Formulaicity and Creativity in Language and Literature

Edited by
Ian MacKenzie and Martin A. Kayman

First published 2018

ISBN 13: 978-1-138-72157-9

Formulaic sequences: a drop in the ocean of constructions or something more significant?

Andreas Buerki

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Formulaic sequences: a drop in the ocean of constructions or something more significant?*

Andreas Buerki

Centre for Language and Communication Research, Cardiff University, Cardiff, UK

ABSTRACT

This article investigates how formulaic sequences fit into a constructionist approach to grammar, which is a major post-Chomskyan family of approaches to linguistic structure. The author considers whether, in this framework, formulaic sequences represent a phenomenon that is sufficiently different to warrant special status or whether they might best be studied in terms of the larger set of all constructions found in language. Based on data drawn from a large corpus of Wikipedia texts, it is argued that it is extremely difficult to form a distinct class of formulaic sequences without creating highly arbitrary boundaries. On the other hand, based on existing theoretical claims that formulaic sequences are the basis of first language acquisition, a marker of proficiency in a language, critical to the success of communicative acts and key to rapid language processing, it is argued that formulaic sequences as constructions are nevertheless significant enough to be the focus of research, and a theoretical category meriting particular attention. These findings have key repercussions both for research primarily interested in formulaic language and phraseology as well as for construction grammatical research.

Introduction

One way of characterising formulaic sequences (FSs) is to say that they are expressions that represent the usual phrasings of a speech community (cf. Burger et al., 1982: 1; Coulmas, 1979; Erman and Warren, 2000; Fillmore et al., 1988; Howarth, 1998: 25; Langacker, 2008: 84; Pawley, 2001). Thus they include idioms (e.g. *live to tell the tale*), collocations (e.g. *hugely successful*), multi-word terms (e.g. *blind spot*), formulas proper, including discourse markers (e.g. *in other words*, a formula introducing a paraphrase), proverbs (e.g. *look before you leap*) and other usual sequences (e.g. *at the expense of X* or fillers like *you know*).

The place of FSs in linguistic theory has traditionally been a difficult one, particularly in regard to the structuralist and later Chomskyan traditions that dominated twentieth-century thinking. This is because FSs run counter to an understanding of linguistic structure as formed by combining words out of a store (the lexicon) according to general, combinatory rules. If

*The article is based in part on work presented at the 6th Formulaic Language Research Network (FLaRN) conference, held at Swansea University, 29 May 2014.

one tries to account for the examples of FSs given above while maintaining a division between rules and words, one soon finds that FSs do not fit in either category: they are not the products of general rules applying to words, and nor do they in general behave like single words. If *hugely successful* is the product of semantically appropriate words arranged according to general rules of combination, why does *greatly successful* not have the same ring to it? Yet it would be awkward to claim that *hugely successful* is on par with a single, word-like unit if only because it displays non-arbitrary internal semantic and formal structure. FSs like *blind spot* might be more amenable to treatment as one big word rather than a combination, but the 'big word' analysis runs into even greater difficulties with FSs like *at the expense of X*, where modifications are clearly possible. The Google n-gram corpus tells us, for example, that the following are common instances of this FS: *at the possible / sole / joint / least / common / private / public / whole / trifling / small expense of X*.[1]

Partly out of this dissatisfaction with a model of linguistic structure that operates with words and rules, a new family of approaches to syntactic structure has emerged and firmly established itself as a field of bustling research activity. This family is now typically referred to as constructionist (or constructivist) approaches, or construction grammars. Constructionist approaches deny that there is in principle a distinction between words and rules, stressing that in reality we find structures on a continuum between these poles, albeit at various connected levels of lexical specificity (or substantivity): whether we have an entirely substantive construction like *blind spot*, or a much more abstract (schematic) construction like the general pattern < adjective + noun > (of which *blind spot* is in some respects a specific case), we are dealing with the same type of thing, namely a construction. If this seems like phraseology taking over the rest of grammar, then in many ways that *is* the idea, although the point is more that constructionist approaches are able to explain FSs and non-FSs using the same theoretical machinery. At the very least, therefore, constructionist approaches appear to offer a proper theoretical home for FSs in an overall theory of grammar, as will be shown in greater detail below. But this raises a different question – the one we will be addressing here: are FSs now merely one type of construction among many along the continuum (a drop in the ocean of constructions), or are FSs still marked out in some way as special, different, demanding particular attention or special treatment? The answer is of considerable importance if there is to be more than a superficial coming together, or indeed integration, of research into formulaic language and constructionist approaches.

Despite its importance, the question has hitherto received scant attention. Although constructionist approaches grew out of a focus on idioms and constructions of limited productivity (e.g. Fillmore, Kay and O'Connor, 1988; cf. Wulff, 2013: 274), it appears that construction grammarians are now most interested in limited-productivity constructions at intermediate levels (represented by constructions like < the Xer, the Yer >) and above, including substantial work at the highly schematic level (cf. Goldberg, 1995; Hilpert, 2014, esp. ch. 2). Constructions at the more substantive level, especially compositional and non-idiomatic ones, have largely been neglected. These phenomena have, on the other hand, been central to work on formulaic language; but where connections to constructionist theory have been made, they have on the whole remained superficial, assuming links to be either unproblematic or of no serious relevance. However, there are now signs that this division in interests is diminishing, making it timely to address this question specifically.

