Oxford Handbooks Online

The Treatment of Multi-word Units in Lexicography

Christiane Fellbaum

The Oxford Handbook of Lexicography
Edited by Philip Durkin

Print Publication Date: Nov 2015 Subject: Linguistics, Lexicography
Online Publication Date: Mar 2016 DOI: 10.1093/oxfordhb/9780199691630.013.31

Abstract and Keywords

This chapter presents a brief typology of multi-word units (MWUs) based on criteria including selectional preference, semantic compositionality, and syntactic well-formedness, focusing on support verb constructions and verb phrase idioms. Three properties of MWUs make lexicographic treatment a challenge. First, their lexical composition tends to be idiosyncratic. Second, many MWUs are semantically non-compositional; this holds in particular for idiomatic expressions. Third, corpus data show that MWUs are not just frozen ‘long words’ but exhibit considerable morphosemantic and lexical flexibility that defies their treatment as lexical units. In light of these challenges, a critical examination is presented of entries for MWUs in a few representative lexical resources.

Keywords: collocation, idioms, compositionality, flexibility, lexical units, multi-word units, support verb constructions, verb phrase idioms

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

25.1 Introduction

THE productive, potentially infinite, combinability of discrete, minimal units into larger ones is a defining hallmark of human language. Words are the building blocks for phrases and sentences, and the encoding and decoding of messages requires access to that component of human grammar where words—or, more formally, form–meaning pairs—reside. The lexicon is large by any count and its boundaries are fuzzy. Both linguistic theorists and lexicographers tend to agree that efficient storage and look-up of words and their meanings require economy; multi-word sequences that obey the principles of well-formedness and semantic compositionality probably have no lexical status and merit inclusion neither in speakers' mental lexicons nor in lexical resources compiled by lexicographers. For example, phrases like *weather forecaster* or *extreme winters*, though they may be frequent, can be composed and understood on the fly, while the meanings of phrases like *fire sale* and *lend a hand* are not straightforwardly (de)composed. For this reason, such semantically idiosyncratic phrases constitute lexical units despite their multi-word make-up. However, we shall see that the boundary between multi-word units (MWUs) with lexical status and freely composed phrases is often not clear-cut.

Analyses of spoken and written corpora reveal a high percentage of MWUs, including collocations and idioms, both in terms of types and tokens (Jackendoff 1997; Moon 1998; Cowie 1992). MWUs are a compelling subject of study not only because of their pervasiveness and universality, but also because they challenge any definition of 'word' and resist clear-cut criteria for integration into both speakers' mental lexicons and lexical resources. On the one hand, speakers recognize MWUs as lexical (p. 412) units paired with distinct meanings, a fact reflected in traditional lexicography and the 'classical' view of MWUs as merely long, fixed 'prefabricated' lexemes. On the other hand, corpus data reveal that many MWUs show considerable variation, and they present strong evidence that many phrasal units are subject to regular morphosyntactic processes that may operate on the phrase as a whole or on internal constituents, independent of their semantic transparency. Semantic wholeness in the face of rich grammatical properties makes the treatment of MWUs in lexical resources a challenge. Before addressing the lexicographical treatment of MWUs, we distinguish and define several key concepts.

25.2 Co-occurrence, Selectional Preference, and Collocation

Words are selective about their context. For example, an English speaker *brushes* his teeth, unlike his French counterpart, who *washes* them. And English speakers *go off on a tangent* more often than they *take off on a tangent*. We talk about a *confirmed bachelor*

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

rather than an *affirmed bachelor*, and *disasters* are *unmitigated* but rarely, if ever, *unrelieved* or *undiminished*, although the less usual phrases would be understood by a cooperative listener.

The idiosyncracies of lexical selection are reflected in the regular and statistically discernible phenomenon of collocation (Firth 1957b; Sinclair 1991; Stubbs 2001; Partington 2004; McEnery and Wilson 1996, inter alia). Church and Hanks (1990) demonstrated the measurability of collocational properties with Mutual Information—a statistical measure based on corpus-based co-occurrence frequencies—that quantifies the idiosyncratic attraction of wordforms to one another beyond syntactic and semantic constraints imposed by subcategorization and selectional restriction rules. Thus, while we understand the meaning of the phrase *powerful tea*, corpus frequency data show that the noun *tea* overwhelmingly prefers to select *strong* as the adjective to express the appropriate meaning.

