (ODE2)

Walter, E., and K. Woodford, eds. 2008. *Cambridge Advanced Learner's Dictionary.* 3rd edn. Cambridge: Cambridge University Press. (CALD3)

Wehmeier, S., ed. 2005. Oxford Advanced Learner's Dictionary of Current English. 7th edn. Oxford: Oxford University Press. (OALD7)

B. Other literature

Atkins, B.T.S., and M. Rundell. 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Béjoint, H. 2000. Modern Lexicography: An Introduction. Oxford: Oxford University Press.

Cowie, A.P. 1999. English Dictionaries for Foreign Learners: A History. Oxford: Oxford University Press.

Cross-lingual Multi-word Expression Lexicons for Language Technology. http://www.cs.vassar.edu/~ide/XMELLT.html. Accessed on 6 November 2009.

Hartmann, R.R.K. 2001. Teaching and Researching Lexicography. Harlow: Pearson Education Ltd.



- Jackson, H. 2002. Lexicography: An Introduction. London, New York: Routledge.
- Landau, S.I. 2001. Dictionaries: The Art and Craft of Lexicography. 2nd edn. Cambridge: Cambridge University Press.
- Martin, W., and B.P.F. Al. 1990. User-orientation in Dictionaries: 9 Propositions. In: *BudaLEX '88 Proceedings*, ed. T. Magay and J. Zigány, 393–9. Budapest: Akadémiai Kiádo.
- Moon, R. 1998a. Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford: Clarendon Press.
- Moon, R. 1998b. Frequencies and Forms of Phrasal Lexemes in English. In *Phraseology: Theory, Analysis, and Applications*, ed. A.P. Cowie, 79–100. Oxford: Oxford University Press.
- Svensén, B. 2009. A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge University Press.
- Vrbinc, A., and M. Vrbinc. 2004. Language Learners and Their Use of Dictionaries: The Case of Slovenia. *Erfurt Electronic Studies in English* 3. 31 pp. http://webdoc.gwdg.de/edoc/ia/eese/eese.html