In the following, I will begin by defining, in more detail, what FSs are, and outlining in particular some of the reasons why they have been identified as a theoretically significant

phenomenon. This is followed by an account of what a constructionist approach to grammar involves in general, and how FSs can be seen to fit into it. The extended investigation entitled *Empirical Boundaries* subsequently explores to what extent FSs can be considered an empirically well-bounded and therefore potentially distinct and special phenomenon from a constructionist viewpoint. In a concluding section, I answer the question in the title of the article by synthesising the findings of the empirical investigation with existing theoretical claims regarding the particular importance of FSs.

Background

Formulaic sequences

FSs have been understood, and consequently defined, in different ways by the various traditions and fields of enquiry in which they have played a role (and indeed across these traditions). Simplifying considerably, three interrelated strands of thinking can be identified. In traditional European phraseology with roots in Soviet scholarship, the triple criteria of polylexicality (involving more than one word), idiomaticity (semantic non-compositionality and/or non-adherence to syntactic regularities) and conventionality (or stability, typically understood to result in fixedness) have been influential in conceptualising FSs (e.g. Burger, Häcki Buhofer and Sialm, 1982).

All of these criteria are recognised as problematic if applied rigidly. The polylexicality criterion appears to be unduly at the mercy of the whims of orthography: *albeit*, *somebody*, *anyway* are all transparently of polylexical origin and more generally 'corpus evidence shows … that the division between multiword and single-word items is blurred, to say the least' (Moon, 1998: 81). It is nevertheless a criterion that is still broadly adhered to in all strands of thinking on FSs, if only for reasons of practicality. The criterion of idiomaticity similarly continues to be of importance as a focal point of research interest within and beyond traditional phraseology, and is also notably prominent in natural language processing (cf. Manning and Schütze, 1999: ch. 5; Sag et al., 2002; Villavicencio et al., 2004). However, it is widely recognised that from a semantic standpoint, it can be exceptionally difficult to make a conclusive assessment of idiomaticity, and such assessments typically depend on an often questionable abstraction of word senses away from contexts (Fleischer, 1982: 38) or judgments on what is literal or metaphorical (Bybee, 2006: 713). From a syntactic viewpoint, the expectation that the whole of syntactic structure is fully regular is arguably quite unrealistic. Secondly, and even more importantly, with the availability of large corpora and the large-scale corpus-linguistic exploration of phraseological phenomena, it has become clear that idiomatic FSs are vastly outnumbered by conventional, non-idiomatic sequences that should nevertheless be considered FSs (among the first studies were Altenberg and Eeg-Olofsson, 1990; Jones and Sinclair, 1974; Renouf and Sinclair, 1991). Finally, conventionality, understood as a high degree of fixedness, has also been shown to be difficult to apply as an absolute criterion – even supposedly fixed phrases and idioms have been shown to be remarkably flexible (e.g. Dutton, 2009; Kuiper, 2007; Langlotz, 2006). And yet again, restrictions on flexibility remain a defining feature of FSs in much current work (cf. Columbus, 2013).

In a second strand of thinking, which could be dubbed the processing view, the aspect of mental processing features prominently: John Sinclair (1991) described relevant entities in terms of 'phrases that constitute single choices' (110), and descriptors such as *prefabs*,

pre-constructed and *pre-fabricated* are frequently used, as for example in the following definition by Alison Wray, who coined the term *formulaic sequence* and originally defined it 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' (2002a: 9).

While the elevation of the manner of mental processing to the status of the most characteristic feature of FSs is important in terms of theory, it also leads to a number of difficulties. Since processing occurs in individuals' heads, formulaicity could be (mis?)understood as primarily a feature of idiolect – a feature, furthermore, which it is difficult to pin down in individuals even in experimental conditions (Durrant and Doherty, 2010), let alone identify in the shared language system, or language that has already been produced – hence Wray's careful wording 'is, or appears to be'. The question of what it is that makes FSs appear prefabricated, however, remains to be explained.

The final line of thinking, which is followed in the present article and found across various traditions, focuses on the aspect of conventionality in relation to speech communities and as manifested in language use. It can be summed up in the characterisation of FSs given above, namely as expressions that represent usual ways of putting things in a community. Charles Bally, one of the fathers of modern research into phraseological phenomena, referred to 'usual groupings' (Bally, 1909: 70, my translation) and 'combinations sanctioned by usage' (73). More recent work in this line of thinking has described FSs as 'conventional phrases' (Pawley, 2001: 122), 'combination[s] of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization' (Erman and Warren, 2000: 30), or 'institutionalised phrases' (Bybee, 2010: 35; cf. also Howarth, 1998: 25 and Brunner and Steyer, 2007: 2). The idea of conventional ways of putting things implies that there are both 'things', i.e. units of meaning (that might contain other units within them), and specific linguistic forms (written or spoken) conventionally associated with those meanings. This leads quite naturally to the statement in (1) which will serve as the characterisation of FSs that underlies the investigations in the remainder of this article:

> (1) Formulaic sequences are phrases that are conventional pairings of form and unit of meaning in a speech community.

Moving beyond views on how FSs are to be characterised, it is also of key importance in the context of the overall question addressed in this article to review claims regarding the importance of FSs as a phenomenon: what theoretical justifications, if any, would suggest that there is anything special about the phenomenon of FSs specifically? Space permits only a brief consideration of four aspects here, but these are, in my view, sufficient to indicate a satisfactory answer to this question.