Statistical analyses of the phenomena of collocation enable the discovery of collocations (e.g. Bond et al. 2003; Baldwin et al. 2003; Fazly and Stevenson 2006; Fazly et al. 2009; Evert 2004/2005; Schone and Jurafsky 2001, inter alia) and to quantify the strength of their co-occurrence as well as the degree of their lexical and syntactic fixedness, a measure of their lexical status. Such analyses show that there are no hard rules to distinguish between merely preferred co-occurrences and more or less fixed collocations that arguably have lexical status. Rather, co-occurrence preferences are situated on a continuous scale of fixedness, and consequently, there is no clear cut-off point that would decide when a phrase qualifies as a unit and for inclusion in the lexicon. Selectional preferences remain a challenge for learners and for machine translation systems, which (p. 413) may successfully identify the appropriate meaning but fail to pick out the most felicitous wordform.

25.3 Collocations, Support Verb Constructions, and Idioms

MWUs are characterized not only in terms of their collocational strength but also in terms of their formal, morphosyntactic properties and their semantic compositionality, that is, the extent to which the meaning of the phrase derives from the meanings of its constituents. We present a brief typology for MWUs in contemporary English, focusing on several sub-classes of collocations and idioms. All are statistically significant co-occurrences of specific lexical items and fall along a sliding scale of syntactic fixedness and semantic non-compositionality.

25.3.1 Collocations

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

Collocations—multi-word combinations that show strong collocational attraction— include noun compounds (*laptop computer, book sale*), light and support verb constructions (*have a drink, take a picture, make a fuss*), syntactically marked phrases such as the propositional phrases consisting of a verb and a bare noun (*in school, to prison*), and verb phrases like *answer the door*.

25.3.2 Noun Compounds

Many noun compounds are semantically opaque *(speed trap, ski bum*) and are clear candidates for inclusion in a dictionary, where they are treated as units or 'long words'. But such compounds, although often semantically idiosyncratic, are highly productive. The semantic relation between their members can vary, but speakers appear to have no difficulty in understanding them. Thus, a *fire sale* is a sale that takes place because of a fire; during a *bake sale*, baked goods are sold (usually for fund-raising purposes); and *garage sales* and *yard sales* occur in garages and yards. In each case, the semantic composition is quite different, though the meanings of the members of the compound are transparent. New compounds are often formed on the basis of a pattern.

Many compounds are short-lived and are closely bound to a specific context; *banana war* is readily understandable when used in discussions of nations competing to export their fruit and *helicopter mother* is interpretable only in the context of the current public debate on child rearing. There is no clear-cut answer as to whether or not (p. 414) such compounds should be included in lexical resources and coverage differs across dictionaries.

25.3.3 Prepositional Phrases

Some collocations are candidates for inclusion in lexical resources due to their syntactic idiosyncracy. These include the class of Prepositional Phrases exemplified by *in class/out of school/to college/in prison/jail/hospital.* The singular noun is always bare, and no adjectival modification is possible, that is, no lexical material can intervene between the preposition and the noun: **in boring class*/**out of high-security prison*/**in local hospital*. This collocational pattern is productive to a limited degree:

in school/in graduate school/in medical school/in kindergarten/in college
in chemistry/Spanish class
*in jail/*in penitentiary/*slammer/*workhouse*
*in hospital/*infirmary/*clinic*
*in court/*in Supreme Court*
*in bed/*on sofa*

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

Because the patterns are not fully productive, dictionary entries for nouns like *court* and *jail* need to indicate their use in such phrases.[1]

25.3.4 Support and Light Verb Constructions

Support Verb Constructions (SCVs) or Light Verb Constructions (Grimshaw and Mester 1988; Kearns 1998/2002, inter alia) are syntactically well-formed verb phrases that are semantically compositional but exhibit strong lexical preferences.[2]

