Linguistica ONLINE. Issue 23. Published: May 7, 2020
http://www.phil.muni.cz/linguistica/art/dickins/dic-005.pdf
ISSN 1801-5336

AN ONTOLOGY FOR COLLOCATIONS, FORMULAIC SEQUENCES, MULTIWORD EXPRESSIONS, COMPOUNDS, PHRASAL VERBS, IDIOMS AND PROVERBS[*]1

James Dickins (University of Leeds, UK, J.Dickins@leeds.ac.uk)

Abstract: This article proposes an ontology (set of entities and explicit statement of the relations between them) for word-sequences (whether continuous or discontinuous) whose unifying feature is the co-occurrence within them of one or more of their words at greater frequency than would be predicted by their overall frequency of occurrence within the language. More precisely, the article proposes a number of possible ontologies, since, for some entities, it presents alternative possible definitions, discussing their merits and demerits. The article focuses almost entirely on English. It begins with a statement of general methodological principles. It argues we should not be attempting to discover what the true meaning of terms is, but to produce 'serviceable definitions' of terms which are at least relatively compatible with those produced by other writers and which can be coherently and explicitly related to other terms within the ontology. What I mean by a 'serviceable definition' is one which is of can be successfully used by researchers for the practical analysis of collocations and the other phenomena considered in this article. Whether the definitions proposed here are therefore 'serviceable' can only properly be judged by their successful deployment in future research.

The article then considers the following: collocations (Section 2), formulaic sequences (Section 3), multiword expressions (Section 4), compounds (sections 5-5.3), phrasal verbs (Section 6), idioms (Section 7) and proverbs (Section 8). Beyond these basic notions, the article considers other possible types of multiword expression (Section 9), further categories deriving from collocation, formulaic sequence and multiword expression (Section 10), semantic correlates of syntactic relationships in multiword expressions (Section 10), notions having fuzzy and discrete boundaries (Section 11), and universal and language-specific categories in the ontology (Section 12). Section 13 is a conclusion.

Key words: ontology, collocation, formulaic sequence, multiword expression, compound, phrasal verb, idiom, proverb

[*] Previously unpublished. [Editor's note]

[1] I thank Eric Atwell, Claire Brierley and two anonymous reviewers for *Linguistica* ONLINE for reading draft versions of this article and making very useful comments on it. These have considerably helped improve the final version. At various points in this final version of the article, I address comments made by the two *Linguistica* ONLINE reviewers on the earlier draft which they read, referring to them, where appropriate, as Reviewer 1 and Reviewer 2.

1. Introduction

This article has developed partly in response to the lack of clear global statements in the literature of the relationships between 'collocation' and a range of related notions.[2] Its aim, accordingly, is to drawn up an ontology (i.e. a set of entities and explicit statement of the relations between them) for such notions, in the hope that this will be useful for future researchers, who may either adopt it, or if they find it inadequate, build their own ontology covering the same or similar notions. The definition of 'ontology' adopted in this article is based on the fairly standard definition in logic, i.e. "set of entities presupposed by a theory" (Collins English Dictionary Online; henceforth CEDO). However, it recognises that a simple statement of the set of entities involved in a system does not tell us anything about how these entities relate to one another; hence the addition of the phrase, "explicit statement of the relations between them".

The article also takes the view that a lack of clear definitional statements is likely to lead to conceptual confusion, with different researchers unknowingly meaning different things by the same terms. The article does not base its view of collocations and related on a particular linguistic theory; and it is beyond its scope to consider collocations and related phenomena in relation to specific theoretical approaches (as done in Gries 2008, for example). The approach adopted is, however, intended to be commonsensical, and as such open to re-interpretation in terms of different theoretical approaches.

The article focuses almost entirely on English; in Section 13, it considers which of the notions which it puts forward may be considered universal, and which are specific to English (and perhaps some other languages). It also has the strictly limited aim of considering collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs, idioms and proverbs from only the perspectives which specifically allow us to differentiate between them. These are mainly statistical, syntactic and semantic (the last largely treated from a purely denotative perspective). This means that many aspects of these notions which are essential in other respects are not discussed. Thus, in the case of idioms and proverbs, for example, I do not discuss issues such as (i) the relationship (e.g. metaphorical, metonymic, or other figurative relationship) between the idiom and proverb sense and the more basic ('literal') sense, (ii) the connotations which these may give to the idiom or proverb; (iii) the 'schematic' patterning of figurative idioms and proverbs (e.g. Kövecses 2010), (iv) the pragmatic and stylistic deployment of idioms and proverbs in different kinds of text, or (v) psycholinguistic issues (for a survey of all these aspects, see Gibbs 2010). These are all very important, but, as noted, inasmuch as they do not serve to differentiate the features which the article considers, they fall outside its scope.

Karl Popper has argued that in seeking to understand a term, we should not ask the question 'What is this really?', i.e. we should not attempt to search for the essence or 'true meaning' of that term – an approach which Popper (1986 [1957]: 26–43) calls methodological essentialism. Rather, we should attempt to provide what could be called a 'serviceable definition' of the term and use this definition as our starting point for the deployment of the concept (the defined term) in subsequent argumentation. To take a specific example, we

[2] This has proved a particular problem for a number of my doctoral research students doing corpus research in Arabic and Arabic-English translation.

should not ask 'What is capitalism really?'. Rather, we should start with a serviceable definition of capitalism, and then use this to investigate relevant phenomena in relation to this definition.

Popper's argument is based on common sense. It is possible to use a term such as 'capitalism' in many different ways, i.e. with many different definitions, whether explicit or implicit. (For a review of some of a number of different – and sometimes clearly incompatible – definitions of capitalism, see Merrill 1995.) The same situation is apparent in linguistics. There are thus, for example, numerous definitions for the term 'morpheme', many of which are clearly incompatible, and use 'morpheme' to mean quite different things for what is meant by the term in other approaches (for a survey, see Bauer 2004: 70–72).

Assuming various definitions of a term to be internally coherent and to apply sensibly to the facts to which they are relevant, there is no point in arguing which is the 'correct' definition. Indeed, this is counter-productive, since it deflects from seeing definitions as used within particular theories systematically, i.e. providing terms for concepts which fit into a systematic theoretical whole. Perhaps worse still, it encourages the epistemologically naïve view that there is necessarily a reality out there which it is our task as researchers simply to discover, rather than recognising the central importance of the theoretical approach which we adopt in shaping that reality.[3] This perspective echoes Saussure's dictum, "C'est le point de vue qui crée l'objet" (Saussure 1975 [1916]: 23), "It is the viewpoint which creates the object" (Saussure 1959: 8), or "it is the viewpoint adopted which creates the object" (Saussure 1983: 8).

There are, however, practical limitations to Popper's objection to methodological essentialism. Popper's views apply very well to abstract concepts, but are less applicable to concrete phenomena. Take the example of a rainbow. In an obvious sense, we all know what a rainbow is: we can point to one in the sky, we can describe what one looks like even if there is no rainbow to look at, we can draw one on a piece of paper. What we cannot, however, know without scientific investigation is the physics and particularly the optics which cause rainbows. In the case of a physical object like a rainbow, a form of methodological essentialism seems to be quite practicable. The basic definition of the phenomena – what they are, how they present themselves – seems obvious. What is of interest, rather, is what 'underlies' these phenomena analytically.

Physical phenomena such as rainbows and abstract concepts such as capitalism and the morpheme are extreme points on a continuum. We can fairly easily think of concepts which are more to the middle of this continuum. Examples include any semi-technical notion for which there is, however, fairly standard agreement among native speakers of a language about what is and is not included under the category concerned. An example might be the notion of a 'hobby'. Native speakers of English have a fairly good idea of what is and is not a hobby. It would be fairly perverse for a researcher to insist on a definition of 'hobby' which was different from that generally accepted by native speakers of English. If, however, there

[3] It might also be argued that it is also epistemologically naïve actively to deny the possibility that there is in fact a reality 'out there' which it is the task of researchers simply to discover. Between the 'absolute realist' and the 'instrumentalist' views, there is a more sophisticated overall perspective which accepts that we can never ultimately know whether what we describe as researchers is reality-as-it-really-is or a version of reality which presents itself as a result of the theory which we use to investigate that reality (cf. Mulder and Rastall 2005).

was no absolutely clear definition of what is and what is not a hobby among native speakers of English, a researcher could provide such a definition, basing themselves on the general views of native speakers, but adding specific criteria of differentiation (e.g. between a hobby and a sport) where this seemed to be necessary for clarity of definition.

In this article, I will consider terms of all three kinds discussed above: 1. Technical terms (of the 'capitalism' or 'morpheme' type), where it makes sense to provide a definition; 2. Semi-technical terms (of the 'hobby' type), where it makes sense to follow general usage, only introducing a new definitional element at the margins, to make plain precisely what is meant by the term in question; 3. Non-technical terms (of the 'rainbow' type), where there is clear existing general agreement about what the term refers to, and where there it would not be appropriate to try and provide a separate definition from this (what Lyons 1991: 32 has termed 'everyday metalanguage'). This division into three groups of terms is, of course, itself somewhat arbitrary. As noted above, there is, in fact, a continuum, such that terms may be more or less technical, or more or less 'everyday'. However, the division into three broad groups seems useful for practical purposes, provided that we remember that this division is a matter of convenience, and that boundaries between these types are, in reality, fuzzy (for further discussion of fuzzy boundaries, see Section 11).

The following are the central terms of these three different groups which I will consider in this article:

1. Technical terms, used in linguistics: collocations, formulaic sequences, multiword expressions;
2. Semi-technical terms, used in grammar teaching, etc: compounds, phrasal verbs;
3. Non-technical terms, in everyday usage (everyday metalanguage): idioms and proverbs.

As noted, these different groups of term require rather different treatment. Group 1, technical terms, are already defined in the literature in different and very often incompatible ways (and the notions they refer to also have alternative terms in some works). This makes it perfectly reasonable – and even necessary – to define them in specific ways for the purposes of this article, accepting that these definitions will necessarily not be compatible with all other definitions given in the literature. Such 'redefinition' (definition for the purposes of this article, and the underlying arguments it supports) should in practice, however, not involve the imposition of a meaning (definition) on the term involved which is so different from other previous definitions that it is likely to confuse readers who have already encountered these previous definitions. A fortiori, such redefinition should not be so far from previous definitions, that it appears a perverse usage of the term in question.

Group 2, semi-technical terms, already have fairly clear and fairly compatible definitions in the literature. Any 'redefinition' in these cases should, if it is to be sensible, only involve clarifying which of minor existing definitional differences this article adopts.

Group 3, non-technical terms, are better not redefined. It is possible to use an existing non-technical everyday term in a new technical sense intended to provide a more precise definition of what the non-technical term refers to. However, there are two points with this.

The first is a general problem that readers are in practice likely to confuse the new technical definition with the existing non-technical one, possibly even where the fact that the

term is being used in a specific technical sense is made plain in the article. The second point – and one which is clearly germane to this article – is that in using an existing non-technical everyday term, one may be intending to define, using technical notions, what is meant by this term in everyday language. This is the case in this article, which in its definitions of 'idiom' and 'proverb' is attempting to answer the questions: *Taking phenomena which are generally identified as idioms, how can we characterise/define idioms in technical linguistic terms?*; and *Taking phenomena which are generally identified as proverbs, how can we characterise/define proverbs in technical linguistic terms?* This is very different from Group 1, technical terms, where we are defining the terms for the purposes of this article, sometimes in ways which are clearly incompatible with the definitions of other writers. It is also different from Group 2, semi-technical terms, where we have at least some freedom to redefine terms.

2. Collocations

There are many different definitions of 'collocation' (cf. Firth 1957: 195; Cowie 1978: 132; Hausmann 1984; Richards, Platt and Webber 1985: 46; Sinclair 1991: 170; Kilgariff 1992: 29; Bahns and Eldaw 1993; Palmer 1993; Herbst 1996: 380; Hill 2000: 51; Lewis 2000: 132; Bartsch 2004: 68; Nesselhauf 2005; Seretan, 2011: 13. For summaries of these, see Bartsch Cai 2017: 4–7; Ruiz Yepes 2017: 11–18). In this article, the term 'collocation' will be used in a sense which is fairly standard in corpus linguistics to mean "a sequence of words or terms that co-occur more often than would be expected by chance within the context of a specific word" (Gómez 2009: 149).

Collocation in this sense is further explained by Lecheka (2015: 2):

> The strength of this kind of attraction between words can be measured through the statistical analysis of corpus data. The purpose of these statistical calculations is to find word pairs with significantly more co-occurrences than what would be expected by chance, given the words' total frequencies in the data. Thus, we can establish the most significant *collocates* of any given word in the language variety that the data represents […]

I will consider the inclusion of the word 'significantly' in Lecheka's characterisation of collocation below in this section. Before that, however, I will make two points in relation to Gómez's definition. Firstly, this definition does not impose a 'span' (or 'window'), i.e. a maximum number of words between a 'node' word (i.e. word of focal interest), and the other word(s) involved in the collocation. Secondly, it does not impose on a collocation that it should have any syntactic 'coherence', i.e. that the words involved in it should form some kind of syntactic unit, or sub-element within a syntactic unit. Thus, according to Gómez, words which occur at any distance apart from one another are technically collocations, provided they co-occur with greater frequency than would be expected by chance, given these words' total frequencies in the data. Similarly, while a form with syntactic coherence, such the noun phrase 'strong tea', involves a collocation of 'strong' and 'tea' under this definition, so does 'weak' and 'tea' in a form such as 'Do me some tea, but don't make it too weak'.

To operationalise 'collocation', i.e. to apply it in practice in a particular piece of research, it is necessary to do two things. The first is to specify a span (specific number of words before or after the node word). If one wanted a term to describe a collocation of this type, it could be called a 'span-defined collocation'. The second necessary step is to identify only those collocations which co-occur with *significantly* greater frequency than would be expected by chance, given these words' total frequencies in the data. This eliminates word co-occurrences which are technically collocations, but whose tendency to co-occur seems only insignificantly greater than would be predicted by their overall frequency of occurrence. In order to achieve this it is necessary to decide statistically what constitutes a significant (as opposed to insignificant) greater-than-chance frequency of word co-occurrence – a figure which would no doubt vary from study to study, depending on the specific goals of the study in question. A collocation of this type could be termed a 'statistically significant collocation' (or just a 'significant collocation'). A collocation which is both statistically significant and defined in terms of span could be called a 'span-defined (statistically) significant collocation'. There are different statistical approaches to collocations, which can yield quite different statistical significance results. A consideration of these falls outside the scope of this article (for discussion, see Evert 2007; Gries 2013).

3. Formulaic sequences

Various terms are used in roughly the same sense as 'formulaic sequence', e.g. 'formulaic expression', 'formulaic language' and 'prefab'; and various definitions have been given for these terms (cf. O'Donnell, Römer and Ellis 2013; Possio 2015: 60). Probably the best known definition is that of Wray and Perkins (2000: 1; also Wray 2002: 9), who define a formulaic sequence as "a sequence, continuous or discontinuous, of words or other elements which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use".

In this article, I shall adopt the following definition of formulaic sequence:

> A formulaic sequence is a collocation, whether continuous or discontinuous, which has syntactic coherence.

I propose this definition in preference to that of Wray and Perkins, in order to remove their notion of 'prefabricated', which has a technical psycholinguistic orientation, and could only be determined – if at all – by detailed psycholinguistic investigation. This is quite different from the current corpus-oriented approach, since corpora cannot directly tell us anything about what is stored and retrieved whole from memory.

Formulaic sequences which have a very high statistical occurrence are likely to occur in standardised contexts. Examples of such formulaic sequences are 'ladies and gentleman' (typically used at the beginning of speeches; cf. Mollin 2014: 149–151), and 'And they all lived happily ever after' (the traditional formula for ending a fairy story in English).

It might be argued that the current definition of 'formulaic sequence' is too broad for what is typically meant by 'formulaic', and that 'formulaic' implies, perhaps amongst other things, a high frequency of usage. If this were felt to be the case, it would be possible to add a further

criterion of statistical frequency to the definition of 'formulaic sequence' along the following lines:

> A formulaic sequence is a collocation, whether continuous or discontinuous, which has syntactic coherence and occurs with statistically greater frequency than a collocation which is not a formulaic sequence.

This would require defining, no doubt on a study-by-study basis, the precise statistical frequency which a syntactically coherent collocation would need to have in order for it to be (also) a formulaic sequence. The relationship between collocations and formulaic sequences can be represented as in Figure 1, which indicates that formulaic sequences are a subset of collocations; i.e. in linguistic-semantic terms, 'collocation' as defined in this article is a hyperonym (superordinate) of 'formulaic sequence'.

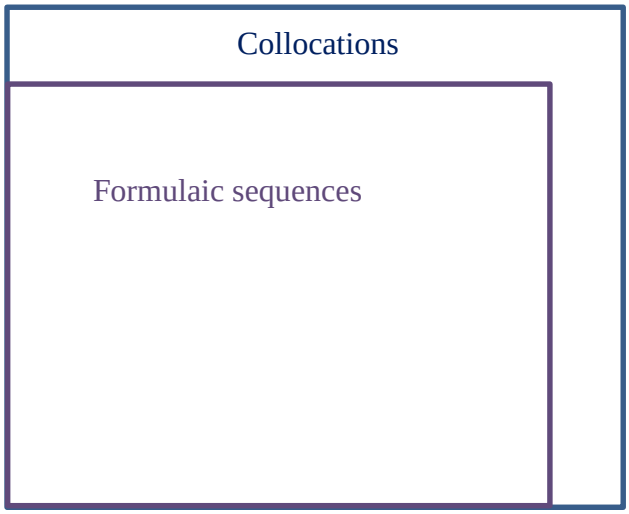

Figure 1: Semantic relationship between collocations and formulaic sequences

In addition to the criteria of syntactic coherence, and possibly 'significant frequency', used for defining 'formulaic sequence' above, we could also choose to regard 'formulaic sequence' rather differently – as a Group 2, semi-technical, term, or even a kind of Group 1, non-technical, term. In this case, we could classify collocations as non-formulaic or formulaic on the basis of native-speaker judgements, rather than on a syntactic (and also perhaps a statistical) basis. This proposed solution, of course, rests on native-speakers being able to make such judgements – and for the judgements which are made by different native speakers to be sufficiently similar that the results have an acceptably high degree of intersubjective acceptance across large numbers of native speakers. (This also raises the question of what constitutes an 'acceptably' high degree of intersubjective acceptance.)