In terms of language processing, psycholinguistic research suggests that, owing to their nature as units, at least some FSs may be processed more speedily than constructions that are assembled at the time of use (Ellis et al., 2009; Tremblay et al., 2011). In fact, it is thought that the use of FSs is nothing less than the mechanism that enables fluency in speech, bypassing unnecessary and cognitively expensive syntactic operations (Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Wray, 2002a: 35–7; cf. also Arnon and Snider, 2010). From a theoretical point of view, therefore, FSs have processing features that distinguish them from other constructions, even if these features are graded and do not therefore result in a sharp binary distinction.

In first language acquisition, cognitive and usage-based models show that acquisition starts with 'lexically specific phrases and [children then] gradually build up a repertoire of increasingly abstract constructions' (Dąbrowska and Lieven, 2005: 437; cf. also Dąbrowska, 2014; Lieven and Brandt, 2011; Tomasello, 2005). FSs are therefore, in an individual sense, where it all starts for language. This again differentiates FSs from other syntactic constructions. The question also arises as to why FSs are evidently not just a stage in language acquisition, but remain a feature at full proficiency. Which sequences remain formulaic and why, and which progress to more abstract syntactic representation only? These questions that centrally involve FSs remain relevant for a fuller understanding of how language works, not least because FSs have typically been looked at more in terms of why they fail to remain abstract and become fixed (e.g. Wray, 2009: 33), though it is clear that processes take place in both directions (cf. Bybee, 2010: ch. 3), at least partly depending on whether one looks at the language of the individual or the shared language. This view from first language acquisition theory therefore confirms that FSs hold special significance among syntactic constructions.

In second language acquisition research, a vast number of findings have shown that a lack of a sufficiently large stock of FSs in one's linguistic knowledge is a particularly noticeable point of divergence between fully proficient (typically L1) speakers and less proficient L2 speakers (Allerton, 1984; Bally, 1909: 70–3; Boers et al., 2006; Jespersen, 1904; Myles, 2004; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Sorhus, 1977; Wray and Perkins, 2000).

Further, research suggests that largely substantive constructions are key to successful communicative acts because they activate a range of social, situational and cultural contextual cues (Erman, 2007: 26; Feilke, 1994; 2003: 213; Wray, 2008: 20–1). Even in lingua franca communication among L2 speakers, communities move fast to establish a stock of FSs to aid mutual understanding, as shown by Barbara Seidlhofer (2009).

The strands of thinking on FSs outlined above highlight various elements of definition and characterisation that are thought important in current linguistic research. While they also demonstrate that there are continuing challenges in seeking to define FSs, considerations from the areas of language processing, first and second language acquisition and communication suggest that FSs are a phenomenon that is marked out on account of its theoretical significance and distinctiveness, if by nothing else.

Constructionist approaches to grammar

This section presents a constructionist approach largely based on Adele Goldberg (2003; 2006) and Martin Hilpert (2014); other flavours of construction grammar might differ in specifics. For a fuller introduction than can be provided here, the reader is referred to those publications, and to Thomas Hoffmann and Graeme Trousdale (2013).

Traditional and in particular Chomskyan approaches to linguistic structure tend to start by establishing general rules that then need a large number of complex constraints to plausibly account for the linguistic knowledge that a speaker is thought to possess in order to produce and understand grammatical utterances in a given language (e.g. phrase structure rules in Chomsky, 1957). In contrast, constructionist approaches start with the particular and build up more general patterns if and to the degree to which more general patterns are required – typically *in addition to* rather than instead of the particular. This leads to a model of linguistic knowledge that incorporates redundancy and a recognition that much (if not

most) of language is not fully regular, as well as a denial of a distinction between rules on the one hand, and particular entities on which rules operate, on the other (Hoffmann et al., 2013: 1). This dichotomy is overcome by recognising constructions as theoretical entities that span the continuum between more rule-like and more fixed entities and crucially everything in between. How so?

Take an utterance such as *I missed the bus*. For the moment, we will assume this is an instance of the construction < I missed the bus >. Any construction, including this one, is held to be a linguistic sign in the Saussurean sense. That is, an arbitrary but conventional pairing of a *signifier* (i.e. a form) with a *signified* (i.e. a unit of meaning) (de Saussure, 1974: 65–70). Though the linguistic sign is generally taken to apply to units of the size of a word or morpheme, constructionist approaches have extended it to constructions. A construction like < I missed the bus > is fully lexically specific, or *substantive* in constructionist parlance. It could be argued that there exists a more abstract (more *schematic* in the jargon) construction < SUBJ. MISS SOMETHING-THAT-RUNS-ON-A-SCHEDULE > which would cover such instances as *He missed the last train* or *passenger Jones will miss his flight*, but not *I miss you*. At an even more schematic level, probably skipping a number of intermediate levels, we have < SUBJECT VERB OBJECT > (the transitive construction), which would cover all the examples mentioned and very many more. These constructions are related in that they are increasingly general with fewer and fewer specifics 'passed on' from more substantive constructions to more schematic ones, but with even the most schematic constructions sharing some aspects with less schematic ones and crucially remaining linguistic signs with a form and meaning part.[2] When looking at an intermediate construction like < SUBJ. MISS SOMETHING-THAT-RUNS-ON-A-SCHEDULE >, it is found that constituents like SUBJ. or SOMETHING-THAT-RUNS-ON-A-SCHEDULE are also constructions. Indeed, although MISS can be realised as a single word (e.g. *missed* as in the examples), single words are still (morphological) constructions and even single morphemes are accorded identical status to constructions, and called constructions as well (e.g. Goldberg, 2006: 5). One might ask why both the mentioned intermediate construction and the fully substantive construction < I missed the bus > should exist if the latter can plainly be derived from the former (and other constructions). The utterance 'I missed the bus' may indeed be a mere instance of the medium-level construction set out above, with no fully substantive construction present in the system (that is, it may be a *construct* rather than a construction). Most construction grammarians would, however, argue that frequent use can be enough to warrant the existence of substantive constructions that are derivable from other existing constructions (Goldberg, 2003: 219; Hilpert, 2014: 12–13). In some speech communities, there will also be the homonymous construction < SUBJ. MISS the [bus / boat] >, 'to miss an opportunity', which cannot be derived from other constructions. Highly schematic constructions like the transitive construction are also needed to enable the flexibility required for the creation of novel utterances. There may be other reasons too, such as 'a vague idea entertained by speakers who are analytically minded enough to see similarities between different kinds of construction' (Hilpert, 2014: 69).