SVCs consist of a 'light' or support verb and a complement that can be an NP, a PP, a double object, or an NP–PP:

give an explanation, have a drink
set on fire, put on hold
give the floor a sweep, keep someone company
take account of something, make a call to somebody

(p. 415) The choice of verb in such phrases does not admit any lexical substitution, even of arguably close synonyms, without a change or loss of meaning. For example, the verb in the phrase *take a bow*, meaning to accept applause or acclaim, does not admit of the substitute *make*; the meaning of *set in motion* is lost in the phrase *place in motion*, and while *give someone company* seems intuitively plausible, the common phrase is *keep someone company*.[3]

Unlike collocations such as *brush one's teeth*, which seem to be unsystematically distributed across the lexicon, SVCs constitute a syntactically, lexically, and semantically well-defined class found in many languages (see, for example, Grimshaw and Mester's (1988) discussion of Japanese *suru* constructions, and similar constructions in Persian). Other, related, examples are *give chase/voice to*, and constructions of the form *have N*, where the noun is a bare verb stem, for example *have a drink/read/smoke/look*. No satisfactory account has been offered so far for the productivity of these phrases and its limitations (but see, for example, Wierzbicka's (1982) discussion of the constructions *have a drink* vs. **have an eat*).

Most of the prepositional phrases and SVCs discussed here may be lexically or syntactically idiosyncratic, but they are usually semantically decomposable ('encoding', in Fillmore et al.'s (1988) terminology) and can be readily interpreted by speakers unfamiliar with the phrasal units in the appropriate context. However, the choice of the particular verb in these constructions to the exclusion of other verbs is unpredictable and requires that SVCs be represented in lexical resources as part of the entry for the noun.

25.3.5 Idioms

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

We next consider different classes of Verb Phrase idioms; each poses different challenges for adequate lexicograpahical representation.[4]

25.3.5.1 Idioms with Irregular Phrase Structure

Some idioms are syntactically idiosyncratic and cannot be assigned to a phrasal category. Their constituents are semantically transparent lexemes, but they are 'unfamiliarly arranged', in the words of Fillmore et al. (1988). While their phrasal irregularity makes these MWUs clear candidates for inclusion in lexical resources, they are also fixed and constitute syntactic units, making their treatment straightforward. Examples are

*nothing doing/nothing much doing (cf. *nothing was doing/*nothing was done)*
*and then some (cf. *and then more)*
*say when (cf. *say what/*when was said)*
*all of a sudden (cf. *some of a sudden)*
*by and large (cf. *by and very large)*

(p. 416)

Fillmore et al. (1988) further distinguish constructions where 'unfamiliar pieces are unfamiliarly arranged'. Such 'decoding' idioms are non-compositional and unanalysable. The 'unfamiliar pieces' in some MWUs could be considered to be so-called 'cranberry-morphemes', whose distribution is strictly limited to the idiom, as in *kith and kin, spic and span*. Speakers do not usually assign a meaning to the nouns outside of the idioms (though *kin* of course has an independent meaning). Because these expressions cannot be assigned to a lexical or syntactic category, the idioms are syntactically irregular and, consequently, frozen:

*kith and kin (cf. *kin and kith, *kith and relative)*
*spic and span (cf. *spic and very span)*

Adequate lexicographical treatment would reflect the structural unity and fixedness of these phrases, treating them as a single lexical item.

25.3.5.2 Partially Lexically Filled Idioms

Many idioms are verb phrases with noteworthy syntactic properties (see Lebeaux 1988, 2000 for a comprehensive classification). A large number of verb phrase idioms are discontinuous, where one internal argument is lexically unspecified (Fillmore et al. 1988 refer to these as 'formal idioms'). These include the many English idioms with a possessive bound to the subject (e.g., *have one's cake and eat it, be on one's last legs, blow one's stack*) or a pronoun bound to a lexically free noun:

cook s.o.'s goose
give s.o. the slip/a hand
take advantage of s.o.
put s.o. on a pedestal