A final redefinition of 'formulaic sequence' might combine aspects of a syntactic plus statistical definition with a native-speaker-judgement, definition. Thus, we could use statistical frequencies to determine 'potential' cases of formulaic sequences, and then use native-speaker judgements to decide whether identified potential cases are in fact to be regarded as cases of formulaic sequences. Having presented these three possible definitions of 'formulaic sequence', I will leave it open which one is the best to adopt.

4. Multiword expressions

The term 'multiword expression' is widely used, and has been defined in different ways. The following, taken from Constant et al. (2017: 840), are illustrative examples of different, and in some cases clearly incompatible, definitions: "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Carpuat and Diab 2010: 242); "a sequence of words that acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002: 1934); "idiosyncratic interpretations that cross word boundaries" (Sag et al. 2002: 32); "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin and Kim 2010: 268). This last definition, with its use of the notion 'decomposed' draws explicitly on the notion of compositionality, i.e. the situation in which "the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them" (Wang 2018: 1), plus, we can add, the semantic correlates of these rules.

In this article I shall define a multiword expression, using the notion of compositionality, as follows:

> A multiword expression (MWE) is an expression consisting of two or more words, which is either: *Type 1*: fully non-compositional, i.e. none of the words has an independent sense; or *Type 2*: in which at least one of the words has a sense which is independent but is only found in the context of this expression; or *Type 3*: in which at least one of the words has a sense which is independent but is only found in definable limited contexts of which this context is one.[4]

It is possible to have a multiword expression which combines both Type 2 and Type 3, i.e. in which at least one of the words has a sense which is independent but this sense is only found in the context of this expression, *and* in which at least one of the words has a sense which is independent but only in definable limited contexts of which this context is one. Such an expression can be termed a *Type 2+3* multiword expression.

With regard to each constituent word of a multiword expression, this means that it may or may not have an independent sense, and that if it does have an independent sense this sense may be found in one context only, or in specific (i.e. definable) limited contexts, or in unlimited contexts. What is meant by 'context' here is more precisely termed 'lexical context', i.e. the context of another word-sense (another word in a specific sense). What is meant by 'unlimited context' is that there is no limit to the words (in particular senses; i.e. word-senses) in the context of which the word in question (in the particular sense in question) may be found. This does not mean that there is no limit to the meanings of other words in whose context the word (in the relevant sense) in question may standardly occur. Thus, the word 'court-martial' as a verb meaning 'try by court martial' (Oxford English Dictionary Online; henceforth OEDO); cf. 'court martial' as a noun meaning 'judicial court, consisting of military or naval officers, for the trial of military or naval offences, or the administration of martial law': OEDO) standardly has to have a word referring to a military institution as its

[4] For a theoretically explicit discussion of what is meant by 'independent sense' from the perspective of extended axiomatic functionalism, which I adopt elsewhere, see Dickins (1998: 241–244); see also the discussion of allosemic amalgamation in relation to morphology in Dickins (2006: 165, 188–189).

subject, whether this be a noun such as '[The, etc.] regiment', '[the, etc.] army', or '[the, etc.] military authorities', or a pronoun (e.g. 'he', 'they') which is co-referential with such a noun, or a proper noun referring to a person who holds an appropriate military rank (e.g. 'Kitchener'). These restrictions do not, however, constitute limited (lexical) contexts, since it is also perfectly possible to have other words as the subject of 'court-martial' which fall outside these categories. Thus, "Your cat court-martialled them" is a perfectly possible English sentence, however bizarre its meaning; and cf. the perfectly semantically reasonable "Your cat did not court-martial them, because cats cannot hold a relevant military rank". This contrasts with true limited (lexical) contexts where only specific words in specific senses are possible in the context of the given word in its specific sense, as discussed in this section below.

The situation with regard to the semantic independence of constituents in given (lexical) contexts can be diagrammatised as in Figure 2.

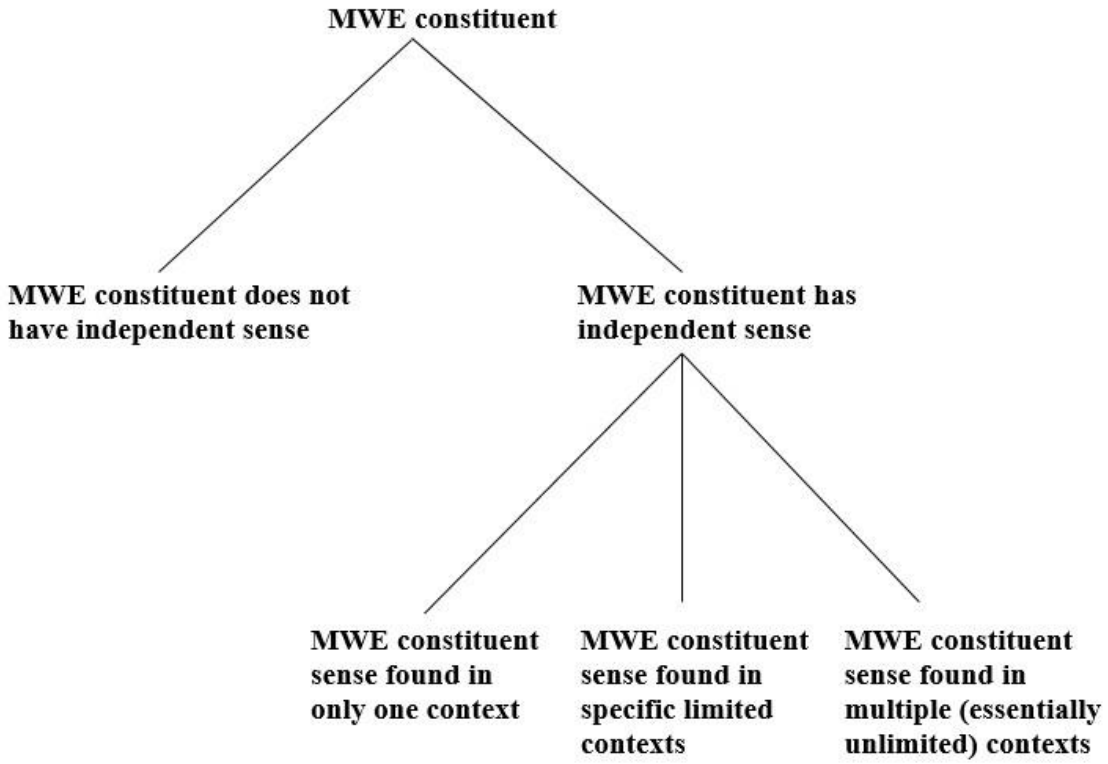

Figure 2: Typology of MWEs according to semantic independence of constituents

Semantic independence is closely related to the principle of compositionality (this section, above). In a compositional expression such as 'kick the football' the meaning of the whole expression is determined by the meaning of its individual expressions – i.e. the meaning of 'kick', the meaning of 'the' and the meaning of 'football' (in the relevant sense of all these words), plus the rules used to combine them: e.g. the fact that 'the football' is the object of 'kick', and that 'the' and 'football' together combine to make a noun phrase, and the semantic correlates of these 'rules' (syntactic relations). Each constituent in a phrase like 'kick the football' (i.e. 'kick', 'the' and 'football') can be said to have a 'free-compositional sense'; i.e. the sense which the constituent has in the complex expression in question, it also has in potentially unlimited other expressions. By contrast, a constituent having a sense which is

found in one context only or in specific limited contexts can be said to have a 'bound-compositional sense'.[5]

Probably most expressions in natural language are fully free-compositional (i.e. each of the words which makes them up has an independent sense). Multiword expressions, as defined in this article, are not. Consider the multiword expression (which is, of course, also an idiom; Section 7) 'kick the bucket', meaning 'die'. Here, it is impossible to say what sense each of 'kick', 'the' and 'bucket' has – because they do not in fact have separate (independent) senses. All we can say is that the entire phrase means 'die'. 'Kick the bucket' in the sense of 'die' is non-compositional. In fact, because none of the constituent words in 'kick the bucket' (= 'die') has a separate, independent sense, the expression is *fully non-compositional*, i.e. it is an example of a Type 1 multiword expression in terms of the definition of multiword expression given at the start of this section. The fact that none of the constituents of 'kick the bucket' has a separate, independent sense is also reflected in the fact that it is not possible to manipulate the multiword expression in any way: we cannot, for instance, say 'The bucket was kicked', or 'Don't go kicking any buckets, please'.[6]

In terms of the semantic independence of the constituents which make it up, 'kick', 'the' and 'bucket', 'kick the bucket' can be analysed as in Figure 3 (next page). Consider now, by contrast with 'kick the bucket', the multiword expressions (also multiword compounds; Section 5.1) *polar bear* meaning 'white carnivorous bear, Thalarctos maritimus […]': CEDO), and *brown bear* meaning 'large ferocious brownish bear, Ursus arctos […]': CEDO), i.e.

[5] A distinction needs to be made between a *constituent* which only occurs in one context, and a *constituent-sense* which only occurs in one context, i.e. a constituent having a sense which only occurs in one context. An example of a constituent which only occurs in one context is the morpheme 'cran', occurring only in the word 'cranberry' (a morpheme occurring in only one context being traditionally known as a 'unique morpheme', 'unique morph' or 'cranberry morpheme'; e.g. Carstairs-McCarthy 2002: 19). An example of a constituent-sense only occurring in one context is 'black' in 'blackbird' (cf. Section 5.2). A constituent only occurring in one context will either have a constituent-sense only occurring in one context or will not have an independent sense. An example of a constituent only occurring in one context having a constituent-sense only occurring in one context is 'cran' in 'cranberry'. Here the fact that 'berry' has a standard sense occurring in essentially unlimited other contexts forces us to conclude that 'cran' also has a sense here (cf. the analysis of 'blackbird' in Section 5.2), and since the constituent 'cran' only occurs this context, its sense also only occurs in this context. Examples of constituents only occurring in one context and not having an independent sense are 'spick' and 'span' in 'spick and span' (assuming 'span' not to be the same word as 'span', whose basic sense is 'interval, space or distance between two points': CEDO). ('And' in 'spick and span' also does not have an independent sense, though it does in unlimited other contexts.)

[6] There is one interesting apparent exception to the claim that 'kick the bucket' as a multiword expression cannot be manipulated in any way. This is the usage 'kick the proverbial bucket' (with 11 results on the IWeb corpus, henceforth IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>). While 'proverbial' in this case formally goes with 'bucket', semantically, it relates to the whole phrase 'kick the bucket'. (Even more curiously, 'kick the bucket' is not a proverb, but a multiword expression, as seen, and also an idiom; Section 7.) Given the oddity of its semantics in relation to its syntax, 'proverbial' in 'kick the proverbial bucket' is not to be taken as the kind of manipulation of a multiword expression which demonstrates the semantic independence of one or more of its constituents. Other idioms similarly allow for the 'insertion' of the word 'proverbial', e.g. 'grasp the proverbial nettle'.

As Reviewer 1 has pointed out to me, it is in fact possible use forms of the type 'The bucket was kicked', or 'Don't go kicking any buckets, please' as semi-jocular 'transformations' of 'he kicked the bucket'. I have argued elsewhere that such forms fall outside the standard conventions of language, and as such are not to be considered in linguistic analysis (Dickins 1998: 324. The example I give there is 'I nearly strangled it', in 'Grasp the nettle!? – I nearly strangled it!').

both ‘polar bear’ and ‘brown bear’ are species of bear. It is possible to say things in English like ‘polar and brown bears’ (15 results on IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>), or equally ‘brown and polar bears’ (20 results on IWeb: 24.9.18: <https://corpus.byu.edu/iweb/>). This demonstrates that ‘bear’ in both ‘polar bear’ and ‘brown bear’ has an independent sense and that this sense is the same in both compounds – the acceptability of forms such as ‘polar and brown bears’ and ‘brown and polar bears’ derives from the fact that ‘bear’ has the same sense in both ‘polar bear’ and ‘brown bear’ (on the basis of the general principles outlined in Cruse 1986: 49–83 for detailed argument of principles in this regard, some problematic cases notwithstanding). This is the sense which ‘bear’ has in multiple (unlimited) other contexts, ‘heavily-built, thick-furred plantigrade quadruped, of the genus *Ursus*’ (OEDO).

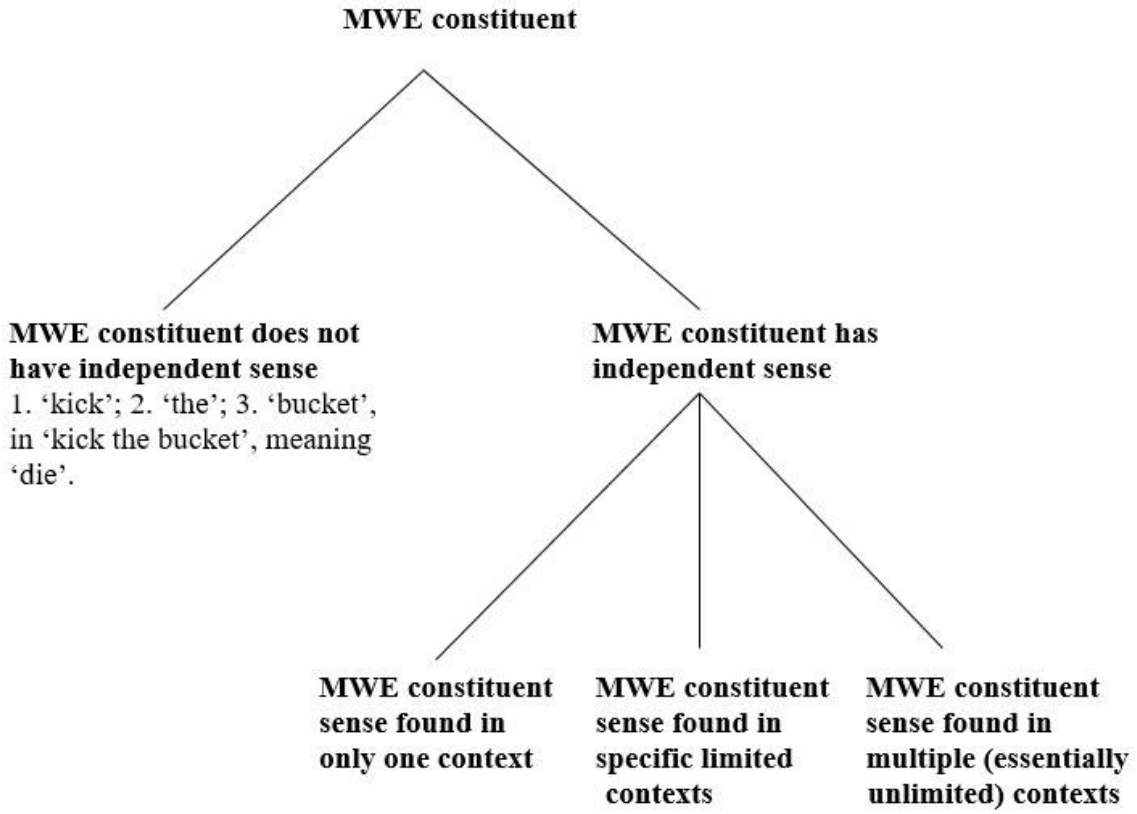

Figure 3: Analysis of the multiword expression ‘kick the bucket’ in terms of the semantic independence of their constituents

This, accordingly, is a Type 2 multiword expression, in terms of the definition of multiword expressions given at the start of this section, i.e. a multiword expression in which at least one of the constituent words has a sense which is independent but is only found in the context of this expression (in the case of ‘polar bear’, this word is ‘polar’, while in ‘brown bear’ it is ‘brown’).[7]

[7] I take it that in cases such as “bears/animals/those [etc.] of the ‘polar’ and ‘brown’ varieties”, the forms ‘polar’ and ‘brown’, with inverted commas around them, are correct; i.e. that this is a case of mention, rather than use, and that a form of this kind is not therefore a counterexample to the claim that ‘polar’ and ‘brown’ only occur in the contexts ‘polar bear’ and ‘brown bear’.

The conclusion that ‘bear’ in both ‘polar bear’ and ‘brown bear’ has the sense it has in multiple (unlimited) other contexts, ‘heavily-built, thick-furred plantigrade quadruped, of the genus *Ursus*’ requires us also to conclude that in ‘polar bear’, ‘polar’ (like the entire compound ‘polar bear’) has the sense ‘white carnivorous bear, Thalarctos maritimus […]’. It correspondingly requires us to conclude that in ‘brown bear’, ‘brown’ (like the entire compound ‘brown bear’) has the sense ‘large ferocious brownish bear, Ursus arctos […]’. ‘Bear’ is a hyperonym of ‘polar’ (also ‘polar bear’) in ‘polar bear’, and ‘bear’ is similarly a hyperonym of ‘brown’ (also ‘brown bear’) in ‘brown bear’.