Constructionist approaches to grammar, then, see linguistic knowledge as a store of constructions (Goldberg, 2003: 219; Hilpert, 2014: 2) offering 'a graceful transition from idiom-like holophrases to fully abstract argument constructions' (Dominey, 2006: 137), including constructions that feature elements of various different levels of schematicity. One of the beauties of this conception is that it can account for and explain the general regularities

evident in language as well as the more usual limited-productivity structures and even the fixed expressions that are the hallmarks of competent, idiomatic language use. Another is that it does away with hyper-compartmentalisation as constructions incorporate morphology, syntax, semantics, phonology and even pragmatics.

A drop in the ocean?

We have identified FSs as phrases that are conventional pairings of forms with units of meaning in a speech community. As such, they easily fit into a constructionist framework, where the remainder of grammar also consists of pairings of form and meaning. In this view, FSs appear as a particular kind of construction, namely the kind that is predominantly lexically substantive. Before considering the position of FSs within this framework in some more detail, it is worth recalling what a radically new position for FSs this inclusion within grammar as bona fide members is: as already outlined, in a model of linguistic knowledge that divides between rules and words (i.e. syntax and lexicon) there is no place for FSs, and they have often been shoved into the lexicon because of their irregularities, typically ignoring their more productive aspects as well as their importance. Traditional phraseology, with idiomaticity seen as an important aspect, has not revolted too much against this placement. As the importance of non-idiomatic FSs was increasingly recognised, new models of the relationship to the rest of linguistic knowledge surfaced. Sinclair (1991) proposed that two principles were responsible for language production and reception abilities: the open-choice principle (the traditional syntax and lexicon model), responsible for novel expressions, and the idiom principle, according to which existing, pre-constructed phrases are called upon to construct or interpret utterances. While speakers can easily and frequently switch between these two principles, Sinclair (114) argued that 'there is no shading of one into another . . . . The models are diametrically opposed' and therefore not integrated. Another proposal of how FSs connect to the rest of grammar is Wray's Needs-Only-Analysis model (2002a; 2008). Wray remains non-committal toward any particular overarching theory (2008: ch. 7), but the model requires that processes of production and reception respect the integrity (i.e. non-analysis) of items that could be analysed as complex, unless there is a specific reason to analyse their internal make-up (2002a: 130). Needs-Only-Analysis is enabled by a heteromorphic lexicon which holds fully fixed items as well as items that are in need of modification, such as items with gaps or ones that need morphological or other fine-tuning (cf. 2008). Here, FSs are grouped with words and morphemes and allowed a degree of underspecification, yet they remain a distinct phenomenon, different in kind from linguistic rules and therefore in a distinct corner.

The constructionist approach, on the other hand, adds FSs to its pool of constructions, entirely on a par with other constructions that collectively represent the totality of linguistic knowledge. This view puts phraseologists into a position where they no longer need to fight to establish a space for FSs in models of linguistic knowledge. If anything, the question is rather whether it still makes sense to think of FSs as a distinct and special phenomenon at all – one that merits particular interest and investigation – or whether FSs do in fact blend in with constructions in general to the degree where there is little justification in looking at them specifically as a group. We have already seen that there are strong theoretical claims that suggest FSs are distinct, but are FSs also marked out as distinct in empirical data? This key question is addressed in the next section and the answers obtained will, together with existing claims about FSs reviewed above, enable us to draw definite conclusions about the place of FSs among constructions in the final section of this article.

Empirical boundaries

This section seeks to ascertain how empirically well-bounded a phenomenon substantive constructions or FSs are, within the larger set of all constructions. It is easier to argue for a distinct role for FSs if they are empirically well-bounded – if they form an easily identifiable set – against the mass of constructions that constitute linguistic knowledge in a constructionist framework. From a constructionist point of view, the distinguishing feature of FSs among all constructions is that they are predominantly lexically substantive. Framed in this way, the hypothesis is that if FSs are empirically well-bounded, schematic elements (SEs), also known as gaps or slots, will not feature prominently among them, though it is clear from previous research that there are FSs that feature SEs (e.g. Altenberg, 1998; Mittmann, 2004: ch. 6; Martinez and Schmitt, 2012). FSs that feature SEs can be seen as marking more peripheral regions, or the fuzzy edges of the phenomenon, where FSs and other constructions start to transition into each other. Below I set out how this hypothesis was tested.