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

These idiomatic constructions are syntactically regular and constitute semantic units, yet they are not fully spelled out lexical units. Their entries need to clearly indicate not only the expression's structure and meaning but also the place of the open position. Thus, a noun phrase with possessive morphology is part of the idiom *cook someone's goose*; this possessive cannot appear in an appositive structure without a loss of the idiomatic reading: (p. 417)

cook Peter's goose
**cook the goose of Peter*

However, a Beneficiary argument can usually be syntactically realized either as an indirect object or in an adjunct:

give the slower students a hand
give a hand to the slower students
give the police the slip
give the slip to the police

25.3.5.3 Schematic Idioms

Fillmore et al. (1988) and Kay and Fillmore (1999) provide a fine-grained analysis of so-called 'schematic idioms'. These are specific syntactic configurations characterized by the presence of a few lexical items (usually function words) and specific meanings. Examples are *the X-er the Y-er* and *not X let alone Y*. The range of lexemes they admit in the unfilled slots is highly constrained by the meaning of the schema. Thus, the two lexemes framing *let alone* must be in some kind of scalar relation, with *Y* expressing a greater value than *X* (*the water in the hotel was not warm, let alone hot; I can't pay the rent let alone make a downpayment on the apartment*).

Schematic idioms are syntactically, lexically, and semantically irregular; their properties must be associated directly with the construction and may be extremely complex, as in the case of the 'Mad Magazine Construction' (*Him a doctor?).* These constructions are fixed and show no syntactic variation, although they allow a range of lexemes whose meanings are constrained by the meaning of the construction. As Fillmore et al. (1988) show in almost forty pages of discussion, a comprehensive account of the lexical productivity and its limitation requires a description that exceeds the format of conventional dictionary entries.

25.3.5.4 Verb Phrase Idioms with Modified Phrase Structure

A large number of idioms are Negative Polarity items (NPIs); in the absence of a negation expression, these phrases lose their idiomatic meanings (e.g. Lichte and Sailer 2004). Examples are:

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

not give a fig/damn/hoot
not have a leg to stand on
horses wouldn't get NP to VP
not be quite right in the head

(p. 418) The challenge for adequate lexicographical treatment of these MWUs is to convey the obligatoriness of the negation while allowing for variation in the specific choice of negation:

nobody gave a fig about the victims
he never had a leg to stand on
I don't know whether he is quite right in the head

A number of NPI idioms are headed by modal verbs, and the absence of an idiom-specific modal changes the meaning of the phrase:

won't hear of it (cf. I don't hear of it)
can't take it with you (cf. she didn't take it with her)

The lexicographer must convey this constraint in the entry of the phrase; the considerable variability of the negation is particularly challenging.

25.4 Semantic Compositionality

Like collocations, idioms vary in the extent to which they are lexically and syntactically fixed. But unlike collocations, which are compositional, idioms are semantically opaque to different degrees. The meaning of syntactically well-formed idioms like *kick the bucket, bite the dust, rock the boat*, and *hit the ceiling* does not arise from the meanings of their constituents and cannot be guessed by speakers unfamiliar with the expression. Consequently, such semantically non-compositional expressions must be represented as units in lexical resources. However, their representation as strings that suggests fixedness does not do justice to the considerable variation with which speakers use these MWUs, as shown in Section 25.4.

Many VP idioms, like *spill the beans* and *let the cat out of the bag* are considered to be at least partially analysable, in that speakers assign a reading to one or more of their constituents. Nunberg et al. (1994) call 'idiomatically combining expressions' or 'internally regular' those idioms whose parts can be given a literal interpretation; in Fillmore et al.'s (1988) terminology, they are 'encoding'. For example, in *spill the beans, spill* can be interpreted as 'reveal' and the *beans* as 'information' or 'secret' (see also Fellbaum 1993); similarly, *cat* in *let the cat out of the bag* refers to sensitive or secret information, while *bag* refers to an (abstract) enclosure or hiding place. Each component of such idioms can be semantically interpreted, and the combination of these meaning