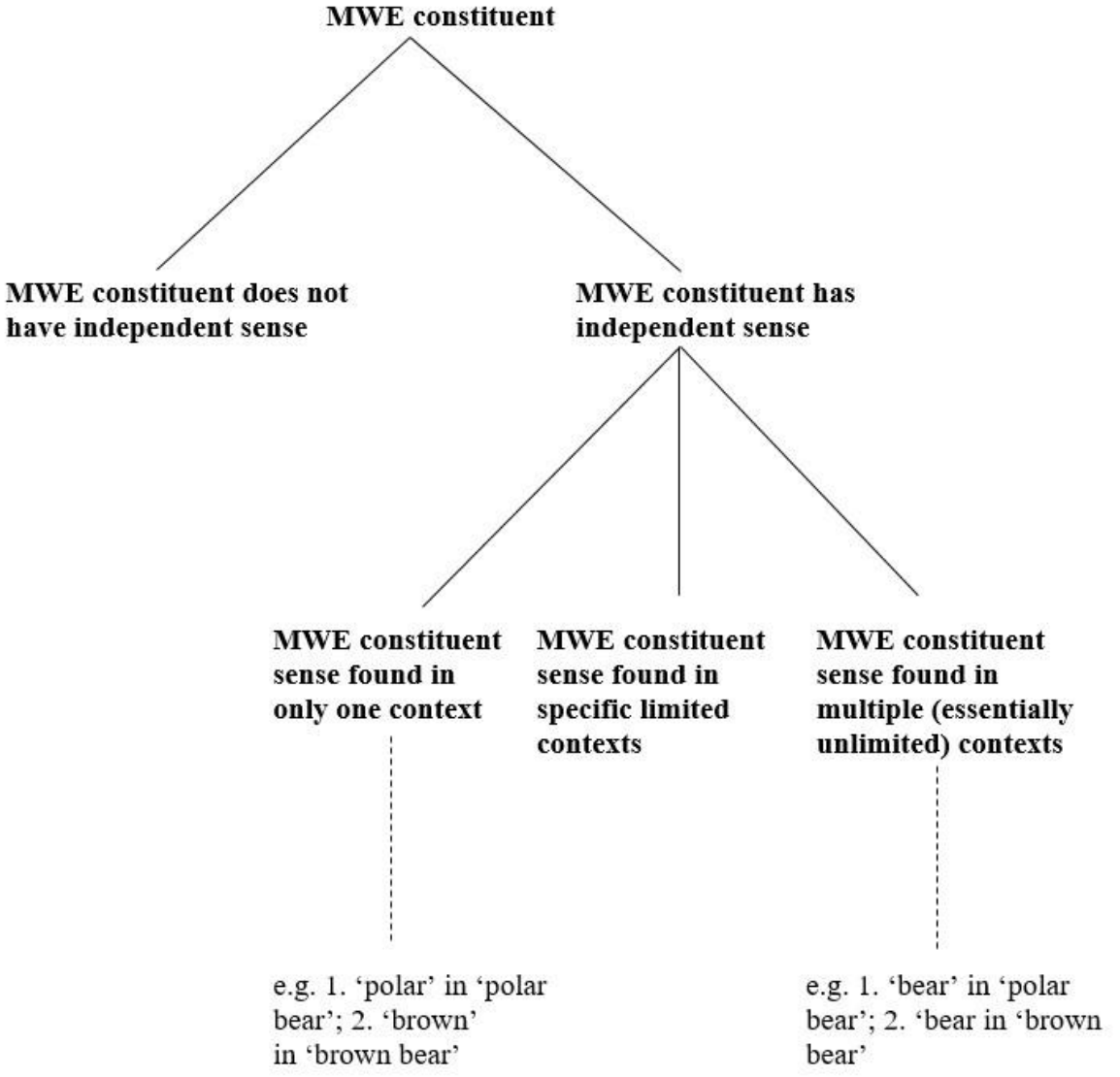

Figure 4: Analysis of the multiword expressions ‘polar bear’ and ‘brown bear’ in terms of the semantic independence of their constituents

This conclusion is supported by a consideration of the semantics of ‘grizzly bear’, i.e. ‘a variety of the brown bear, Ursus arctos horribilis’. The fact that ‘grizzly’ here means the same as it does in the entire compound ‘grizzly bear’ is shown by the fact that we can say ‘grizzly and polar bears’ (17 results on IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>), i.e. by the same procedure which was used above to show that bear’ in both ‘polar bear’ and

'brown bear' has the same sense that it does in each of those compounds. In the case of 'grizzly bear', however (unlike that of 'polar bear' and 'brown bear'), this conclusion is secondarily demonstrated by the fact that it is possible to use 'grizzly' on its own, without a following 'bear', as a noun to mean the same as 'grizzly bear', e.g. 'I've just seen a grizzly'.

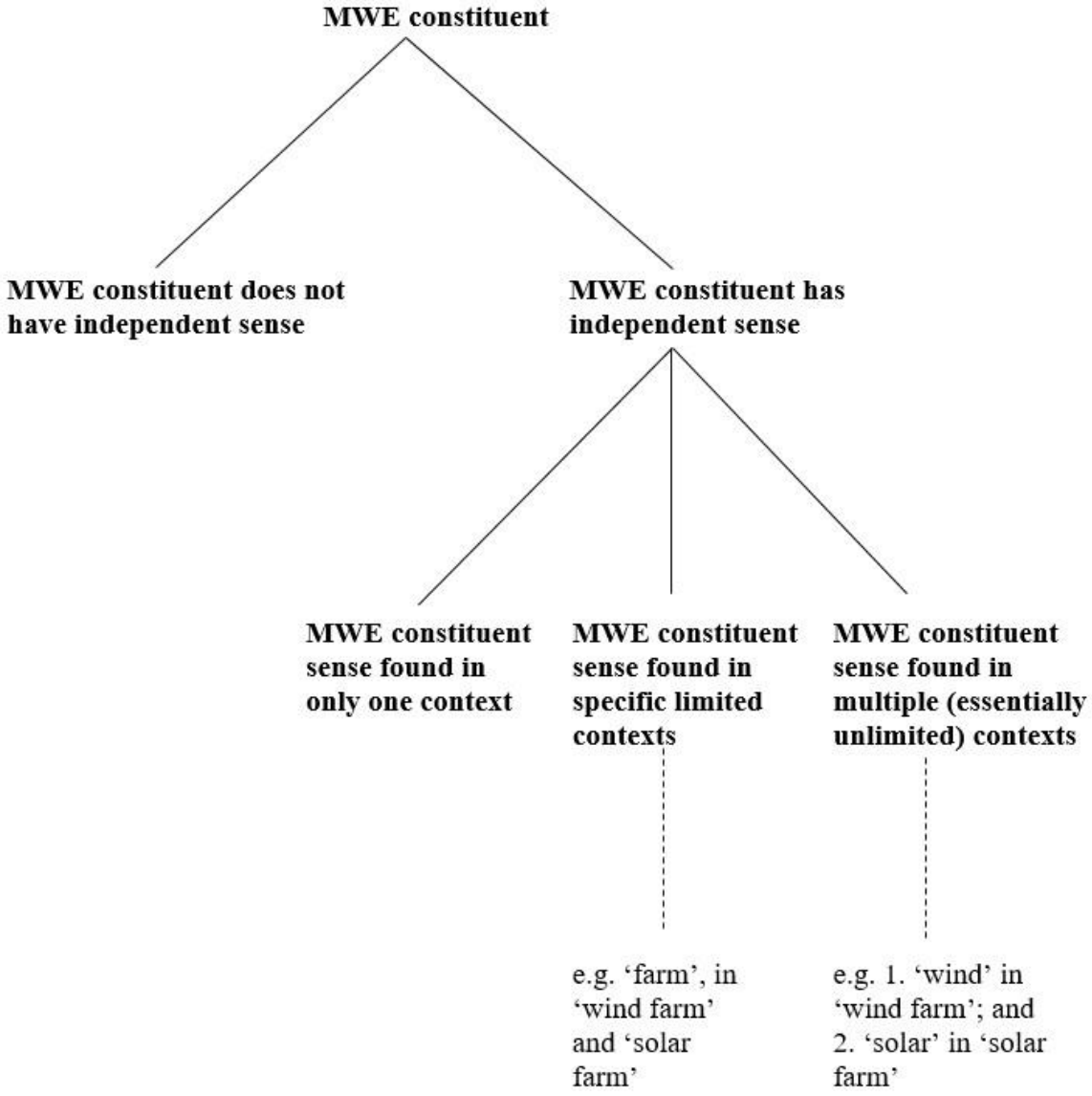

Figure 5: Analysis of the multiword expressions 'wind farm' and 'solar farm' in terms of the semantic independence of their constituents

In terms of the semantic independence of the constituents which make them up 'polar bear' and 'brown bear' can be analysed as in Figure 4 (previous page).

Examples of Type 3 multiword expressions, in which at least one of the words has a sense which is independent but only in specific limited contexts, are 'wind farm' and 'solar farm' (more precisely, these are multiword compounds: Section 5.1). Here, 'farm' has an independent sense, as can be seen from the fact that it is possible to things like 'wind and solar farms' (242 results on IWeb: 24.9.18: <https://corpus.byu.edu/iweb/>) and 'solar and wind farms' (101 results on IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>). However, a form

such as ‘farms for wind energy’ does not appear possible (no results on IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>). ‘Farm’ is thus only found in specific limited contexts in the (putative) sense ‘array of machinery for producing energy from a …. source’ (or similar), whereas ‘wind’ in ‘wind farm’ and ‘solar in ‘solar farm’ have a sense which is found in multiple (essentially unlimited) contexts.

In terms of the semantic independence of the constituents which make them up, ‘wind farm’ and ‘solar farm’ can be analysed as in Figure 5 (previous page).

As noted above, the definition which I have adopted in this article for multiword expression is different from that adopted by a number of other writers – and in fact rather narrower than that adopted by some. In particular, many writers, following Sag et al. (2002), extend the notion of ‘multiword expression’ to include expressions with only pragmatic and even statistical idiosyncrasies. Under the definitions adopted in this article, such expressions are analysed as collocations or formulaic sequences, rather than multiword expressions.

The relationship between collocations, formulaic sequences and multiword expressions can be represented as in Figure 6. This indicates that multiword expressions are a subset of formulaic sequences,[8] formulaic sequences themselves being, as noted (Section 3), a subset of collocations. In linguistic-semantic terms, ‘collocation’ as defined in this article is thus a hyperonym (superordinate) of ‘formulaic sequence’, which is a hyperonym (superordinate) of ‘multiword expression’.

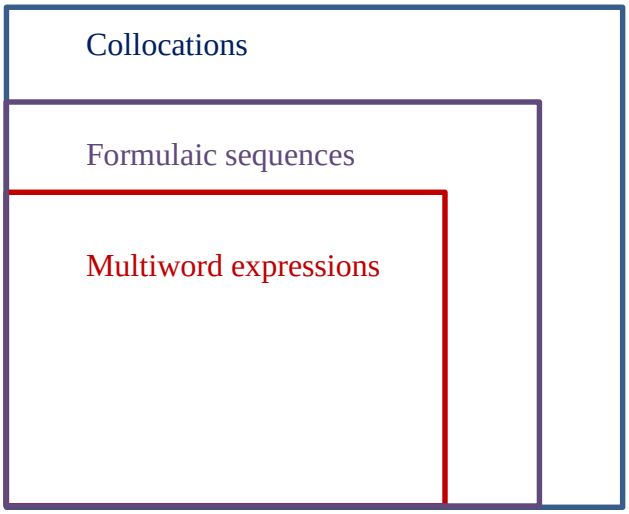

Figure 6: Semantic relationship between collocations, formulaic sequences and multiword expressions

5. Compounds

Compounds are expressions which may consist of one word, or two or more words and which contain at least one bound-compositional constituent (whether this occurs in only one context, or in specific limited contexts). A compound can be defined as follows:

[8] It would not be possible to regard multiword expressions as a subset of formulaic sequences if we were to include as part of the definition of the latter a relationship to specific situations or types of discourse. This is partly why I have chosen not to define ‘formulaic sequence’ in relation to situation/type of discourse in Section 3.

> A compound is a grammatical entity which consists of two or more elements each of which can appear as a separate word (in other contexts), and which is not fully free-compositional.

This definition is fairly compatible with standard definitions of 'compound'; e.g. "'the formation of a new lexeme by adjoining two or more lexemes" (Bauer 2003: 40), "A term used widely in descriptive linguistic studies to refer to a linguistic unit which is composed of elements that function independently in other circumstances" (Crystal 2008: 96).

Multiword compounds, i.e. compounds which consist of more than one word, are a type of multiword expression, while single-word compounds (compounds consisting of only one word) are not. Both multiword compounds and single-word compounds can be analysed in terms of the semantic independence of their constituents.

5.1 Multiword compounds

A multiword compound can be defined as follows:

> A multiword compound is a grammatical entity which is written (orthographically) as two or more words and consists of two or more elements, each of which can also appear as a separate word (in other contexts), and which is not fully free-compositional.

Consider in this respect, the examples 'sleeping policeman' meaning 'ramp in the road intended to jolt a moving motor vehicle, thereby encouraging motorists to reduce their speed' (OEDO),[9] 'wind farm', 'solar farm', 'polar bear' and 'brown bear' (these last four already discussed in Section 4, under multiword expressions). In terms of the semantic independence of their constituents, these examples can be analysed as in Figure 7 (next page).

As 'polar bear', 'brown bear', 'wind farm' and 'solar farm' have already been discussed in Section 4, in this section I will only consider 'sleeping policeman'.

[9] As Reviewer 2 has pointed out to me, there is an issue with how we determine that *sleeping* in *sleeping policeman* is a separate word, occurring also in other contexts, rather than something that just happens to be phonologically and orthographically the same as *sleeping* in *sleeping student* (i.e. one who sleeps). The same goes for *up* in *give up* in the case of phrasal verbs, and so on. It is often assumed in corpus linguistics that we just know this. As I have argued elsewhere, however, its determination, if it is to be rigorous, relies on a conception of what constitutes a word (or similar), and therefore ultimately on a particular theory of language (linguistic theory) within which the notion 'word' (or similar) is defined. This is an issue which I have addressed from the theoretical perspective of extended axiomatic functionalism in Dickins (1998: 227–240; also 187–198). One useful criterion (though not the only one) for claiming that two words – and by extension phrases – are the same is the presence of a figurative (e.g. metaphorical) relationship between them. This obtains, for instance, in the case of *sleeping policeman* in the sense of 'slumbering law enforcement officer' vs. *sleeping policeman* in the sense 'ramp in the road intended to jolt a moving motor vehicle, thereby encouraging motorists to reduce their speed' (cf. Dickins 2005; 2018), though it is not found with the same clarity, and in some cases not at all, in other examples discussed in this article. Given the theory-neutral orientation of this article, I assume that the analyses which I give of what do and do not consider to be words (etc.) are 'commonsensical', and leave it to the reader to decide whether they agree with these analyses or not.

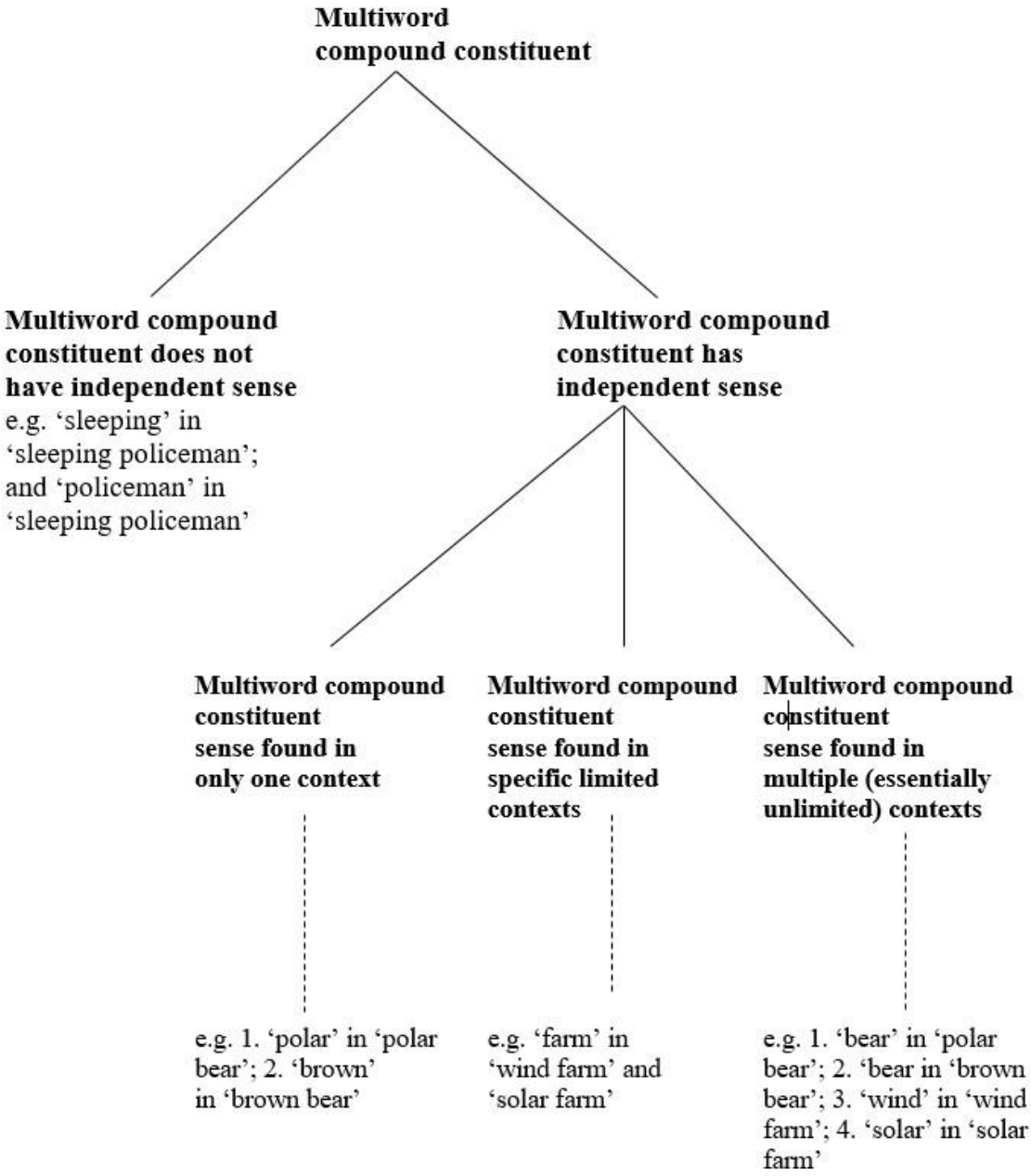

Figure 7: Analysis of multiword compounds according to semantic independence of constituents

Compounds involving constituents which do not have independent senses must logically have at least two such constituents – these two constituents (neither with an independent sense) together forming a larger constituent which does have independent sense. Thus the entire compound 'sleeping policeman' has an independent sense 'ramp in the road intended to jolt a moving motor vehicle [etc.]' (OEDO), but neither of the constituents which make it up, 'sleeping', or 'policeman', has an independent sense.[10]

[10] It is essential to distinguish between compositionality and semantic motivation. While neither of the elements 'sleeping' and 'policeman' in 'sleeping policeman' are semantically independent (the compound is semantically non-compositional), there is a clear metaphorical semantic motivation to the term 'sleeping policeman'.

5.2 Single-word compounds

A single-word compound can be defined, as follows:

> A single-word compound is a grammatical entity which is written (orthographically) as a single word and consists of two or more elements each of which can appear as a separate word (in other contexts), and which is not fully free-compositional.