Data and procedure

FSs in a corpus of 29 million words of English-language Wikipedia texts were extracted and the degree to which they contained SEs was assessed via a random sample. The data were obtained by downloading the English- language Wikipedia dump (a file containing all Wikipedia pages) of 4 February 2013, splitting it into individual articles, selecting a subset of 63,075 individual articles (out of the total 4.2 million) and converting Wikipedia's XML format into plain text, removing all non-textual information. This was accomplished using WikiExtractor 2.2 (Attardi and Fuschetto, 2012). The median document length was 461 words after conversion. Only documents with a minimum length of 10 words were included. The total word count for all documents was 29,077,310.

For the purposes of automatic identification and extraction, and in line with the characterisation of FSs as 'phrases that are conventional pairings of form and unit of meaning in a speech community', I operationalised FSs, following Buerki (2012: 269), as in (2) below.

(2) Frequent word sequences forming a semantic unit.

This operationalisation measures conventionalisation via frequency of occurrence; *frequent* was defined as occurring at least twice per million words. *Word sequences* were defined as sequences of consecutive orthographic word forms of two to nine words in length. The following two exceptions to the use of word form sequences were allowed: 1) the label NUM was used in place of numbers; 2) the names of the 12 months of the year were replaced with the label MONTH. A *semantic unit* was defined as a word sequence possessing the sort of semantic unity typically found in words and structurally complete phrases. Semantic unity was also attributed to sequences that, while lacking this unity, can acquire it through the addition of a single, semantically or formally restricted SE (such as when *in search of* does not form a full semantic unit unless an SE on its right edge is added, i.e. *in search of X* where *X* is restricted semantically to something prized that is being pursued). In restricting the number of this type of SE to a single X, this operationalisation errs on the side of caution: if more Xs were allowed in a sequence, the degree of measured schematicity would be higher. In addition, although more word sequences could be identified as formulaic, work on the data set indicated that an undesirable decrease in the reliability of inter-subjective identification of FSs would result if more Xs were allowed.

The extraction of FSs from corpus data proceeded in the three steps outlined in Figure 1. The extraction of recurring word sequences (step 1) was carried out using the N-Gram Processor (Buerki, 2013). Sequences across sentence (and sentence-equivalent) boundaries were blocked and an additive stop list (eliminating sequences that consist entirely of stop-listed items) was used. The stop list contained the 150 most frequent word forms of English according to the Leipzig Corpus Portal (anon, 2001). In step 2, various lengths of word sequences were consolidated into one list using SubString 0.9.5 (Buerki, 2011). At step 3, a lexico-structural filter with 52 filter entries was applied to remove word sequences with properties that are likely to render them non-compliant with (2) above. A detailed discussion of the procedure (applied to a different data set) is found in Buerki (2012). In total, this procedure automatically extracted 31,075 sequence types out of the corpus.

Extraction accuracy was established as follows. A random sample (n = 1000 types) of automatically extracted sequences was rated for compliance with the operationalisation in (2) by the author and independently by a research assistant. Since frequency was automatically controlled, sequences only needed to be assessed for semantic unity. For this purpose, raters were first given a set of instructions describing the notion of semantic unity and showing examples of word sequences with and without semantic unity. The two-page handout used for this purpose is supplied as part of the online supplementary materials for this article. After a short test run, raters were asked to judge word sequences as either possessing semantic unity or not possessing it. This was done using a simple computer program which presented raters with a word sequence and asked for the rating. Ratings by the two raters agreed in 77.1% of the 1000 cases rated, indicating that the notion of semantic unity is sufficiently inter-subjectively robust for the purposes at hand. Conferring with each other subsequently, raters were able to resolve rating differences on the basis of the guidelines in an additional 178 of the 229 cases where initial ratings were different. The author made the final decision in the remaining cases. In this manner, 763 of the 1000 types in the sample were deemed to comply with the operationalisation set out in (2), resulting in an extraction accuracy of 76.3%. Comparisons of extraction accuracy and recall among automatic extraction procedures for phraseological phenomena are highly problematic owing to differing underlying operationalisations, but this accuracy figure is comparable to similar documented procedures. A discussion of issues involved in assessments of accuracy and recall is to be found in Buerki (2012), and in more general terms in Christopher Manning and Hinrich Schütze (1999: ch. 5).

Most of the analyses reported below are based on a random sample of 1000 sequences out of the total 31,075 automatically extracted sequences. There are principally two reasons why the use of a large corpus is nevertheless advantageous: first, all of the extracted sequences were used in one of the analyses (see below) and second, identifying and extracting FSs, based in part on frequency considerations as they are, can be accomplished with

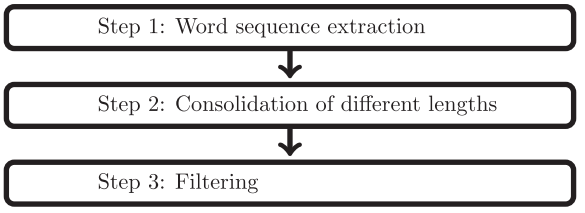

Figure 1. Steps of the extraction.

greater accuracy and recall if a larger corpus is used, even if only a subset of extracted sequences is subsequently analysed.