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

constitutes a string whose form and meaning correspond in a one-to-one fashion to that of the idiom. By contrast, Nunberg et al. (1994) note that 'internally irregular' idioms like *kick the bucket* and *buy the farm* are not compositional, as the constituents of a transitive verb construction cannot be (p. 419) semantically interpreted and re-assembled to map onto the literal reading of the intransitive verb, 'die'.[5]

25.4.1 Metaphors in Idioms

Metaphorical idioms (e.g. Wood 1986) are compositional in that one or more of their constituents can be interpreted as a metaphor independent and outside of the idiomatic context. For example, *fire* is a conventional metaphor and readily interpretable as a potential or real danger in idioms like *play with fire, pull the chestnuts out of the fire, be in the line of fire*; the same metaphor is reflected in the meaning of *get burned*, 'suffer a setback' (Lakoff and Johnson 1980). By contrast, idiom components like *cat* in *let the cat out of the bag* are context-specific metaphors; *cat* cannot be freely used to mean 'secret information'. While one might argue for dictionary entries for conventional metaphors like *fire* that exhibit a certain degree of distributional freedom, entries for highly context-specific metaphors like *cat* do not seem warranted.

25.4.2 Variation in Idioms

Perhaps the greatest challenge for the lexical treatment of idioms is that very few are completely fixed but allow for variations, often as rich as freely composed strings. Consequently, lexical entries casting them as frozen sequences of words, while correctly capturing their semantic unity, do not do justice to their usage.

Fillmore et al. (1988) note that most idioms allow at least verbal inflection, such as variation in tense and number of the lexically unfilled subject. The misconception that idioms are frozen is probably due to the fact that much of the literature on idioms and collocations is based on data derived via introspection. Moon (1998) was one of the first comprehensive studies of English idioms based on corpora, and her data challenge a simple integration of idioms into any theoretical grammatical framework. Neumann et al. (2004) and Fellbaum (2006, 2007) examine German idioms using a one billion word corpus. Search queries that allow for the retrieval of lexical and syntactic variations of the idioms' canonical forms (Herold 2007) return numerous examples of variations and demonstrate that most idioms participate in the regular grammatical processes associated with free language. Strikingly, the data refute the prevailing view that an idiom's variability is entirely conditional on its semantic transparency, as articulated in, for example, Nunberg et al. (1994). Flexibility cannot be straightforwardly accounted for in terms of semantic compositionality.

(p. 420) 25.4.2.1 Lexical Variation

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

Corpus data show that in many idioms a constituent can be exchanged for another, semantically related lexeme. Quite often, such substitution has a playful character and alludes to a specific event or state of affairs. For example, when the secret life of golf champion Tiger Woods was revealed, newspaper readers encountered headlines such as *Who let the Tiger out of the Bag?* Besides playing on the golfer's name, the use of *tiger* rather than *cat* here implies the considerable magnitude of the scandal. Such variations are highly context- and situation-specific, tend to be found in the public media, and are usually short-lived.

Besides substitution of the idioms' noun components, corpus data show that adjectives are often added to the nouns (Fellbaum and Stathi 2006; Stathi 2007). Ernst (1981) calls 'external modification' cases where the adjective in fact modifies the entire idiom, much like an adverb, as in

Carter doesn't have an economic leg to stand on

Many people were eager to jump on the horse-drawn Reagan bandwagon

These examples show further that many of the attested variations of common idioms play on specific situations, events, and people.

Another kind of variation, where a speaker adapts a noun phrase's determiner or number, is found particularly often with metaphors and nouns that can be mapped onto a referent. An attested example is *more than one cat was let out of the bag that night*.

Lexical variation extends to changes of category of the major constituents. Thus, Moon (1998) cites corpus data where a verb has been turned into a noun:

lose face/loss of face

waste one's breath/a waste of breath

break the ice/ice-breaker

25.4.2.2 Syntactic Variation

Idioms are also subject to considerable syntactic operations. Moon (1998) cites corpus examples of passivization of English idioms (*Mary's teeth were gnashed as the home team went down in defeat*), relativization (*That is a bullet on which the Arthur Golds of this world have steadfastedly failed to bite*), and pronominalization (*if there is ice, Mr. Clinton is breaking it*). Other examples found on the Web include the following:

Beyoncé has finally cozied up to *the cat that was let out of the bag* months and months ago

If those experiences were the apex of their lives, why wouldn't they just go on repeating them, over and over, until *the bucket was finally kicked*?