Single-word compounds are by definition not a sub-type of multiword expression. In terms of the semantic independence of their constituents, however, single-word compounds can be analysed the same way as multiword expressions. This is illustrated in Figure 8 (next page) in relation to the following single-word compounds: 'ladybird' i.e. 'any of various small brightly coloured beetles of the family Coccinellidae' (CEDO), 'blackbird' 'common Eurasian thrush, Turdus merula, of which the male has black plumage and a yellow bill and is noted for its melodious song, and the female is dark brown' (OEDO), and 'trustful' and respectful'.[11]

'Ladybird' provides an example of a single-word compound in which the constituents do not have an independent sense. Like multiword compounds, single-word compounds involving constituents which do not have independent senses must logically have at least two such constituents – these two constituents (neither with an independent sense) together forming a larger constituent which does have independent sense. Thus the entire compound 'ladybird' has an independent sense 'Any of numerous small, domed beetles of the family Coccinellidae', but neither of the constituents (morphemes) which make it up, 'lady', or 'bird', does.

'Blackbird' provides an example of a second type of single-word compound constituent – one which has an independent sense, but this sense is only found in the context of this compound. Paralleling the analyses in Section 5.1 of 'polar' (in 'polar bear') and 'brown' (in 'brown bear'), in 'blackbird', both 'black' and the entire compound 'blackbird' have to be regarded as having the sense 'common Eurasian thrush, Turdus merula [etc.]'. I will discuss the ramifications of this analysis further below. For the moment, we should note, however, that 'black' in 'blackbird' does not have the sense "Of the darkest colour possible, that of soot, coal, the sky on a moonless night in open country, or a small hole in a hollow object; designating this colour; (also) so near this as to have no recognizable colour, very dark" (OEDO); i.e. 'black' in 'blackbird' does not have the standard colour sense of 'black'. This can be seen from the fact that not all blackbirds are black; in fact the female is brown (e.g.

Just as a policeman may do, a 'sleeping policeman' controls traffic speed, and like someone who is sleeping a 'sleeping policeman' does not actively intervene. Many of the examples analysed in this article (e.g. 'polar bear', 'brown bear', 'blackbird', 'wind farm', 'solar farm', are semantically motivated, in that they can be seen to be figuratively related (typically metaphorically related) to other more basic senses of the elements which they are made up of. Such semantic motivation, however, is entirely independent of the analysis of the semantic compositionality of the elements which make up these words and phrases.

[11] I have categorised 'trustful' and 'respectful' as (single-word) compounds, on the basis that they both consist of elements which can, in other contexts, occur as independent words – 'trust' and 'ful(l)' in the case of 'trustful' and 'respect' and 'ful(l)'. The fact that 'full' is written with two 'l's as a full word and one in this context is to be regarded simply as an idiosyncrasy of spelling. In terms of some definitions of 'compound', 'trustful' and 'respectful' are not compounds. However they are, in terms of the definition given at the start of this section.

https://www.rspb.org.uk/birds-and-wildlife/wildlife-guides/bird-a-z/blackbird/), and an albino blackbird would be white, while a blackbird which had fallen into a can of red paint would be red.

'Trustful' and 'respectful' provide an example of a third type of single-word compound constituent – one which has an independent sense which is found in more than one context (distinguishing it from the second type of single-word compound constituent above), but only in a limited number of contexts. In 'trustful' and 'respectful', the 'ful' constituent means 'having/exhibiting/showing' (or similar). This is not what 'full' means (in any of its senses) as an independent word; and 'ful' only has this sense in a specific number of words in which it occurs. (In many words in which 'ful' occurs, it does not have this sense, e.g. 'wonderful'.)

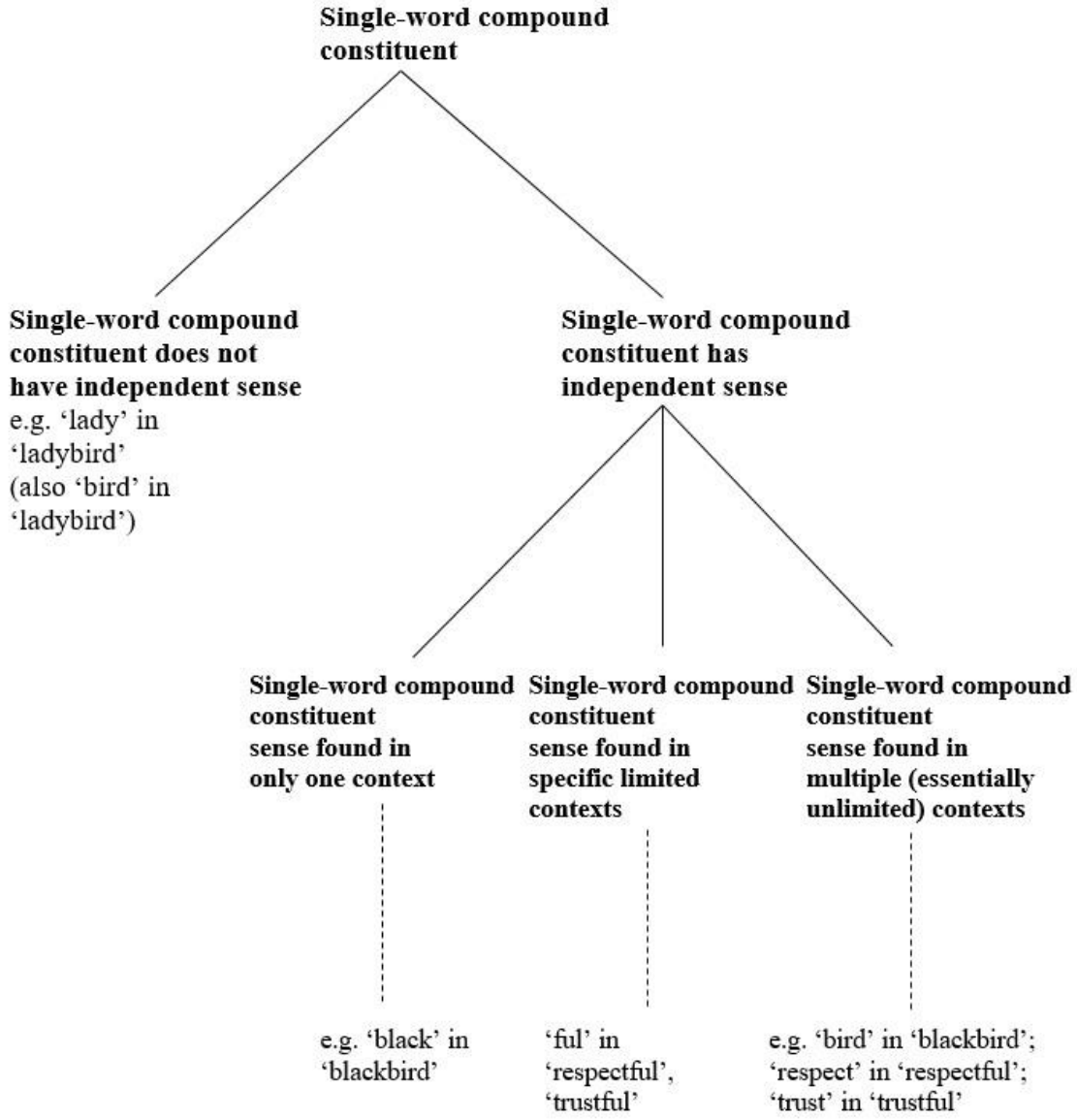

Figure 8: Analysis of single-word compounds consisting according to semantic independence of constituents

A final type of single-word compound constituent is one which has an independent sense, this sense being found in multiple (essentially unlimited) contexts. Examples are 'trust' in

'trustful' and 'respect' in 'respectful', where 'trust' and 'respect' mean what they mean (in the relevant sense) as independent words. Another, somewhat more subtle, example is 'bird', in 'blackbird'. Here, we should analyse 'bird' as having the sense 'Any feathered vertebrate animal: a member of the second class (*Aves*) of the great Vertebrate group […]' (OEDO), i.e. the same sense as it has in multiple (unlimited) other contexts. The reason for this is as follows.

Given, as argued above, that (i) the constituent (morpheme) 'black' in 'blackbird' has the same sense ('common Eurasian thrush, Turdus merula […]') as does the entire compound 'blackbird', and that (ii) 'bird' in 'blackbird' has the same sense which it has in multiple other contexts ('Any feathered vertebrate animal […]'), it follows that 'bird' in 'blackbird' is a hyperonym of 'black' in 'blackbird', as well, of course, of 'blackbird' itself. This proposed hyperonym-hyponym relationship can be represented as in Figure 9.

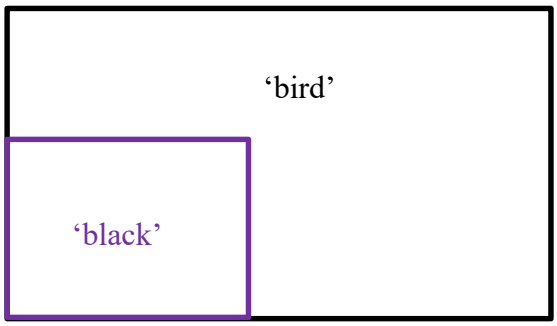

Figure 9: Hyperonym-hyponym relationship between 'bird' and 'black' (also 'blackbird') in 'blackbird'

Because 'black' (also 'blackbird') further delimits the sense of 'bird' here (as a hyponym of 'bird'), the fact that 'bird' (as the hyperonym/superordinate) has a wider sense than 'black' makes the element 'bird' irrelevant to the overall sense of 'blackbird'. This overall sense is simply defined by the sense of 'black' (a sense which, as noted, is only found in the context of 'blackbird').[12]

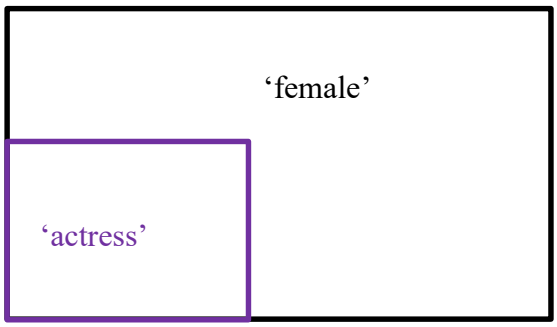

Figure 10: Hyperonym-hyponym relationship between 'female' and 'actress' in 'female actress'

[12] It should be noted that in Dickins (1998: 227–230; 433–435) I come to quite different conclusions regarding the semantic analysis of 'blackbird'. I now believe those conclusions to be wrong.

The reasonableness of this argument can be seen by considering a form such as ‘female actress’, e.g. as in ‘she’s a female actress’, where ‘female’ is a hyperonym and ‘actress’ a hyponym: all actresses are female, but not all females are actresses. The fact that ‘female (as the hyperonym) has a wider sense than ‘actress’ makes it irrelevant to the overall sense of ‘female actress’. This overall sense is simply defined by the sense of ‘actress’. This can be represented as in Figure 10 (previous page), paralleling Figure 9.

The analysis of ‘blackbird’ in this section parallels that of ‘polar bear’ and ‘brown bear’ in Section 5.1, where it was argued that ‘polar’ and ‘brown’ are hyponyms of ‘bear’, i.e. that ‘polar’ and ‘brown’ independently denote specific types of bears. The general proposal that compounds may consist of hyperonym-hyponym pairs in which the overall compound is also a hyponym of one of the constituents making up the compound was supported for multiword compounds by the analysis in Section 5.1 of ‘grizzly bear’.

Figure 11 shows the semantic relationships between collocations, formulaic sequences (formulaic language), multiword expressions, and compounds.

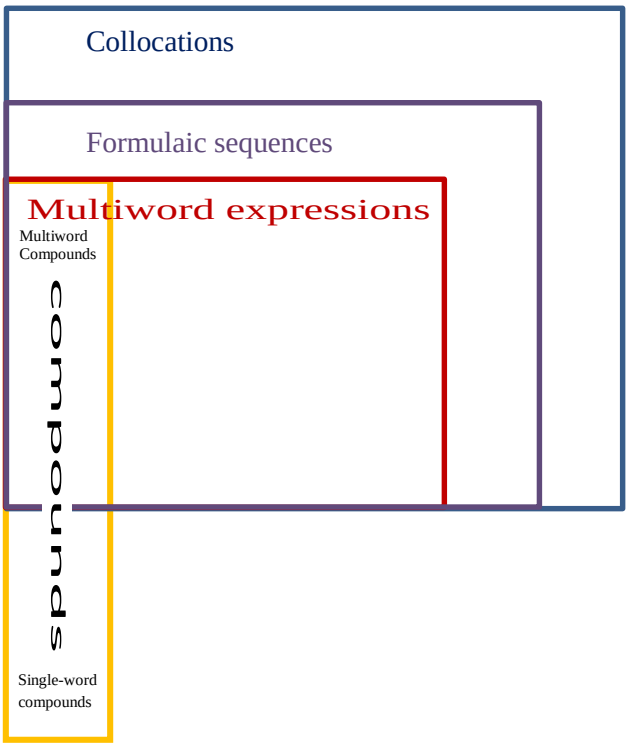

Figure 11: Semantic relationship between collocations, formulaic sequences, multiword expressions, and compounds

As noted above, while multiword compounds are a sub-type (subset) of multiword expressions, single-word compounds, by definition, are not. We should also acknowledge the fairly ad hoc nature of the distinction between multiword and single-word compounds. In many case, a compound can be written as a single word, or two words, or as two words with a hyphen between them; e.g. ‘desertsurfing’, ‘desert surfing’, and ‘desert-surfing’.

5.3 Some issues with the definition of 'multiword compound'

It might be felt that the definition of 'multiword compound' given above is too restrictive, as it excludes all forms which are fully free-compositional. There are two other alternative definitions of compound, therefore, which might be felt to be closer to standard usages of 'compound' in linguistics, which I will consider here.

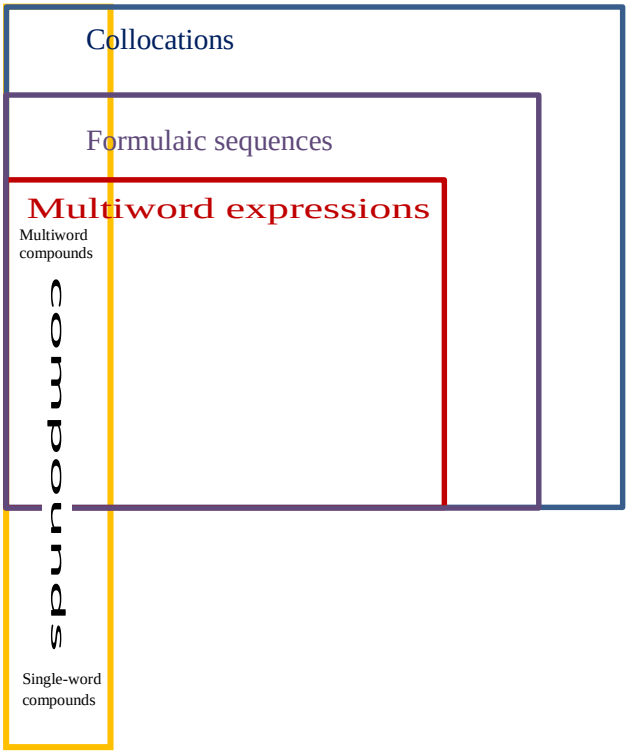

Figure 12: Semantic relationship between collocations, formulaic sequences, multiword expressions, and compounds defined to include all collocations

The first is to extend the notion of 'multiword compound' to include collocations of the grammatically appropriate kind (and therefore, by definition, also include formulaic sequences of the grammatically appropriate kind). This would give a scope for multiword compound (and compounds more generally) as in Figure 12. According to this definition, compounds would include any relevant form in which words co-occur with a greater frequency than is predicted from the overall frequency of occurrence of the words in isolation. Thus a commonly occurring collocation such as 'door key' (765 results on IWeb: 26.9.18: <https://corpus.byu.edu/iweb>) would count as a compound under this approach, notwithstanding the fact that it is fully free-compositional (cf. Section 11). By contrast a form such as 'cave key' (e.g. 'key for opening up a cave, via opening an iron door with railings which has been placed in front of it' (21 results on IWeb: 26.9.18: <https://corpus.byu.edu/iweb>), which we can take to not be a collocation, would be excluded. 'Compound' as defined in this way is represented in Figure 12. (Overall, 'door' occurs 1,561,423 on the c. 1.4 billion-word IWeb corpus, while 'cave' occurs 199,087 times; i.e. 'door' occurs around 7.84 times more frequently than 'cave'. However, 'door key' with 765 occurrences occurs around 36.43 times more frequently than 'cave key' with 21 occurrences. 15 of the 21 occurrences of 'cave

key' occur in the walkthrough for the computer game 'Turok 2: Seeds of Evil': <http://www.the-spoiler.com/ACTION/Acclaim/turok.2.1.html>).

A further extension to the definition of 'compound' would be to treat any form which is of the grammatically appropriate kind as a compound, regardless of whether it involves a collocation or not. Under this definition, 'cave key' as well as 'door key' would count as a compound. 'Compound' under this definition can be represented as in Figure 13.