To determine the degree to which FSs featured SEs, a random sample of 1000 extracted sequence types (the same sample as was used to establish the accuracy of the extraction procedure, above) was assessed regarding the inclusion of SEs, that is, constituents which are not lexically fixed, but fixed at a more abstract level. A higher degree of schematicity would indicate a less well-bounded phenomenon that blends in with other, more abstract constructions, whereas a low degree of schematicity would indicate that FSs as a group are well-bounded vis-à-vis more schematic constructions. Sequences in the sample that did not comply with the operationalisation in (2) were excluded from analysis, resulting in 763 sequences assessed. Three kinds of SEs were considered:

- The labels NUM and MONTH. These are schematic for any particular cardinal or ordinal number and any of the 12 months of the year (e.g. *as many as NUM* or *on NUM MONTH NUM it was announced that X*).
- SEs that were needed to complete the semantic unit at either edge of an extracted sequence (e.g. *a selection of X; X for Best Actress*). Table 2 lists further examples. Importantly, SEs of this kind were assigned only if absolutely necessary to complete the semantic unit. For example, the sequence *the band played* occurs in the corpus both transitively (*the band played a pre-recorded version of 'Feel Good Inc.'*) and intransitively and therefore it cannot be classed as featuring an SE of this kind at its right edge – *the band played* is a bona fide semantic unit even without addition. SEs of this kind were marked as present at the same time as semantic unity was assessed. Agreement between the two raters regarding the placement of these elements in sequences deemed by both to possess semantic unity was 81%. Where there were differences between raters, the final decision on placement was made by the author.
- Optional medial SEs. An optional medial SE was deemed present if a search in the source corpus revealed additional sequences identical to the sequence under investigation except for the presence of medial extra material and if the number of additional sequences of this sort was at least 10% of the frequency of the sequence under investigation and above a frequency of 10, whichever is higher. Only medial material of one to three words in length was considered. For example, the sequence *the film industry* occurred 65 times in the corpus. A search revealed an additional 35 sequences with an optional medial SE (realisations given in square brackets): *the [Bengali / West German / growing Hindi / television and / general / early 1960s Malayan] film industry*. Since 35 is more than 10% of 65 and also more than 10, the original consecutive FS was considered to have an optional medial SE. Frequencies were not adjusted, that is, they continue to represent the number of times an FS occurred in its continuous form, excluding the additional occurrences *with* optional SEs in place.

The number of sequences in the sample containing SEs of one or more of any of the above three kinds was counted and set in relation to the total number of sequences in the sample.

To further the detailed analysis of SEs, all FSs were categorised using a working typology of six FS categories. While FSs have been categorised in very many different ways (for overviews see e.g. Fleischer, 1982: ch. 3; Granger and Paquot, 2008) and categorisations have been criticised for producing neither discrete categories nor fully coherent taxonomies (e.g.

Wray, 2002a: ch. 3), the typology outlined below is important in illustrating that FSs are a rather heterogeneous collection of items, not least with respect to their degree of schematicity. Recognising that there are other valid ways of categorising FSs, the typology comprises the following:

- Formulas: FSs that perform functions, including discourse functions, e.g. *I'm sorry* (apologising); *yours faithfully* (ending a letter); *in summary* (introducing a concluding section) (cf. esp. Coulmas, 1979, 1981).
- Collocations: a base and one or more collocates that are used together, such as the base *teeth* which is used with *brush* (cf. Hausmann, 1991).
- Multi-word terms: typically express a single concept (e.g. *dual carriageway*), often technical or semi-technical terms (e.g. *aspect ratio*; *the proletariat*) and some proper names (e.g. *the Nintendo DS*). Phrasal verbs and periphrastic expressions are also part of this category.
- Idioms: FSs with clearly non-compositional meaning (e.g. *pull someone's leg*). Although many idioms feature sanctioned variants, the semantics typically depend on the exact words used.
- Proverbs: 'short sentences of wisdom' (Mieder, 2004), e.g. *garbage in, garbage out*.
- Usual sequences: other word sequences that instantiate usual ways of putting things – these are word sequences that are largely compositional in their semantics and may include collocations or multi-word terms within them (e.g. *a wide variety of X*, where *wide variety* is a collocation, or *in Eastern Europe*, where *Eastern Europe* is a multi-word term).

It was not expected that the sample would contain many examples, if any, of idioms and proverbs as they have been shown to be rare in actual language use (Moon, 1995: v; 1998: 81). Further, as these are prototype categories with fuzzy boundaries, occasionally an FS could reasonably be assigned to more than one of them. However each FS was only assigned to one category, disregarding any secondary possibilities. Examples for each category are shown in Tables 2, 3 and 4. The full data set of analysed FSs is part of the online supplementary materials for this article.

Finally, the FS density of the corpus was determined by counting the words bound up in all automatically extracted sequences and reducing this count to the accuracy percentage established in the sample, then comparing it to the total word count of the corpus. For the purpose of this calculation, SEs were not included in the word count for FSs (since they were of variable extent), save for the NUM and MONTH labels, which were counted as part of the word count of words bound up in FSs.

Results

In terms of the working typology of FSs, as expected, no idioms and proverbs were among the extracted sequences in the sample. Counts in the remaining categories are shown in column 2 of Table 1 and reveal usual sequences to be the most common category of FSs in the sample (64% of types), followed by multi-word terms (24%), collocations (11%) and the 'formula' category (1%). For all figures in Table 1, the numbers for tokens (taking into account how often individual FSs occurred) presented a closely similar picture and are therefore not shown.

Table 1. Formulaic sequence types with schematic elements (SEs) by category (sample n = 763 types).