(p. 421)

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

The evidence clearly refutes the common claim that idioms are fixed unless their components can be semantically interpreted. Cases like the passivization of the semantically unanalysable *bucket* also call into question the notion of a continuum of flexibility that interacts with semantic transparency, as proposed by Abeillé (1995), who based her analysis of a number of French idioms, and Dobrovol'skij (1999), inter alia.

25.5 Consequences for the Lexical Representation of MWUs

The wide range of MWUs examined here all share the property of being semantic units; as such, they must be included in lexical resources.[6] But, given the considerable variation of many MWUs, their representation as 'long words', which suggests fixedness, seems inadequate. At the same time, it is clearly impossible to give a comprehensive account of the variability and its limitations. We examine three kinds of lexicons and their treatment of MWUs.

25.5.1 Virtual Lexicons

I call 'virtual lexicons' models of the lexicon constructed by linguists and psycholinguists. One challenge for these lexicons is to account in a systematic and comprehensive way for people's linguistic behaviour with respect to MWUs. Virtual lexicons tend to discuss a small number of cases and do not strive to construct large-scale resources. Soehn (2006), working in the Head-driven Phrase Structure Grammar (HPSG) framework, borrows the well-established lexicographic principle of cross-listing. He distinguishes between metaphoric components of idioms and semantically opaque ones (called 'listemes', following diSciullio and Williams's (1987) distinction between regular, class-based, and irregular, listed items in the lexicon). Constituents of decomposable idioms are encoded in a lexical entry together with a literal string (e.g. *spill, divulge*). In Soehn's model of the lexicon, the string in an idiom selects its complement (*beans*) via its listeme value.

(p. 422) 25.5.2 Traditional Paper Dictionaries

I sample a few dictionaries focusing on collocations that specifically target learners of English. Benson et al.'s (1986) *BBI Combinatory Dictionary of English: A Guide to Word Combinations* is an exemplary resource for learners of English. *BBI* lists grammatical and lexical collocations. Eight types of grammatical collocations are distinguished, based on their syntax (e.g. preposition + noun, as in *by accident, at anchor* and predicate adjective + *to* + infinitive, as in *ready to go* and *difficult to convince*). The patterns applying to each headword are indicated by means of a code. Lexical collocations are phrases characterized not by their syntactic structure but by their lexical make-up, which is more

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

or less fixed, arbitrary, and not predictable. Thus, *BBI* list the verbs *commit* and *attempt* in the entry for *suicide*, implicitly guiding the learner to avoid producing combinations like *perpetrate suicide*.

The *Oxford Collocations Dictionary for Students of English* (2009) by Colin McIntosh lists for each of its 9,000 headwords the most common collocates (lexical collocations), with illustrative examples drawn from the two billion word Oxford English Corpus, and specifies the part of speech. For example, the entry for *ablaze* lists the verbs *be* and *set*, and specifies that *with* is the head of the prepositional phrase following *ablaze* (*every window was ablaze with light*).

Boatner et al.'s *Dictionary of American Idioms* (1975) is not based on corpus data, reflecting an older lexicographic tradition. It treats idiomatic MWUs as fixed long words. While such information can convey their meaning to a user who encounters an unknown expression, it does not tell him how to productively use the expression. Moreover, if the speaker encounters a token of the expression that includes a lexical variation, he may not find the MWU in its canonical form in the dictionary. Like many resources, Boatner et al. make use of cross-listings. Thus *give up the ghost* can be found under the entry for *give*, and looking up *ghost* refers the user back to the entire idiom. But, as in the case of paper dictionaries that were compiled under economic constraints, cross-listing is not consistent; thus, *sell down the river* is included in the entry for *sell* but not for *river*.