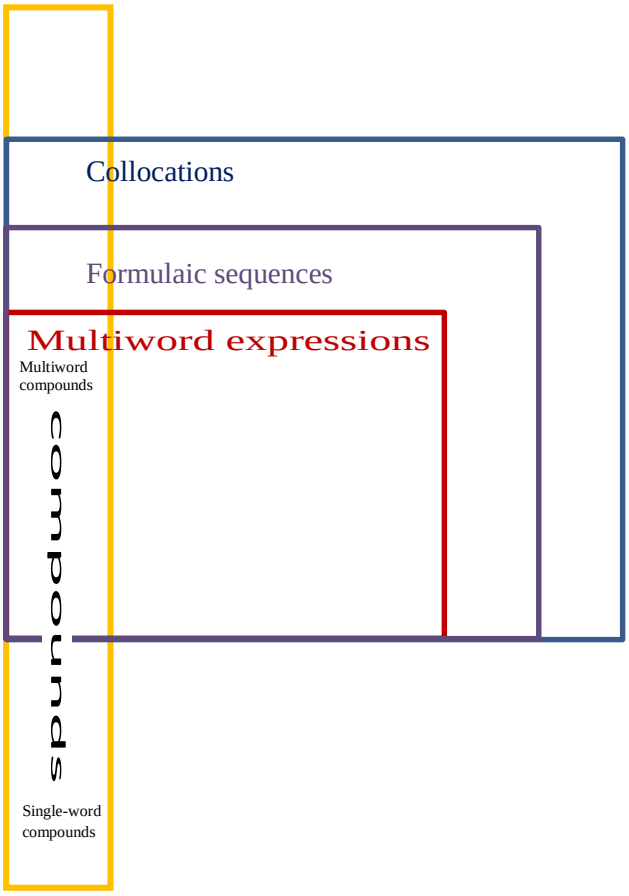

Figure 13: Semantic relationship between collocations, formulaic sequences, multiword expressions, and compounds defined to include all grammatically appropriate word co-occurrences

Under this definition, the notion 'compound' simply refers to specific set of grammatical structures. Since multiword compounds under this definition extend beyond the scope of collocations, the notion of 'compound' is not relevant to defining phenomena which fall within the scope of the notion of 'collocation', which is the basic concern of this article. Accordingly, 'compound' under the definition represented in Figure 13 is not a useful notion for this article.

There is, however, a further rider to this. If we decide in linguistics 'compound' is standardly used to describe a set of purely grammatical entities, we may be best advised to agree to use 'compound' in this way. In this case, we could use the term 'MWE-compound' to refer to what has been earlier termed a 'compound' in this article (e.g. in Figure 12). Other terms would also be available, such as 'formulaic compound' to describe any relevant form

which also falls within the category of formulaic sequence, and 'collocational compound' to describe any relevant form which also falls within the category of collocation.

Having raised these possibilities, I will in the remainder of this article continue to use 'compound' as represented in Figure 12, i.e. to refer only to appropriate grammatical structures which are also multiword expressions or not fully free-compositional single-word compounds.

6. Phrasal verbs

A phrasal verb can be defined as follows:

> A phrasal verb is a phrase that consists of a verb with a preposition or adverb or both, and is not fully free-compositional.

This is fairly close to definitions of phrasal verb used elsewhere, e.g. "a phrase that consists of a verb with a preposition or adverb or both, the meaning of which is different from the meaning of its separate parts" (Cambridge Dictionary Online). A phrasal verb, under the definition given here, is a type (sub-type) of multiword expression. Under this definition, if something is classified as a phrasal verb, it is not also to be classified as a multiword compound; i.e. 'multiword compound' and 'phrasal verb' are defined in this article as disjunct (non-overlapping) classes.

Examples of phrasal verbs are 'give up' (= 'leave off; to cease from effort, leave off trying; to stop': OEDO), 'fish out' (= 'find or extract'), 'turn on' (= 'to excite, interest, fill with enthusiasm; to intoxicate with drugs, to introduce to drugs; to arouse sexually': OEDO, i.e. the antonym of 'turn off').

Phrasal verbs can be analysed according to the semantic independence of their constituents as in Figure 14 (next page).

'Give up' is an example of a phrasal verb in which neither of the individual constituent words has an independent sense.

'Fish out', meaning 'find or extract' (as in "The supervisor's role is to fish out any old candle wicks using the small sticks if the wax is from old candle remnants": IWeb: <http://www.youthwork-practice.com/ideas-kids-crafts/Candle-Making.html>), provides an example of a second type of phrasal-verb constituent. Here 'fish' has an independent sense, but this sense is only found in the context of this phrasal verb. 'On' and 'off' in 'turn on' (= 'excite, interest, fill with enthusiasm', etc.), and 'turn off' (= 'put (a person) off, repel, disillusion, cause to lose interest', etc.) provide further examples of phrasal-verb constituents having independent senses which are only found in the context of a particular phrasal verb. The word 'turn' here can be glossed as meaning '(un)excite, (dis)interest', i.e. as subsuming semantically (being the hyperonym/superordinate of) both the active emotion of 'excite/interest' and the passive one of '*un*excite/disinterest'. The fact that 'turn' has an independent sense in 'turn on' (= 'excite, interest, fill with enthusiasm', etc.), and 'turn off' ('put (a person) off, repel, disillusion, cause to lose interest', etc.) can be seen from the acceptability of things like 'turns you on or off' (5 occurrences in the relevant sense on IWeb, 24.9.18: <https://corpus.byu.edu/iweb/>), e.g. "When you see other presenters, notice what they say

that turns you on or off. Adapt what you learn to your own presentations and style": <https://www.earlytorise.com/zen-and-the-art-of-speaking-at-seminars/>. The fact that 'turn' has an independent sense in the phrases 'turn on' and 'turn off' in the relevant senses further requires us to conclude that so do 'on' and 'off'. This, however, does not seem to be a sense which occurs elsewhere. Thus, we cannot use 'on' on its own to mean 'excited, interested', etc. or 'off' on its own to mean 'unexcited, disinterested', etc. Nor do there seem to be other multiword expressions in which 'on' and 'off' mean 'excited, interested', etc. and 'unexcited, disinterested', etc. 'On' and 'off' in 'turn on' and 'turn off', in the relevant senses, thus only occur in the independent sense which they have here in the context of 'turn'.

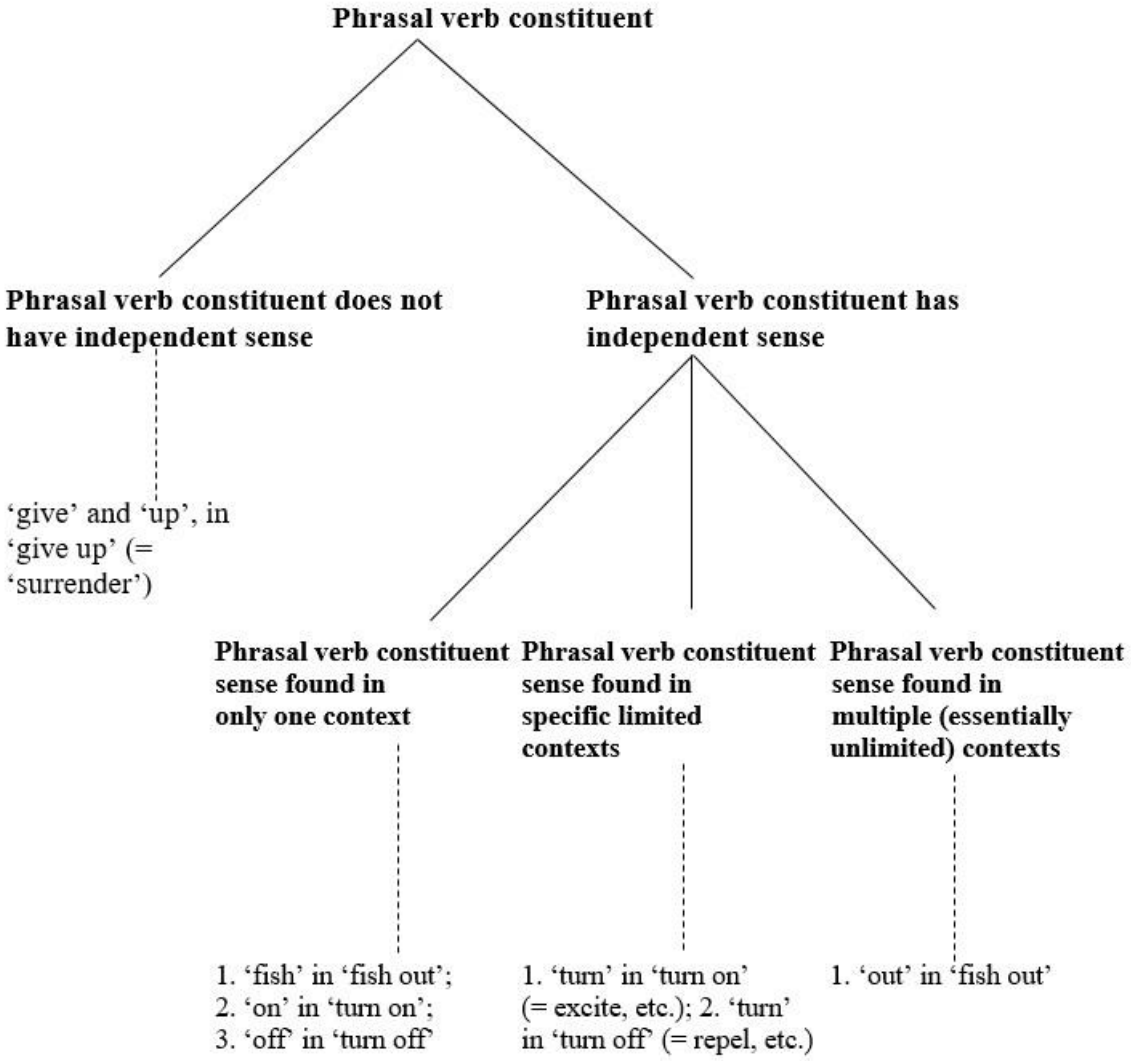

Figure 14: Analysis of phrasal verbs according to semantic independence of constituents

'Turn', meaning '(un)excite, (dis)interest', etc. in 'turn on/off' (in the relevant senses) is an example of a phrasal verb constituent having an independent sense occurring only in limited contexts. 'Turn' seems in fact to only occur with this sense in the two contexts of 'on' and 'off'. There do not seem to be any other multiword expressions in which it has the same sense.

A final type of phrasal-verb constituent is one which has an independent sense, this sense being found in multiple (essentially unlimited) contexts. An examples is 'out' in 'fish out' (= 'find, extract'). Thus, 'out' can occur in the sense which it has in 'fish out' (roughly 'in/to [the] outside of') in essentially unlimited contexts, for example with verbs such as 'walk out' (e.g. 'he walked out of the room'), 'pull' (e.g. 'she pulled the hamster out of the hole'), or even without a verb, e.g. 'out of there, please!'.[13]

Figure 15 shows the semantic relationships between collocations, formulaic sequences, multiword expressions, compounds and phrasal verbs.

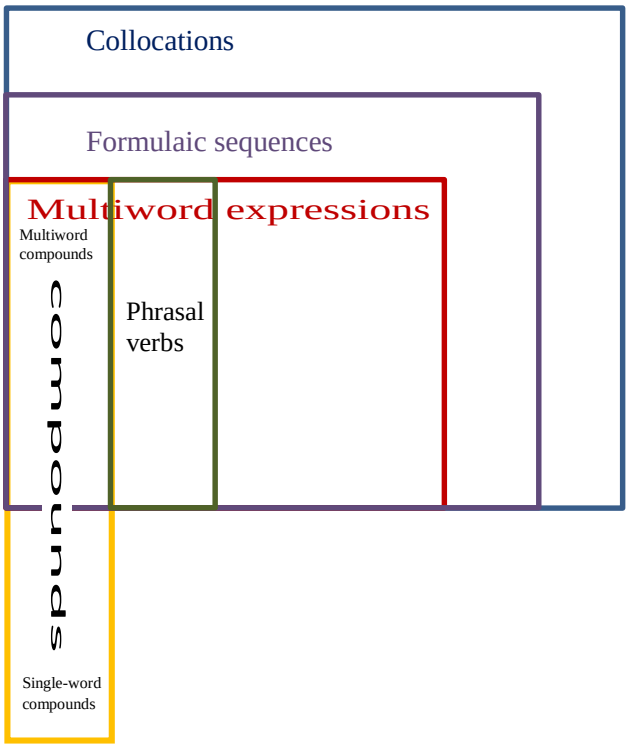

Figure 15: Semantic relationship between collocations, formulaic sequences, multiword expressions, and compounds and phrasal verbs

7. Idioms

An idiom is a particular kind of multiword expression. Precisely defining idioms has, however, proved problematic. For attempts, see Grant (2003), Grant and Bauer (2004), Liu (2008) and Wulff (2008). I will define 'idiom' as follows:

> An idiom is a phrase that is not a compound, is not a phrasal verb, is non-clausal, and is not fully free-compositional.

[13] The account of the semantics of phrasal verbs which I have given here is not complete (see also Section 1). In addition to issues of the semantic independence of the components of phrasal verbs, which involve denotative meaning, there are also issues of connotative meaning, most prominently what Hervey and Higgins term 'reflected meaning' (e.g. Dickins, Hervey and Higgins 2017: 103-104), and, associated with this, metaphor (cf. Dickins, Hervey and Higgins 2017: 194-210, and, rather more rigorously, Dickins 2005; Dickins 2018).

This definition excludes compounds (Section 5), phrasal verbs (Section 6) and also proverbs, which are clausal (Section 8). This is an attempt to mirror what is generally meant by idioms in everyday language (idioms being classified as non-technical in Section 1). Although most idioms in English are figurative (mainly metaphorical, but also sometimes metonymic, etc.), figurativeness is not a defining feature of idioms. Thus '(as) sure as eggs is eggs' meaning 'absolutely sure' is an idiom, but is not figurative, as is 'by and large' (assuming we classify this as an idiom, rather than a multiword compound; this section, below).

In terms of the semantic independence of their constituents, idioms can be analysed as illustrated in Figure 16.

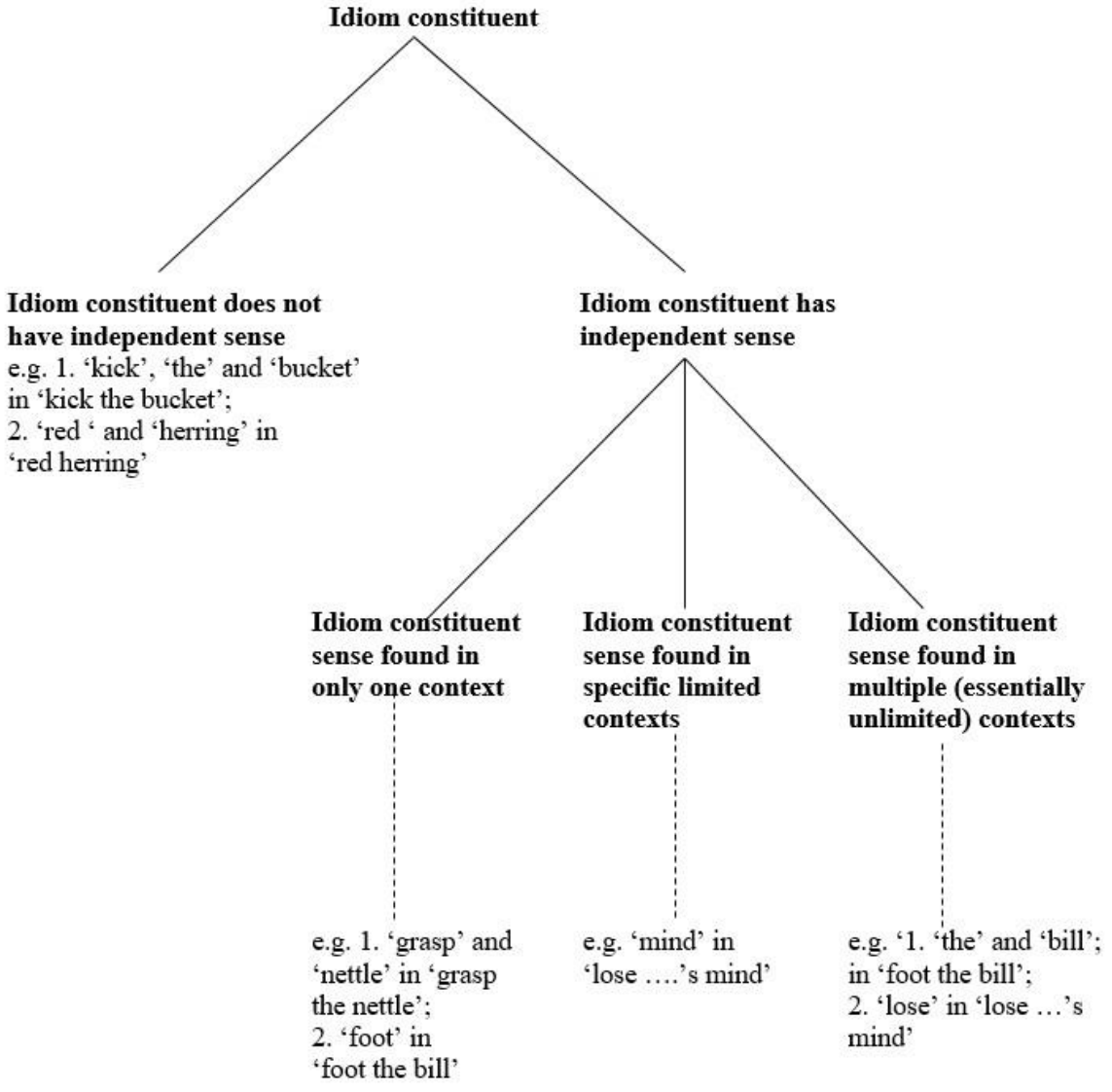

Figure 16: Analysis of idioms according to semantic independence of constituents

As with other multiword expressions, constituents in idioms may not have an independent sense, i.e. the constituent cannot be analysed semantically independently from other constituents in the idiom. Alternatively, the constituent may have an independent sense, i.e. it can be analysed semantically independently from other constituents in the idiom. Where a constituent has an independent sense, this may occur either (i) only in the context of this idiom,

(ii) in a limited number of other contexts as well, or (iii) in multiple contexts (essentially unlimited in number).

Idioms involving constituents which do not have independent senses must have logically at least two such constituents – these two constituents (neither with an independent sense) together forming a larger constituent which does have independent sense. Thus the entire idiom 'kick the bucket' has an independent sense (= die), but none of the constituents which make it up (taking each word in the idiom to be a separate constituent), 'kick', 'the' or 'bucket', has an independent sense. Where idioms are made up entirely of constituents which do not have independent senses, these idioms cannot be changed in any way (although 'additions' may be made, e.g. for tense, as in 'he kicked the bucket'). So, for example, one cannot say 'the bucket was kicked', or 'They both kicked buckets' (cf. Section 4).

Another example of idiom constituents which do not have an independent sense are 'red' and 'herring', in 'red herring', i.e. 'anything that diverts attention from a topic or line of inquiry' (CEDO). 'Red herring' is interesting because although it is typically classified as an idiom, it could arguably also be classified as a compound (for the analysis of compounds, see sections 5–5.3).