Category	n	Non-optional		Optional		Any SE		> 1 SE	
Formulas	8	6	(75)	1	(13)	6	(75)	2	(25)
Collocations	87	73	(84)	35	(40)	76	(87)	32	(37)
Multi-word terms	183	34	(19)	40	(22)	70	(38)	6	(3)
Usual sequences	485	168	(35)	225	(46)	337	(69)	67	(14)
All	763	281	(37)	301	(39)	489	(64)	107	(14)

Note: Percentages in parentheses, relative to column 'n'.

Table 2. Examples of formulaic sequences with non-optional schematic elements.

Category	Examples	Examples of schematic element realisations
Formulas	this means that X	young children learn difficult tasks better if...
	collectively known as X	romance languages / cipher notations / the judiciary
	with the exception of X	glycine / the Finns / a few small private railway lines
	to note that X	'downloading' is not the same as 'transferring'
Collocations	threatened with X	extinction / disaster / lawsuits / excommunication
	tired of X	his music / her detective Poirot / taking care of Gregor
	relationship with X	his mother / a German-born English woman / Clara / him
	imprisoned for X	armed robbery / doing so / endangering state security
Multi-word terms	Ariane NUM	1 / 2 / 3 / 42P / 42L
	School of X	economics / business / journalism / design / law
	average annual X	growth rate / consumption / temperature / inflation
	X for Best Actress	Golden Globe Award / BAFTA Award / nomination
Usual sequences	placed NUM	5th / 6th / 4th / 15th
	spent most of X	his childhood / the rest of his life / 1917 / their time
	it is unknown X	how many more perished / who are financing the PVV
	as many as NUM	50,000 / 35 / six / a couple of tens of thousands

Note: Examples for schematic element realisations are non-exhaustive.

In terms of non-optional SEs, of the 763 FSs in the sample, 281 (37%) contained at least one SE (95% confidence interval: 33–40%). As shown in Table 1, the SEs were not spread equally across the different categories of FSs: multi-word terms contained the fewest (only 19% of types contained one), other attested categories contained considerably more non-optional SEs. Table 2 shows typical examples of SEs in the sample.

Corpus searches for optional slots revealed that many of the extracted FSs contained optional SEs. There was some variation among different categories of FSs in terms of the occurrence of optional SEs: multi-word terms again show a low proportion of FSs with optional SEs, although figures for the 'formula' category are the lowest. Overall, at 39% (95% confidence interval: 36–43%), the proportion of FSs with optional SEs was similar to the proportion with non-optional SEs. Examples of optional SEs in various categories of FSs are shown in Table 3. Similar to the case of sequences containing non-optional SEs (cf. Table 2), FSs with optional SEs differ widely in the degree to which they restrict those elements: some sequences only permit what seems a finite number of substantive elements or only one particular element. *The Nobel Prize*, for example, allows two optional elements but the first is restricted to a year (e.g. *The 2001 Nobel Prize*), the second to either *Peace* or *Memorial* (e.g. *The Nobel Peace Prize*). A similar case is the sequence *are divided into NUM*, where the only optional element recorded in the corpus was *further* (i.e. *are further divided...*). A medium level of restriction could be described as cases where optional SEs form a restricted set of semantically related items, as in the sequence *relies [heavily / largely / substantially / entirely / only] on X*. One could argue that *relies* in *relies on X* is an SE as well because the sequence is also instantiated as *rely on X*, the two being instances of a single type. For the present investigation, schematicity at the morphological level is disregarded since it rarely plays a

Table 3. Examples of formulaic sequences with optional schematic elements.

Formula	With the [further / occasional / sole / possible / notable] exception of X
Collocations	noted [in an article / by one researcher / the fact / in 2004 / at the time] that X drawing [heavily] on X is [broadly / entirely / relatively / also / not / otherwise] consistent with X varies [inversely / significantly / nonlinearly / strongly / periodically] with X
Multi-word terms	the [UK album / top of the / US and UK / German / pop / Oricon] charts the [Warsaw Treaty / largest standing / regular / police and / Iranian] armed forces the [Provisional] IRA
Usual sequences	was [later / mistakenly / struck and / captured and / also / reportedly] killed a [wide / narrow / diverse / small / limited / largely random] selection of X was [speedily / famously / subsequently / first] translated into X on [false / supposed / fabricated / 11 / two / fresh] charges of X could [no longer / not / probably / still / well / only / nominally] be considered

Note: Examples for schematic element realisations are non-exhaustive.

Table 4. Examples of formulaic sequences without schematic elements.

Formulas	in this respect / a few days later
Collocations	in parentheses / high frequency / most visited / most valuable / took charge / on film / several ways / in large numbers / much less
Multi-word terms	quality of life / Julius Caesar / hard drives / Football Association / Buenos Aires / House of Commons / App Store / natural history / show up / stay on / has grown
Usual sequences	in large part / not fully / on the radio / after the match / in many cases / on his own / at the time of his death / listed below / earthquake and tsunami / in its entirety

major role in predominantly isolating languages like English. Cross-linguistically, it seems clear that schematicity at the morphological level needs to be accounted for as well, further increasing the overall degree of schematicity of FSs. Examples of sequences that impose only very general restrictions are also found, e.g. *as a [PhD / compromise / strong / moderate / presidential / possible] candidate*. In some cases, the nature of the optionally inserted material is similar in character to an interruption, because very little is shared between fillers, e.g. *lawsuit [for trademark infringement / filed / brought / in July 2011] against X*. Care was taken, however, not to posit an SE in cases where inserted material altered the meaning such that it was incompatible with the base FS, e.g. *for a population of less than X* is not an instance of *for less than X*. Overall therefore, a continuum of degrees of restriction applying to SEs was found, some being fairly substantive (with a choice of only a few specific instantiations, or the presence versus absence of a single lexically specific optional element), others far more schematic. This again illustrates the partial and graded degrees of productivity typical of constructions in general.