25.5.3 Electronic Lexical Resources for Collocations

Modern lexicographic resources are no longer bound to the paper format. Electronic resources are not only unconstrained as concerns their size but they offer the possibility of rich cross-referencing and linking among entries and thus more points of access to a given wordform or meaning. Moreover, the availability of electronic corpora now makes corpus-based lexicography possible, reflecting a broad, varied speaker community rather than data constructed by the lexicographer. In fact, the division between corpus and lexicon melts away, and the lexicographer's task can be reduced to the annotation of corpus data rather than the crafting of lexical entries.

An example of a corpus-based lexical resource for idioms is the German collocation database (Fellbaum 2006, 2007). One thousand selected German verb phrase idioms were analysed by searching a one billion word corpus (Klein and Geyken 2010). The corpus was searched using flexible regular expressions involving wordforms (p. 423) characteristically, although not necessarily, associated with a given idiom (e.g. German *Gras, beissen*, lit. 'grass', 'bite' corresponding to English *bucket, kick* in the idiom *ins Gras beissen*). Corpus tokens showed that idioms that are traditionally represented as fixed strings in fact exhibit a rich variety that extends not only to syntax (passivization, focusing, etc.) but also to lexical substitution and modification (see Section 25.4.2). The corpus examples were manually sorted and classified according to their lexical and morphosyntactic signatures. An online interface allows the user to search for particular

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

expressions and all variations with illustrative corpus examples that provide an impression of the idioms' flexibility, although no exhaustive account of their full use can be given.

The online *Oxford Collocation Dictionary for Advanced Learners of English* (ozdic.com) gives rich information about a word's collocational properties grouped by part of speech. For example, the entry for *heart + Verb* lists many expressions, including idiosyncratic combinations and idioms, with corpus examples:

> jump, leap, lurch, miss/skip a beat Her heart leapt with joy. | ache My heart aches when I think of their sorrow. | desire sth everything your heart could desire | sink | go out Our hearts go out to (= we sympathize deeply with) the families of the victims.

The *Oxford Collocation Dictionary for Advanced Learners of English* includes a separate category PHRASES. Looking up *hand* and *ear*, one finds *fall into the wrong hands* and *fall on deaf ears*, respectively. But even digital resources do not offer comprehensive cross-listings.

25.5.4 Corpus-based Dynamic Resources

Corpus data not only aid learners by giving them direct access to attested examples, but are important in providing lexicographers with representative data on which to base a lexical entry. The Sketch Engine (Kilgarriff et al. 2004) is an important tool that, for a given target word, returns frequent and prototypical examples from several corpora. Lexicographers can choose and add corpora so that the resource has a dynamic character.

The Sketch Engine provides 'sketches' of the keyword, consisting of corpus lines (key-word-in-context, or KWIC lines). It provides a frequency score, based on the corpus, of words that co-occur with the keywords. Moreover, the collocates are classified in terms of their syntactic relation to the keyword (modifier, modified-by, object-of, etc.). Importantly, the Sketch Engine allows one to compare several words with respect to their frequent collocates. This reflects both the uniqueness of a given word in an MWU—and hence the fixedness of the expression—and the substitutability of similar lexemes in a collocation. Unlike the frozen, stative database of the German collocation project, the Sketch Engine allows the user to create a customized corpus from selected corpora and corpus examples. The larger the corpus, the more likely the inclusion of less prototypical and infrequent variations, which tend not to be found in resources where (p. 424) the lexicographer necessarily has to omit marginal examples. Thus, resources like the Sketch Engine potentially expose the user to the full range of variation of an MWU.