A second type of idiom constituent has an independent sense, but this sense is only found in the context of this idiom. An example is 'grasp' in 'grasp the nettle'. Here, 'grasp' has the sense 'tackle', 'deal with' or similar. This is not, however, a sense which 'grasp' has in any other context. Similarly, 'nettle' in 'grasp the nettle', has the sense 'difficult problem' or similar, a sense which it does not have in any other context. The fact that these two constituents (words) have independent senses in the idiom 'grasp the nettle' is shown by the fact that the constituents in the idiom (unlike those in 'kick the bucket') can be reorganised grammatically and be further modified. For example, it is possible to say things like 'That's one nettle which you are just going to have to grasp', or 'Eventually the British government grasped the nettle of Irish peace' (cf. also Dickins 1998: 241–243, 324, 435).

There are also idioms in which only one element has an independent sense, this sense being only found in the context of this idiom. An example is 'foot' in 'foot the bill' (= 'pay or settle (a bill, esp. one which is large or unreasonable, or which has been run up by another party'). Here, 'bill' means what it does in unlimited other contexts. 'Foot', however, is only found with this sense in this one phrase[14].

A third type of idiom constituent is one which has an independent sense this being found in more than one context (distinguishing it from the second type of idiom constituent above), but only in a limited number of contexts. An example is 'mind', meaning roughly 'rational faculties', in 'lose …'s mind'. This seems to occur in only one other context: 'out of …'s mind'. That 'mind' has an independent sense here is shown by the demonstration that 'lose' has an independent sense – in fact a sense which occurs in unlimited contexts, e.g. 'lose ….'s sanity/rational faculties/self-control/temper'. It is noteworthy, also, that while it is possible

[14] It could be argued that there is marginally one other context in which 'foot' occurs in this sense. This is with 'it', for example in 'I can't pay this bill. You're going to have to foot it'. As this example shows, however, not only does 'bill' have to be somewhere in the general discourse-context; the word 'it' also has to refer specifically to 'bill'. It is, of course, also possible to say things like, 'I've been footing all the bills for months now', with 'bills' in the plural. Here, however, it seems sensible to say that 'foot' still occurs in the context of 'bill', albeit that 'bill' itself is in the plural form (i.e. it has the plural suffix 's').

to say 'regain ….'s sanity/rational faculties/self-control/temper', it is not possible to say *'regain …'s mind'.

A final type of idiom constituent is one which has an independent sense and this is found in multiple (essentially unlimited) contexts. Examples are 'the' and 'bill' in 'foot the bill', and 'lose' in 'lose …'s mind' (discussed immediately above). Nunberg, Sag and Wasow (1994) refer to idioms in which none of the constituents has an independent sense as 'idiomatic phrases', while Sag et al. (2002) refer to them as 'non-decomposable idioms'. Nunberg, Sag and Wasow (1994) refer to idioms in which each of the elements has an independent sense as 'idiomatically combining expressions' or 'idiomatic combinations', while Sag et al. (2002) refer to them as 'decomposable idioms'.[15]

Figure 17 shows the semantic relationships between collocations, formulaic sequences (formulaic language), multiword expressions, compounds, phrasal verbs and idioms.

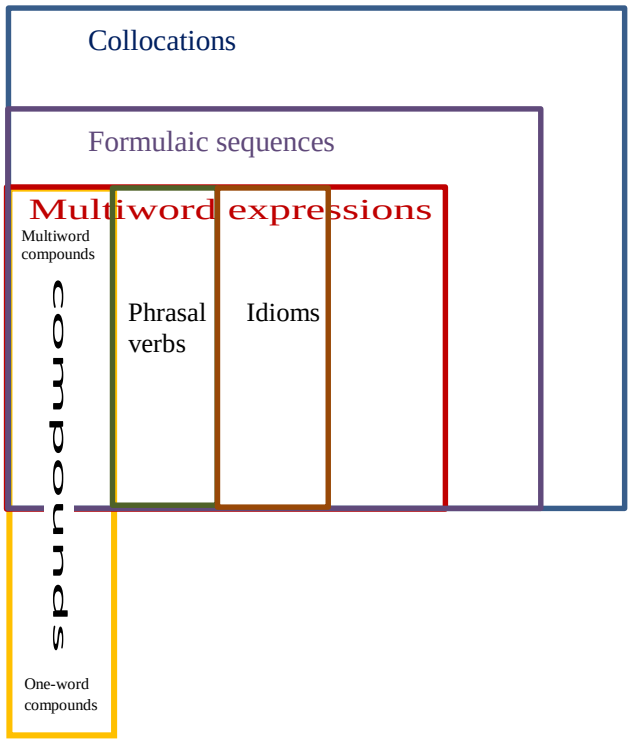

Figure 17: Semantic relationship between collocations, formulaic sequences (formulaic language), multiword expressions, compounds, phrasal verbs and idioms

In this section, I have defined an idiom as 'a phrase that is not a compound, not a phrasal verb, is non-clausal, and is not fully free-compositional'. I have also suggested that 'idiom' is a Group 1 term (Section 1), i.e. a non-technical everyday term. Given that 'compound' (sections 5–5.3) and 'phrasal verb' are semi-technical terms (Group 2 terms; Section 1),

[15] As with phrasal verbs, the account of the semantics of idioms which I have given here is not complete (see also Section 1). In addition to the denotative semantic independence of the components of idioms, there are also connotative issues of 'reflected meaning' (e.g. Dickins, Hervey and Higgins 2017: 103–104), and, associated with this, metaphor (cf. Dickins, Hervey and Higgins 2017: 194–210, and, rather more rigorously, Dickins 2005; Dickins 2018).

which we have some liberty in defining for our purposes, there does not seem a problem in defining these two notions in such a way that they are fully distinct (disjunct, in set-theoretical terms).

The greater problem is between i. compounds and idioms; and ii. phrasal verbs and idioms. Given that 'idiom' is a Group 1 term, a non-technical everyday term (Section 1), we are not in practice at much liberty to redefine it for academic purposes. The kind of issues this raises can be seen in relation to 'sleeping policeman', analysed in Section 5.1 as a multiword compound, and 'red herring', analysed in this section as an idiom. Both might be thought of as figurative (though this is not, in the current approach, a defining feature of either compounds or idioms): 'sleeping policeman' is fairly clearly metaphorical, while 'red herring' is, from a figurative perspective, puzzling, because there is no clear reason why a 'red herring' ('something that diverts attention from a topic or line of inquiry') should be figuratively related to a red-coloured herring (fish). It is also not immediately clear why 'sleeping policeman' should be analysed as a compound and 'red herring' as an idiom.

A partial grammatical solution could be found by further defining compounds as being members of specific word-classes: nominal, adjectival, adverbial, verbal, etc. This would make it possible to unambiguously categorise 'grasp the nettle' as an idiom, rather than a compound, because it involves more than one word class, having one constituent 'grasp', which is verbal, and another '(the) nettle', which is nominal. This does not help in the case of 'sleeping policeman' and 'red herring', however; both are nominal.

One way round this would be to rely on native speaker judgements, assuming fairly consistent judgements are made by native speakers in regard to idioms and non-idioms. Thus, we could define a compound as any multiword expression which is not regarded by native speakers as an idiom or a phrasal verb. Thus, if native speakers typically regard both 'sleeping policeman' and 'red herring' as idioms, we would class these both as idioms. If they regard only 'red herring' as an idiom, we would then regard (classify) 'sleeping policeman' as a compound (to be further discussed below).

We could adopt the same procedure with regard to phrasal verbs and idioms, i.e. taking native speakers' views on the nature of what is and is not an idiom into account first, and then classifying as a phrasal verb any relevant example which is not generally considered by them to be an idiom. Under this approach, if native speakers were to regard 'turn on' as an idiom, but 'give up' as not an idiom, the former would be classified as an idiom and the latter as a phrasal verb. This approach, however, seems rather messy, since 'phrasal verb' is fairly well understood as a semi-technical term, and what we have identified as phrasal verbs are perhaps only marginally regarded in everyday understanding of idioms as also being idioms. It may be better, then, on this basis to treat 'phrasal verb' and 'idiom' as discrete sets.

The distinction between compounds and idioms (e.g. putatively 'sleeping policeman' and 'red herring'), by contrast, seems much more problematic. The relative technicality of 'sleeping policeman' might make us somewhat more inclined to regard it as compound rather than an idiom. However, we can see that there are likely to be a very large number of similar examples in which a division such as technical vs. non-technical (etc.) is very blurred. An alternative to the definition of 'idiom' given at the start of this section and visualised as in Figure 17, therefore, might be to define idiom and compound as overlapping, i.e. to accept that a phrase can be both a compound and an idiom. We could, of course, also do the same

with ‘phrasal noun’ and ‘idiom’, making them overlapping classes, such that a phrase can be both a phrasal noun an idiom. These possible redefinitions are represented in Figure 18.

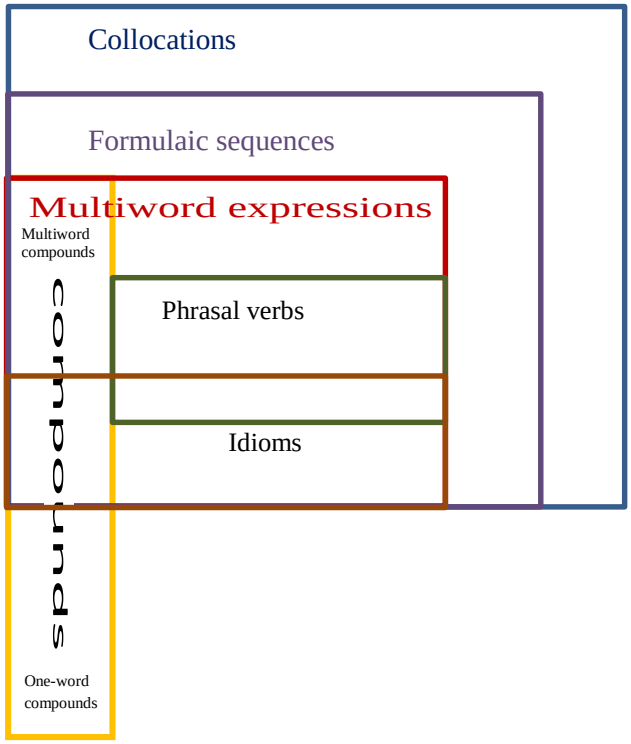

Figure 18: Semantic relationship between collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs and idioms – with compounds and idioms, and phrasal verbs and idioms defined as overlapping classes

In the remainder of this article, I will stick with the definitions of compound, phrasal verb and idiom given at the start of sections 5, 6 and 7.

8. Proverbs

Unlike idioms, proverbs are clausal, i.e. they can stand on their own as sentence[16] (though they may also occur as a clause within a larger sentence, e.g. ‘Too many cooks spoil the broth, as they say’). OEDO defines a proverb as ‘A short, traditional, and pithy saying; a concise sentence, typically metaphorical or alliterative in form, stating a general truth or piece of advice; an adage or maxim’. Issa (2014, Chapter 2) provides a useful survey of different definitions of proverbs, as well as views on their typical (though not necessarily defining) features.[17]

[16] Not all proverbs have the standard form of sentences in English, and in particular they do not all contain a main verb, for example ‘any port in a storm’. All proverbs, however, can stand on their own as sentences, i.e. they can appear as complete utterances.

[17] Reviewer 1 has pointed out to me that there is an interesting difference between the definition of ‘proverb’ given here and the definitions of other categories discussed in this article (collocations, formulaic sequences,

Since 'proverb' is a non-technical term (Section 1), I will adopt the view here that proverbs are what native speakers consider to be proverbs (apart from the stipulation, above, that proverbs are clauses – which would, I believe, be reflected in terms of their identification of proverbs in practice by native speakers). This, of course, means that the boundaries between what is and is not a proverb will be somewhat fuzzy (Section 12): we cannot expect all native speakers to recognise exactly the same things as proverbs (and non-proverbs).

Some proverbs are fully free-compositional – i.e. all the words which make them up are used in the same sense in which they are used in other contexts. Examples are 'Honesty is the best policy', 'A little learning is a dangerous thing' and 'Better late than never'. Most proverbs, however, contain at least some words which are not completely free-compositional, i.e. they include words which are not found in the same sense in unlimited contexts. Examples are 'Too many cooks spoil the broth', 'A stitch in time saves nine', and 'Birds of a feather flock together' (all of which are fully bound-compositional; i.e. none of the constituent words in them has an independent sense). Fully free-compositional proverbs do not belong to the multiword expression category, though they do belong to the category of formulaic sequences (given the inevitably high levels of collocation they involve). Proverbs which contain at least some words which are not completely free-compositional belong to the multiword expression category.

In terms of the semantic independence of their constituents, proverbs can be analysed as illustrated in Figure 19 (next page).

'One swallow doesn't make a summer' is an example of a proverb in which none of the individual constituent words has an independent sense (the proverb's meaning being along the lines 'It should not be assumed that something is true just because there is one piece of evidence for it').[18] 'Silence is golden', meaning 'Silence is virtuous/preferable (to speaking)/to be enjoyed, etc.', provides an example of a second type of proverb constituent. Here 'golden' has an independent sense, but this sense is only found in the context of this proverb.[19] I have not been able to find an example of a third type of proverb constituent, where this constituent has a sense which only occurs in specific limited contexts.

The final type of proverb constituent, which has an independent sense and this sense is found in multiple (essentially unlimited) contexts, is illustrated by the proverb 'honesty is the best policy'. Here all the constituent words have the same sense as they have in unlimited

multiword expressions, compounds, phrasal verbs and idioms). In the case if proverbs alone, situational/discoursal considerations are invoked, e.g. in the OEDO definition 'A short, traditional, and pithy saying [...] stating a general truth or piece of advice […]'. There is thus inconsistency in the criteria used for identifying proverbs as opposed to these other categories. However, given that 'proverb' is a non-technical term (Section 1) – and that proverbs can, I think, can only be reasonably identified on a non-technical basis – the characterisation of proverbs in situational/discoursal terms is justified.

[18] I believe that the analysis that none of the constituents in 'one swallow doesn't make a summer' has an independent sense is correct. However, as Reviewer 2 has pointed out to me, this might be queried. He/she notes, "One wonders whether this is really true for *one* (and perhaps also for the negation). After all, the author invokes the idea of oneness in the characterization of the proverb's meaning ("It should not be assumed that something is true just because there is **one** piece of evidence for it")."

[19] Other similar senses of 'golden', such as 'Of time, an opportunity: Of inestimable value; exceedingly favourable or propitious' (OEDO), as in 'a golden opportunity' are, I believe, distinct from the sense of 'golden' in 'silence is golden', having a clearly different semantic range.

other contexts, i.e. the proverb is fully free-compositional. Another example is provided by 'silence' and 'is' in 'silence is golden'.[20]

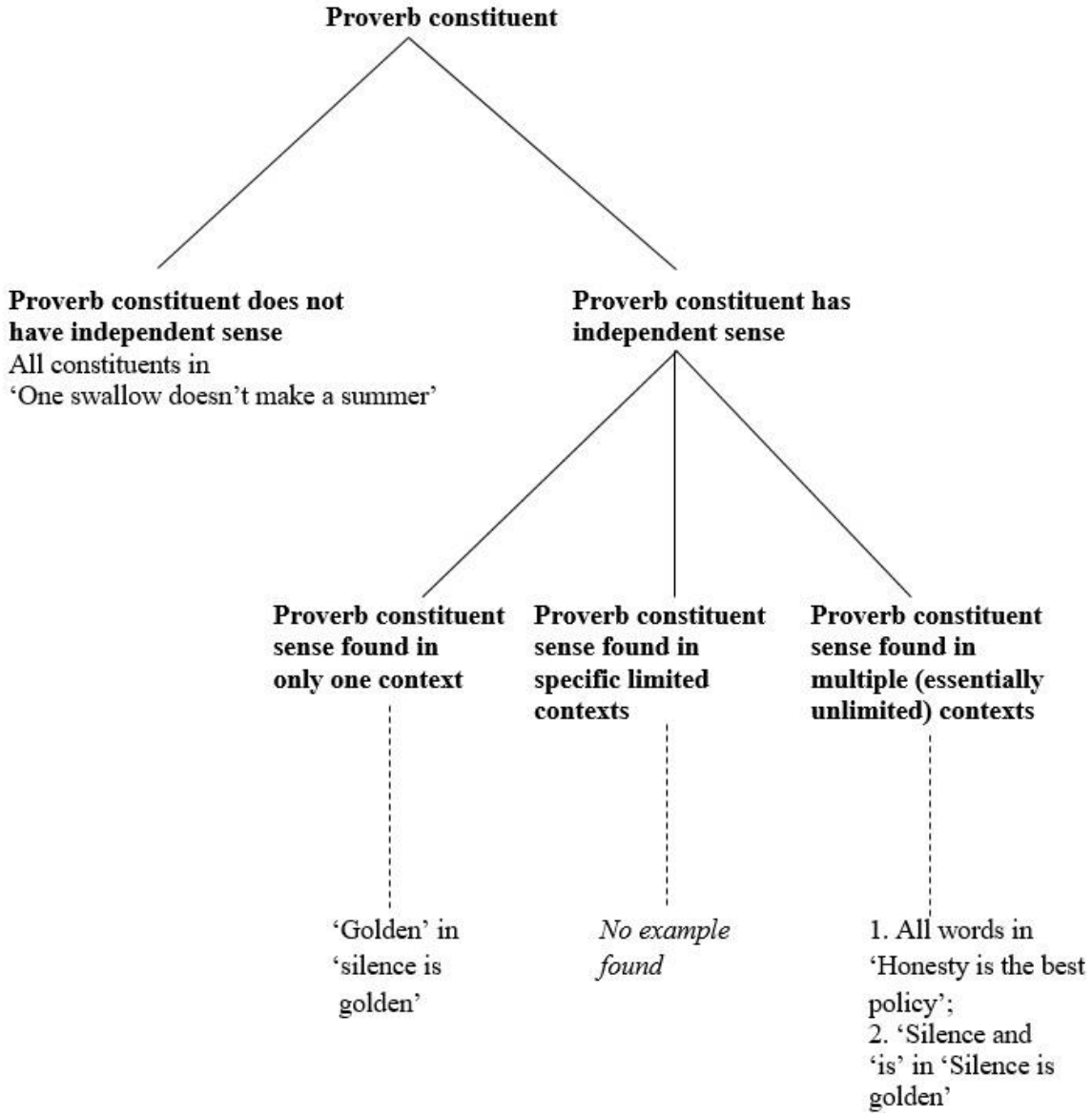

Figure 19: Analysis of proverbs according to semantic independence of constituents

Figure 20 (next page) shows the semantic relationship between collocations, formulaic sequences (formulaic language), multiword expressions, compounds, phrasal verbs, idioms and proverbs.