Several resources were built with Sketch Engine corpus query technology. One is Dante (Database of Analysed Texts of English, <http://www.webDante.com>), a lexicon based on a 1.7 billion word corpus that gives fine-grained descriptions for tens of thousands of

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

English idioms, phrasal verbs, and compounds. For example, the entry for the keyword *benchmark* contains the Light Verb Construction (called ‘chunk’ in Dante) *set a benchmark.*

An improvement on static lexical resources like the German collocation dictionary is the automatic collocation dictionary ForBetterEnglish.com (Kilgarriff et al. 2008). A user can enter a word into the Web-based interface and ForBetterEnglish returns the most frequent collocates, based on large corpora, each with the particular syntactic configuration. Thus, the query *disaster* returns the information that, as an object, this word occurs most frequently with the verbs *avert, spell, loom*; as a subject, is is most often followed by the verb *strike*. Its most frequent modifiers are *natural, humanitarian, unmitigated, man-mde, impending, ecological. Disaster* compounds most frequently with *recovery, preparedness, relief* as a modifier and with *tsunami, earthquake, mining*, and *WTC* as the head. For each collocation, a representative corpus example is given.

Electronic, corpus-based lexicography opens up new possibilities for a more comprehensive representation of speakers’ language. Not only does it allow access to the language of a broad swathe of a linguistic community and no longer rely on the introspection and idiolects of a handful of lexicographers, but it can display the full spectrum of syntactic and lexical variations that must be accounted for when one tries to describe the lexicon as a complex component of our linguistic knowledge.

25.6 Summary and Conclusion

The typology of multi-word collocations and idioms presented here shows that those that exhibit a certain degree of syntactic and semantic idiosyncracy and thus deserve inclusion in a lexical resource do not easily fall out. As lexical items, many are only partially filled and allow of considerable variation; this is true even for semantically non-compositional idioms that are subject to all regular grammatical processes.

Their status as lexical units on the one hand, and their considerable flexibility on the other hand make many MWUs a challenge for lexical treatment. There are no easy solutions for an optimal way to to represent MWUs in a way that could inform a speaker—and especially a learner—about the full range of their use. The challenge of representing not only their meaning but also the extent and the limits of their usage entails crossing the traditional boundary between the lexicon and the morphsyntactic component of speakers’ grammatical knowledge. On the basis of some specific cases, it was shown how some leading lexical resources try to meet this challenge; each necessarily has some deficiencies. Resources that combine corpus data with lexicographic descriptions carry great promise for doing justice to speakers’ rich use of MWUs.

Notes:

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

([1]) British and US English differ slightly in their use of these phrases; thus *in hospital* is not common among American English speakers.

([2]) The distinction between SVCs and LVCs is not clear-cut and we will ignore it here, referring to all MWUs under consideration as SVCs.

([3]) Some state-denoting complements co-occur with different verbs, reflecting regular aspectual alternations between causative, inchoative, and stative: *put/be/keep on hold, put/be on fire*; however, not all SCVs show all alternates: **go/*get/*keep on fire*.

([4]) Fixed expressions of the following kind will not be considered here: proverbs like *the early bird gets the worm*, phrases like *mum is the word, the shoe is on the other foot*, and *when the cows come home/when hell freezes over*, and routine formulae that often have a pragmatic point (Fillmore et al. 1988) such as *have a good one, take care, let's not go there*, and similes such as *sharp as a whistle*.

([5]) Interestingly, idioms in other languages also encode 'die' with transitive verbs; cf. German *den Löffel abgeben* ('hand in the spoon') and French *casser sa pipe* ('break one's pipe').

([6]) There is also rich evidence that speakers store multi-word expressions as units in their mental lexicons, notwithstanding the fact that they can serve as input to all the grammatical processes available to simple words.

Christiane Fellbaum

Christiane D. Fellbaum is a Senior Research Scientist in the Computer Science Department at Princeton University. Her research focuses on lexical semantics, the syntax–semantics interface and computational linguistics. She is one of the developers of WordNet, a large lexical database that serves as a resource for computational linguistics and many natural language processing applications. She is a founder and president of the Global WordNet Association, which guides the construction of lexical databases in many languages. She pursued her interest in multi-word expressions and idioms in the context of a large-scale corpus project on German collocations at the Berlin-Brandenburg Academy of Sciences.

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018

PRINTED FROM OXFORD HANDBOOKS ONLINE (www.oxfordhandbooks.com). © Oxford University Press, 2018. All Rights Reserved. Under the terms of the licence agreement, an individual user may print out a PDF of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice).

Subscriber: University of Edinburgh; date: 22 June 2018