[20] As with phrasal verbs and idioms, the account of the semantics of proverbs which I have given here is not complete (see also Section 1). In addition to the denotative semantic independence of the components of proverbs, there are also connotative issues of 'reflected meaning' (e.g. Dickins, Hervey and Higgins 2017: 103–104), and, associated with this, metaphor (cf. Dickins, Hervey and Higgins 2017: 194–210, and, rather more rigorously, Dickins 2005; Dickins 2018).

9. Other types of multiword expression?

In this article, I have assumed that there are no other types of multiword expression (as defined in this article: Section 4) in English in addition to compounds, phrasal verbs, idioms and some proverbs[21]. This would need to be tested in future research. There are, of course, other terms for phenomena which are deemed to be similar to proverbs, such as 'aphorism', 'maxim', 'saying' and 'adage'. These typically belong under the category of 'formulaic sequence' and are not multiword expressions.

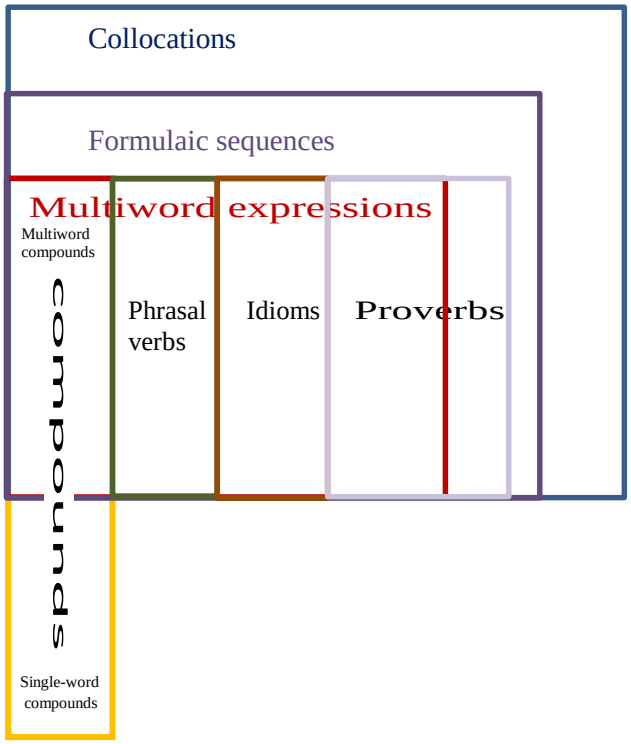

Figure 20: Semantic relationship between collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs, idioms and proverbs

10. Further categories deriving from 'collocation', 'formulaic sequence' and 'multiword expression'

I have established the basic categories of collocation, formulaic sequence and multiword expression. I have also established the more specific categories of phrasal verb and idiom overlapping with multiword expression, and of proverb overlapping with multiword expression and non-MWE formulaic sequence (see below).

[21] I have excluded proper names from consideration in this article; these are, for example, included as a type of multiword expression by Sag, Baldwin, Bond, Copestake and Flickinger (2002) in their discussion of multiword expressions. They have, however, specific features which I believe mean that they need to be analysed separately from the phenomena considered in this article.

On the basis of the basic categories of collocation, formulaic sequence and multiword expression we identify further analytically useful categories, as follows:

1. **Non-formulaic collocation**; i.e. a collocation which is not a formulaic sequence (formulaic sequences of course also include multiword expressions, which themselves properly include multiword compounds, phrasal verbs, idioms, and proverbs) (see Section 10).
2. **Non-MWE formulaic sequence** (where 'MWE', as noted in Section 4, stands for 'multiword expression'); i.e. a formulaic sequence which is not a multiword expression (multiword expressions also include multiword compounds, phrasal verbs, idioms, and fully free-compositional proverbs).
3. **Non-MWE collocation**; i.e. a collocation which is not a multiword expression (multiword expressions also include multiword compounds, phrasal verbs, idioms, and fully free-compositional proverbs).

Figure 21 shows the scope of category 1 *non-formulaic collocations*: this is the area with the green background.

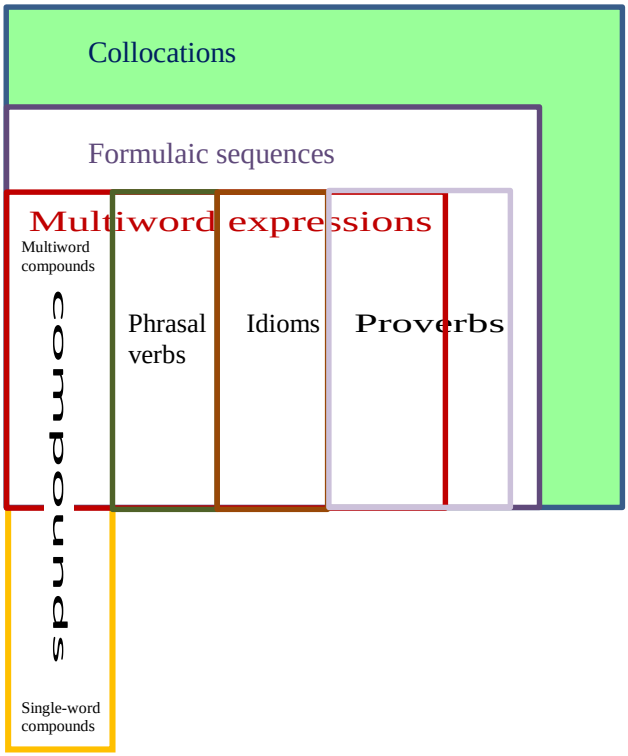

Figure 21: Scope of non-formulaic collocations

Category 1 (Figure 21), *non-formulaic collocations*, comprises fully free-compositional collocations which, however, lack the syntactic coherence (and possibly also the statistical frequency; see Section 3) to count as formulaic sequences. Because of their lack of syntactic coherence (and possibly the fact that the statistical occurrence of the collocational elements is not very significantly higher than their statistical occurrence across all contexts), these are the kind of collocations which may not be recognised by native speakers as such, and are likely to be only revealed by statistical computational analysis.

Figure 22 shows the scope of non-MWE formulaic sequences – the area with the green background.

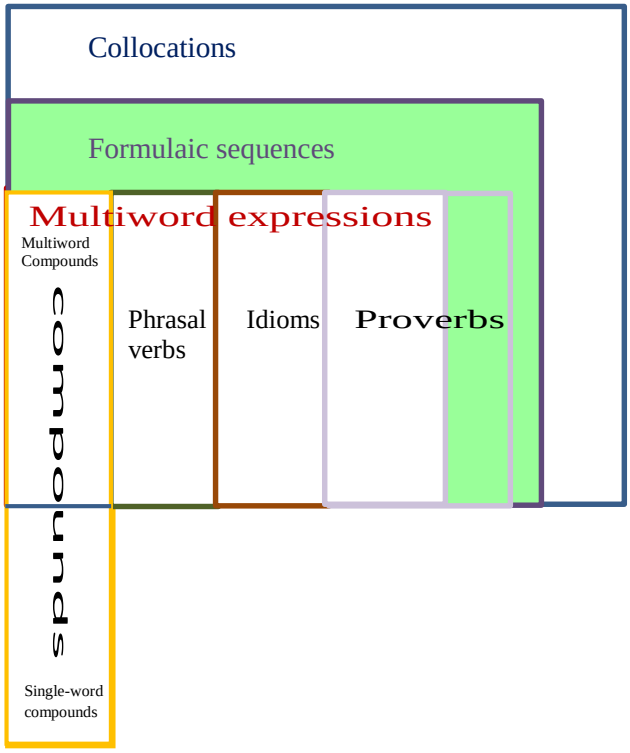

Figure 22: Scope of non-MWE formulaic sequences

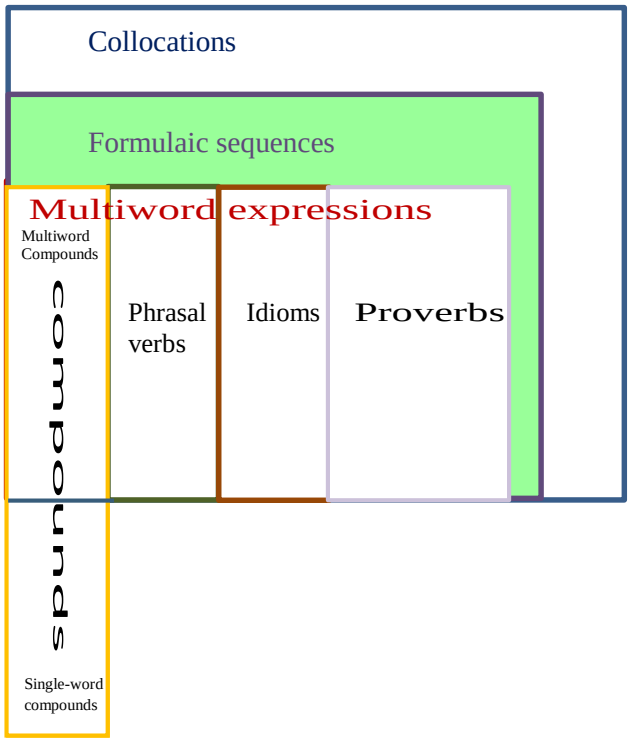

Figure 23: Scope of non-MWE, non-proverbial formulaic sequences

Category 2 (Figure 22), *non-MWE formulaic sequences*, comprises fully free-compositional collocations whose syntactic coherence (and possibly also frequency of occurrence) means they are classified as formulaic. One potential practical issue with category 2, *non-MWE formulaic sequences*, is that it cuts through proverbs: those proverbs which are fully free-compositional are included in it, while those that are not (i.e. which are multiword expressions) are excluded from it. As noted (Section 1), 'proverb' is an everyday term, for which native speakers are likely to agree on the ostensible definition (i.e. they are likely to agree about are and are not cases of proverbs), even if they cannot provide an abstract definition of what a proverb is in principle. 'Proverb' is, accordingly, for native speakers a fairly natural class (for which see the discussion in Hervey 1982: 17–18, of Peirce 1960) and not one which one would want to split up by another taxonomy. Accordingly, instead of the category *non-MWE formulaic sequences*, it would probably be preferable to operate with an otherwise identical category which excludes all proverbs, i.e. *non-MWE, non-proverbial formulaic sequences*. This is represented by the green-shaded area in Figure 23 (previous page).

Figure 24 shows the scope of category 3 above, *non-MWE collocations*. This category is useful for distinguishing between those collocations and formulaic sequences which are fully free-compositional and those which are not.

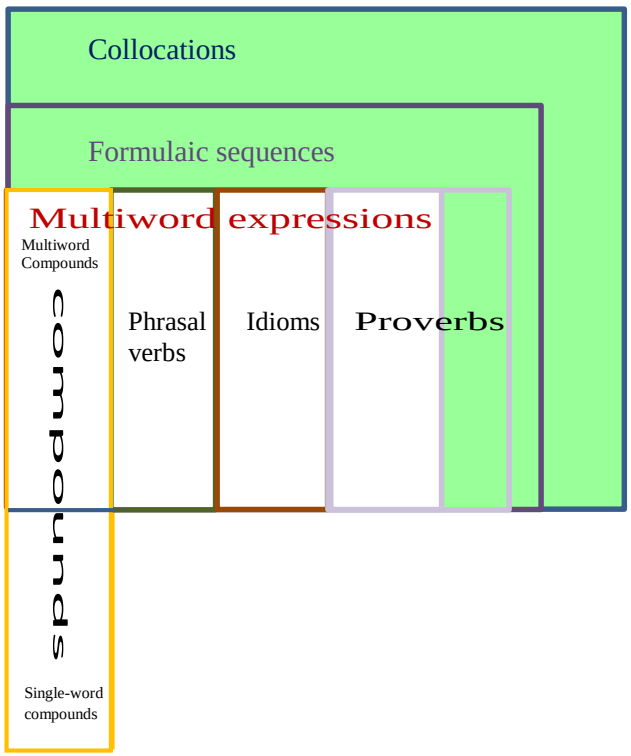

Figure 24: Scope of Non-MWE collocations

As with category 2, *non-MWE formulaic sequences*, category 3, *non-MWE collocations* seems unnatural because it cuts through proverbs, including some and excluding others. As with category 2, therefore, it seems better to operate with an otherwise identical category here which excludes all proverbs, i.e. *non-MWE, non-proverbial collocations*. This is represented by the green-shaded area in Figure 25 (next page).

In the approach known as ‘phraseology’, the term ‘collocation’ is sometimes defined to mean roughly what is meant in this article by ‘non-MWE collocations’, such that collocations “differ from other types of phraseological units that exhibit a fixed form and a non-decomposable, unitary meaning” (Pastor 2017: 29). For an introduction to phraseology, see Gries (2008: 3–26). Cruse (1986: 40) also defines ‘collocation’ in this way.

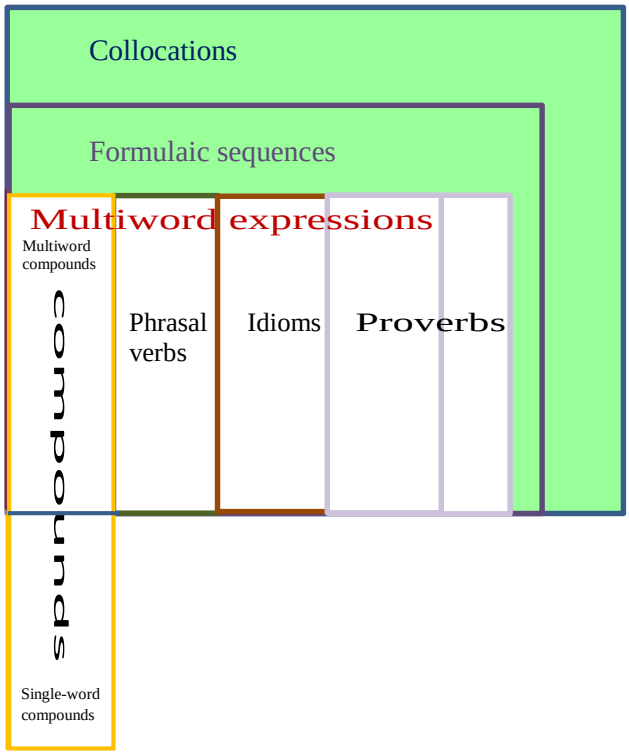

Figure 25: Scope of Non-MWE, non-proverbial collocations

The category of non-MWE, non-proverbial collocations allows for the investigation of all those fully free-compositional collocations, whether weak or strong (excluding proverbs), ignoring the question (which may be a moot one) of whether the collocations in question constitute formulaic sequences. Other derived categories in addition to those identified in this section could, of course, be established, depending on specific research concerns.

11. Semantic correlates of syntactic relationships in multiword expressions

Up till now I have focused on constituents, and largely ignored the semantic correlates of syntactic relationships. It is clear that these semantic correlates, even in fully free-compositional expressions are varied and can be complex (e.g. Ó Séaghdha 2008). For instance, a ‘door key’ is normally a key for opening a door, while an ‘ignition key’ is normally a key for turning on the ignition. It is, however, perfectly possible to say ‘door and ignition keys’, demonstrating that ‘key’ has the same sense in both ‘door key’ and ‘ignition key’, a sense which is also found essentially unlimited other contexts. It is also clear that ‘door’ and ‘ignition’ in ‘door key’ and ‘ignition key’ have the same sense as they do in unlimited other contexts, as can be seen from the possibility of phrases such as ‘keys for the door, the ignition,

the garage, the house, the garden shed [etc.]'. English (other languages also) allows for semantic narrowing in given phrases of the full possible denotative range which can be expressed by, for example, noun-noun expressions. It falls outside the scope of this article to investigate this in further detail, but it is an issue which needs to be addressed for a full understanding of the compositionality or non-compositionality of collocations. It may be that a distinction needs to be made between forms like 'door key' and 'traffic lights'. 'Door key' arguably involves an extra-linguistically determined (or, perhaps better, 'motivated') restriction on typical range of reference – what Dickins, Hervey and Higgins refer to as associative connotative meaning (Dickins, Hervey and Higgins 2017: 97–99; for a more theoretically grounded account, see Dickins 2014). 'Traffic lights', by contrast, seems to have a sense which is much more specifically denotatively fixed (cf. Sag, Baldwin, Bond, Copestake, and Flickinger 2002). In 'traffic lights', just as much as 'door key', however, the two constituent words both seem to have an independent sense. This is borne out by the fact that it is possible to say things like 'traffic lights and signals' (15 results on IWeb, 1.10.18: <https://corpus.byu.edu/iweb/>), where at least some of the examples are to be read as equivalent to 'traffic lights and traffic signals'; e.g. "Ensuring that traffic flows as smoothly as possible at all times, including rush-hours, calls for planning, studying traffic volumes at various times of day, and knowing how to install and coordinate traffic lights and signals, signs and other traffic control devices": <https://www.cityoftulsa.org/government/departments/streets-and-stormwater/streets/>. It is also borne out by the fact that one can also say things like 'street and traffic lights' (21 results on IWeb, 1.10.18: <https://corpus.byu.edu/iweb/>) or 'traffic and street lights' (4 results on IWeb, 1.10.18: <https://corpus.byu.edu/iweb/>), where at least some of the examples are to be read as equivalent to 'traffic lights and street lights'; e.g. "The road system and infrastructure would likely need major upgrades for driverless vehicles to operate on them. Traffic and street lights, for instance, would likely all need altering": <https://axleaddict.com/safety/Advantages-and-Disadvantages-of-Driverless-Cars>.

It is possible that what differentiates 'traffic lights' from 'door key' semantically is the greater specificity of the denotative relationship associated with the noun-noun syntactic relationship in the former as compared to the latter. It is, finally, worth stressing, regardless of issues of the semantic correlates of syntactic relationships between elements in phrases of this kind, that these correlates fall into definable patterns. This is evidenced by the oddity of a neologism such as 'tamper-evident' as in 'tamper-evident packaging' (i.e. packaging where it is/will be evident if it has been tampered with). Here, the words 'tamper' and 'evident' have their standard senses, but the semantic correlates of the syntactic relationship involved falls outside the standard range permitted by the English language.

12. Fuzzy boundaries and discrete boundaries

In understanding categories, it is essential to distinguish between those which involve discrete boundaries and those which involve fuzzy boundaries. Fuzzy boundaries involve situations in which it is not entirely clear which of two classes a particular entity is to be assigned to. This situation is to be distinguished from semantic overlap, where a particular entity can be assigned to two classes simultaneously. Semantic overlap is illustrated by 'doctor' and

'genius'. Some, but not all, doctors are geniuses, and some, but not all, geniuses are doctors. We can accordingly say 'he is both a doctor and a genius' (or 'he is both a genius and a doctor'). This situation contrasts with, for example, 'cup' and 'mug'. Labov (1972) showed that English speakers distinguish cups from mugs according to a variety of features including the shape of the vessel concerned. There are also numerous objects which a speaker might describe as a cup or a mug (and one might even say, for example, 'You could call that a cup or a mug'). However, 'cup' and 'mug' are not a case of semantic overlap: one cannot say 'that is both a cup and a mug' (or 'that is both a mug and a cup'). 'Cup' and 'mug' are, in abstract semantic terms, discrete (in set-theory 'disjunct') notions. In the real world, however, the boundary between 'cup' and 'mug' is fuzzy. (Cf. Dickins 2014: 20, for further discussion of abstract semantic disjunction vs. real-world (realisational) 'semantic overlap', i.e. fuzzy boundaries.)

We have already seen that in perhaps the most basic distinctions made in this article – those between everyday terms, semi-technical terms and technical terms (Section 1) – we are dealing with fuzzy boundaries. It will not in all cases be clear whether a particular term should be regarded as everyday or semi-technical, or as semi-technical or technical. We can consider in the light of the same distinction between fuzzy boundaries and discrete boundaries the fundamental concepts defined in this article: collocations, formulaic sequences, multiword expressions, compounds, phrasal verbs, idioms and proverbs.

At the outer boundary of these concepts is the distinction between collocation and non-collocation. If we use statistical frequency as our sole criterion, this is, properly speaking, a discrete boundary. Technically, as soon as we have a frequency of co-occurrence between two or more words which is higher than predicted by their frequency of occurrence as individual words, we have a collocation. In practice, we are unlikely to choose to investigate collocations where this frequency of co-occurrence is not significantly greater than the frequency of occurrence of the individual words; but this is a matter of research focus, not of the definition of what constitutes a collocation.

The distinction between non-formulaic collocation and formulaic sequence, by contrast, is potentially complex. If we use syntactic coherence as our sole criterion, we have a discrete boundary, as we do if we also add to this statistical frequency. The statistical frequency boundary is, however, conventional, because we could have chosen to draw it somewhere else. Finally, as soon as we introduce criteria involving native-speaker judgements, the boundary becomes fuzzy (with different native speakers no doubt making different judgements in different cases).

Since non-MWE collocations involve only fully free-compositional constituents and multiword expressions (MWEs) have at least one constituent which is not fully free-compositional, the boundary between the two is non-fuzzy – at least assuming that it is unambiguously possible to determine what is a fully free-compositional constituent and what is not.

The relationship between multiword expression and multiword compound, multiword expression and phrasal verb, and multiword expression and idiom are clear: all of multiword expression, multiword compound and idiom are sub-types (properly included in) multiword expression. There are no complications between multiword expression and these properly included concepts in terms of fuzzy boundaries.

The boundaries between compounds, phrasal verbs and idioms are more problematic. It seems possible to define compounds and phrasal verbs in a way which both makes the two

notions disjunct (in set-theoretical terms) and the boundary between them non-fuzzy, and does not noticeably conflict with the general understanding of these terms (cf. Section 7). The boundary between compounds and idioms can be made disjunct, though that boundary is likely to be fuzzy: even if we establish two disjunct classes, thereby stipulating that an example can only be a member of one class, there are going to be cases where we are not sure which class we wish to assign that example to. If, of course, we were to define 'compound' and 'idiom' as overlapping classes (Section 7), some examples would, by definition, potentially be members of both classes.

Similarly even if the boundary between phrasal verbs and idioms is made disjunct, this boundary is likely to be fuzzy, with examples in which it is not entirely clear what is the more appropriate category. If we were to define 'phrasal verb' and 'idiom' as overlapping classes (Section 7), some examples would, by definition, potentially be members of both classes. Finally, the boundary between compounds/phrasal verbs/idioms on the one hand, and proverbs on the other is non-fuzzy: only proverbs are clausal, and there should, therefore, be no indeterminate cases. 'Proverb', of course, overlaps with multiword expression and non-MWE formulaic sequence, meaning that a particular proverb may be also a multiword expression or a non-MWE formulaic sequence.

13. Universal categories and language-specific categories

The following categories are universal in that we would expect to find them in any language: collocation, formulaic sequence, multiword expression. For a natural language not to have collocations would involve the vanishingly unlikely situation that there was no variation in the statistical frequency of words regardless of what other words they occurred in the context of. Corresponding arguments apply to formulaic sequences. 'Multiword expression', as defined in this article, is a notion derived from the fundamental possibilities of the relationships between syntax and semantics. It is possible for 'languages' not to have multiword expressions (logical languages do not have them, for example). All natural languages, however, seem in practice to have multiword expressions.

The following categories are better regarded as non-universal: compound, phrasal verb, idiom and proverb. 'Compound' is a semi-technical term (a Group 2 term; Section 1), which is applied to specific lexical features in English. However, this may not be an appropriate term to use with other languages, for instance because their lexical features are very different from those of English. Alternatively, if we use the same term, we need to be aware that what we are referring to are not necessarily the same types of features as in English. Either way, 'compound' is not a universal category, meaning that we cannot simply transpose the ontology given in this article from English to other languages. The same applies to 'phrasal verb', which is not a category found in many languages; and even in languages where a category exists which can reasonably be given this name, it may well be very different in key respects from 'phrasal verb' in English. 'Idiom', as an everyday non-technical (Group 1; Section 1) term, is, similarly, very unlikely to have simple correspondents in many other languages, and can therefore be said to not exist as a 'ready-made' concept in these languages.

Although sayings which we might identify as 'proverbs' may be found in all natural languages (cf. Issa 2014: Section 2.2), we should not take it that the category of 'proverb' as

defined in (and applied to) English is itself universal. Arabic, for example, makes a distinction between a *maṯal*, typically translated as 'proverb' and a *ḥikma* (literally 'wisdom'), i.e. a wise saying. Not all examples of what Arabs regard as *maṯal* seem to be what English speakers would think of a proverbs, while *ḥikma* is a category which does not really exist for English. The notion 'proverb' cannot therefore be simply transposed from this ontology for English to one for Arabic, or by extension to other languages.

14. Conclusion

The article has attempted to establish an ontology (with discussion of some alternatives) for collocations, formulaic sequences, multiword expressions, compounds, idioms and phrasal verbs in English. I have stressed the 'constructive' nature of this: particularly in the case of technical terms, it is not a question of saying 'what do these terms mean?', but rather: 'what can we use these terms to mean, such that we arrive at an overall ontology which is both coherent and useful for our purposes?'. I have also stressed that different kinds of word/phrases for notions in the ontology should, for practical reasons, be treated rather differently: we have more leeway to redefine technical terms for our own purposes (particularly when these terms already have multiple and contradictory definitions in the literature) than we do to redefine everyday terms which already have a fairly generally agreed (if probably rather vague) meaning. The resulting ontology is not intended to be definitive. We might expect it to be improved on through further consideration of the same areas of phenomena. It is also only of value for investigating the areas of phenomena which it seeks to define. For other, even closely related, areas of analysis, other ontologies involving some other terms and notions would need to be established.

References

BAHNS, Jens and Moira ELDAWa. 1993. 'Should we teach EFL students collocations?'. *System*, 21(1). Pp. 101–114.

BALDWIN, Timothy and Su Nam KIM. 2010. 'Multiword expressions'. In Nitin Indurkhya and Fred Damerau (eds.), *Handbook of natural language processing* (2nd ed.). Boca Raton: CRC Press. Pp. 267–292.

BARTSCH, Sabine. 2004. *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Narr.

BAUER, Laurie. 2004. *A glossary of morphology*. Edinburgh: Edinburgh University Press.

CAI, Ying. 2017. *Second language acquisition of Chinese verb-noun collocations*. University of Massachusetts Amherst: Masters thesis.

CALZOLARI, Nicoletta, Charles FILLMORE, Ralph GRISHMAN, Nancy IDE, Alessandro LENCI, Catherine MACLEOD, and Antonio ZAMPOLLI. 2002. 'Towards best practice for multiword expressions in computational lexicons'. In *Proceedings of LREC 2002*. Canary Islands. Pp. 1934–1940.

CARPUAT, Marine and Mona DIAB. 2010. 'Task-based evaluation of multiword expressions: a pilot study in statistical machine translation'. In *Proceedings of NAACL/HLT 2010*. Los Angeles. Pp. 242–245.

CARSTAIRS-MCCARTHY, Andrew. 2002. *An introduction to English morphology*. Edinburgh: Edinburgh University Press.

CONSTANT, Mathieu, Gülşen ERYIĞIT, Johanna MONTI, Lonneke van der PLAS, Carlos RAMISCH, Michael ROSNER, and Amalia TODIRASCU. 2017. 'Multiword expression processing: a survey'. *Computational Linguistics*, 43(4). Pp. 837–892.

COWIE, Anthony. 1978. 'The making of dictionaries and reference works'. In Peter Strevens (ed.), *In honour of A.S. Hornby*. Oxford: Oxford University Press. Pp. 127–139.

CRUSE, David Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

CRYSTAL, David. 2008. *A dictionary of linguistics and phonetics* (6th edn.). Oxford: Blackwell.

DICKINS, James. 1998. *Extended axiomatic linguistics.* Berlin and New York: Mouton de Gruyter.

——. 2005. 'Two models for metaphor translation'. *Target* 17(2). Pp. 227–273.

——. 2006. 'The verb base in Central Urban Sudanese Arabic'. In Janet Watson and Lutz Edzard (eds.), *Arabic grammar as a window on Arab humanism: essays in honour of Michael Carter*. Wiesbaden: Otto Harrassowitz Verlag. Pp. 155–195.

——. 2014. 'Associative meaning and scalar implicature: a linguistic-semiotic account'. *Linguistica* ONLINE, <http://www.phil.muni.cz/linguistica/art/dickins/dic-003>.

——. 2018. 'Tropes and translation'. In Adelina Hild and Kirsten Malmkjaer (eds.), *The Routledge handbook of translation studies and linguistics*. London and New York: Routledge. Pp. 208–222.

DICKINS, James, Sándor G.J. HERVEY and Ian HIGGINS. 2017. *Thinking Arabic translation* (2nd edn.). London and New York: Routledge.

EVERT, Stefan. 2007. 'Corpora and collocations: extended manuscript': <http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf>.

FIRTH, John Rupert. 1957. *Papers in linguistics 1934–1951*. London: Oxford University Press.

GIBBS, Raymond W. Jr. 2010, 'Idioms and formulaic language'. In Dirk Geeraerts and Hubert Cuyckens (eds.), *The Oxford handbook of cognitive linguistics*. Oxford and New York: Oxford University Press. Pp. 697–725.

GÓMEZ, Pascual Cantos. 2009. 'Attempting to model sense division for word sense disambiguation'. In Maria Manuela Cruz-Cunha, Eva F. Oliveira, Antonio J. Tavares, and Luis G. Ferreira, (eds.), *Handbook of research on social dimensions of semantic technologies and web services*, Vol. 2. (2 Volumes). Hershey: IGI Global. Pp. 126–157.

GRANT, Lynn. 2003. *A corpus-based investigation of idiomatic multiword units*. University of Wellington: PhD thesis.

GRANT, Lynn and Laurie BAUER. 2004. 'Criteria for re-defining idioms: are we barking up the wrong tree?'. *Applied Linguistics*, 25(1). Pp. 38–61.

GRIES, Stefan Th. 2008. 'Phraseology and linguistic theory: a brief survey'. In Granger, Sylviane and Fanny Meunier (eds.), *Phraseology: an interdisciplinary perspective*. Amsterdam and Philadelphia: John Benjamins. Pp. 3–26.

GRIES, Stefan Th. 2013. '50-something years of work on collocations: What is or should be next ...'. *International Journal of Corpus Linguistics*, 18(1). Pp. 137–165.

HAUSMANN, Franz J. 1989. 'Le dictionnaire de collocation'. In Franz J. Hausmann, Herbert E. Wiegand, and Ladislav Zgusta (eds), *Wörterbücher, dictionaries, dictionaries: ein internationales Handbuch zur Lexikographie*. Berlin: de Gruyter. Pp. 1010–1019.

HERBST, Thomas. 1996. 'What are collocations: sandy beaches or false teeth?'. *English Studies*, 77(4). Pp. 379–393.

HERVEY, Sándor G. J. 1982. *Semiotic perspectives*. London: George Allen and Unwin.

HILL, Jimmie. 2000. 'Revising priorities: from grammatical failure to collocational success'. In Lewis, M. (ed.), *Teaching collocation*. Hove: Language Teaching Publications. Pp. 47–69.

ISSA, Huwaida Jaber. 2014. *Proverbs, modified proverbs and curses in two novels of the Syrian coast*. University of Leeds: PhD thesis.

KILGARRIFF, Adam. 1992. *Polysemy*. University of Sussex: PhD thesis.

KÖVECSES, Zoltan. 2010. *Metaphor* (2nd edition). Oxford: Oxford University Press.

LABOV, William. 1972. 'The boundaries of words and their meanings'. In Charles James N. Bailey and Roger W. Shuy (eds.), *New ways of analysing variation in English*. Washington DC: Georgetown University Press. Pp. 340–373.

LEHECKA, Tomas. 2015. 'Collocation and colligation'. In Jan-Ola Östman and Jef Verschueren (eds.), *Handbook of pragmatics online*. Amsterdam and Philadelphia: John Benjamins. <https://benjamins.com/online/hop/articles/col2>. Pp. 1–23.

LEWIS, Michael. (ed.). 2000. *Teaching collocations: further developments in the lexical approach*. Hove: Language Teaching Publications.

LIU, Dilin. 2008. *Idioms: description, comprehension, acquisition, and pedagogy*. London and New York: Routledge.

LYONS, John. 1991. *Natural language and universal grammar: essays in linguistic theory 1*. Cambridge: Cambridge University Press.

MERRILL, Michael. 1995. 'Putting "capitalism" in its place: a review of recent literature'. *The William and Mary Quarterly*, 52(2). Pp. 315–326.

MOLLIN, Sandra. 2014. *The (Ir)reversibility of English binomials: corpus, constraints, developments*. Philadelphia: John Benjamins.

MULDER, Jan W.F. and Paul RASTALL. 2005. *Ontological questions in linguistics*. Munich: Lincom.

NESSELHAUF, Nadja. 2005. *Collocations in a learner corpus*. Philadelphia: John Benjamins.

NUNBERG, Geoffrey, Ivan A. SAG, and Thomas WASOW. 1994. 'Idioms', *Language* 70(3). Pp. 491–538.

Ó SÉAGHDHA, Diarmuid. 2008. *Learning compound noun semantics (Technical report, number 735)*. Cambridge: Computer Laboratory, University of Cambridge. <https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-735.pdf>.

O'DONNELL, Matthew Brook, Ute RÖMER, and Nick C. ELLIS. 2013. 'The development of formulaic sequences in first and second language writing: investigating effects of frequency, association, and native norm'. *International Journal of Corpus Linguistics*, 18(1). Pp. 83–108.

PASTOR, Gloria Corpas. 2017. 'Collocational constructions in translated Spanish: what corpora reveal'. In: Ruslan Mitkov (ed.). *Computational and corpus-based phraseology*. Berlin: Springer. Pp. 29–40.

PEIRCE, Charles Sanders. 1960. *Collected papers of Charles Sanders Peirce* (vols. 1–8). Hartshorne, C. and Weiss, P. (eds.). Cambridge (Mass.): Harvard University Press.

POPPER, Karl. 1986 [1957]. *The poverty of historicism*. London and New York: Ark.

POSIO, Pekka. 2015. 'Subject pronoun usage in formulaic sequences. Evidence from Peninsular Spanish'. In Orozco, Rafael, Carvalho, Ana Maria & Shin, Naomi (eds.), *Subject pronoun expression in Spanish: a cross-dialectal perspective*. Washington: Georgetown University Press. Pp. 59–74.

RICHARDS, Jack C., John PLATT, and Heidi WEBER. 1985. *Longman dictionary of applied linguistics*. London: Longman.

RUIZ YEPES, Guadalupe. 2017. 'Hybrid Methods for the Extraction and Comparison of Multilingual Collocations in the Language for Specific Purposes of Marketing'. In *Proceedings of Computational and corpus-based phraseology: recent advances and interdisciplinary approaches*. London: EUROPHRAS. Pp. 11–18. <http://www.tradulex.com/varia/Europhras2017-II.pdf>.

SAG, Ivan A., Timothy BALDWIN, Francis BOND, Ann COPESTAKE and Dan FLICKINGER. 2002. 'Multiword expressions: a pain in the neck for NLP'. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*. Mexico City. Pp. 1–15.

SAUSSURE, Ferdinand de. 1959. *Course in general linguistics* (translated by Wade Buskin). New York: McGraw-Hill Paperbacks.

——. 1975 [1916]. *Cours de linguistique générale*. Payot: Paris.

——. 1983. *Course in general linguistics* (translated by Roy Harris). Duckworth: London.

SINCLAIR, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

WANG, Xiaofei. 2018. *Quotation and truth-conditional pragmatics*. London and New York: Routledge.

WULFF, Stefanie. 2008. *Rethinking idiomaticity: a usage-based approach*. Continuum: London and New York.

WRAY, Alison 2002. *Formulaic language and the lexicon.* Cambridge University Press: Cambridge.

WRAY, Alison and Michael PERKINS. 2000. 'The functions of formulaic language: an integrated model'. *Language and Communication*, 20(1). Pp. 1–28.