Looking at FSs containing any kind of SE (column 5 of Table 1) reveals that overall, no fewer than 64% of FS types (95% confidence interval: 61–67%) contain at least one SE (optional or not). Notably, this column does not add up to the sum of the preceding two columns because some FSs contain both non-optional and optional SEs. Again, there are notable differences between different categories of FSs.

As seen from the figures in the rightmost column of Table 1, 14% of sequence types in the sample featured multiple SEs (95% confidence interval: 12–17%). Given that bigrams (sequences comprising two words) made up 69% of the 763 FSs in the sample, this is again a remarkable figure.

Finally, an estimate of FS density based on automatically extracted sequences, adjusted for accuracy, suggests that around 55% of running words in the corpus are part of FSs as defined in (2) above. The calculation runs as follows: (total words part of automatically extracted sequences * accuracy based on sample) / total words in the corpus = (20,887,932 * 0.761)/29,077,310 = 0.54667.

Discussion

Looking to ascertain how well FSs are empirically bounded as a distinct phenomenon among the group of all constructions, this analysis measured the degree to which FSs as a group contain SEs. It is important to note that, to give FSs the best chance at standing out as distinct, we took as a starting point a very conservative operationalisation of FSs (only consecutive word form sequences were extracted). This is despite the general acknowledgement that FSs are more flexible than that (e.g. Becker, 1975: 62; Wray, 2002a: 269; 2006: 592). Additionally, the door was left open only a tiny bit to the possibility of SEs occurring by the inclusion of NUM and MONTH labels, while optional and non-optional SEs remained very tightly restricted. Given these restrictions, conditions were favourable to FSs showing a clear central area of lexically substantive constructions, with the likelihood of a fuzzy area of partly schematic constructions at the edges. What the results of the analysis have shown, however, is that the door left open a tiny bit was forced wide open by the empirical data and that SEs are rife among FSs: roughly two thirds of FS types in our sample contained at least one SE. The 95% confidence interval indicates that the figure will lie between 61% and 67% in all similar data outside the sample. While the figure for FSs containing only non-optional SEs is lower at 37%, the inclusion of optional SEs is important as they are a notable feature of some FSs but not others and disregarding them would therefore be unfaithful to the empirical reality of FSs as they occur in language.

The high degree of schematicity among FSs, then, shows them to blend in with other constructions that make up linguistic knowledge in the constructionist view of grammar, indicating that they are a very poorly bounded phenomenon. If fully lexically substantive constructions are taken as the prototypical FS (their substantiveness being the only realistic distinguishing feature from a constructionist point of view), then only 36% of FS types in our sample are actually 'typical' – the overwhelming majority inhabit the fuzzy transitional area to more schematic constructions. Even in the case of the most substantive category of FS, the multi-word term, 38% of types (95% confidence interval: 31–46%) contained SEs, making them constructions that allow various interesting substantive patterns (cf. Table 2). It therefore seems clear that FSs are not sufficiently empirically well-bounded to stand out as a natural group, deserving of special treatment or special interest on account of their distinctiveness as a phenomenon.

At this juncture, it is worth briefly addressing possible objections to the method and the conclusions: it may be thought that the method of extraction determines the result such that, if no SEs had been allowed, none would have been found, and if more had been allowed, more would have been found. While in general extraction parameters greatly influence resulting sequences, a very restrictive definition was deliberately used because such a definition had the best prospects of success in delineating FSs as an empirically well-bounded phenomenon. The understanding of constructions (including FSs) as semantic units necessitates the appearance of SEs in certain FSs (cf. Table 2) even if no elements such as NUM or

MONTH had been allowed. Similarly, through the tight restrictions outlined and checks for plausibility in each instance, it was made certain that optional SEs are not extraction artefacts. Consequently, the results are likely to be conservative with regard to the true degree of schematicity. Notably, the goal was not to ascertain whether SEs would be found in FSs (this was clear from the outset), but to assess the degree to which FSs featured SEs.

Conclusions on the place of FSs among constructions

If FSs are placed within a constructionist approach to grammar, they slot in as predominantly lexically substantive constructions, their substantivity being the only formal distinguishing feature vis-à-vis other constructions. The careful analysis of a large sample of empirical data has now comprehensively shown that this distinguishing feature does not produce an empirically well-bounded phenomenon. Even if one allowed for a prototypical core with a fuzzy periphery, the periphery would cover most of the area. So formulaic sequences dissolve into the ocean of all constructions, although, as was shown, quantitatively they are much more than a drop with more than half of running words being part of FSs. From a constructionist perspective at least, FSs are therefore best seen as a convenient label for a theoretically defined portion of a larger phenomenon rather than a self-contained phenomenon – different in degree, rather than type. To construction grammarians, this is not a surprise – the fluidity of levels of schematicity and their ubiquitous intermixing in single constructions is what constructions are all about and FSs, where acknowledged, were simply taken to be part of this universe. To phraseologists and others interested in FSs, the magnitude of integration and lack of empirical boundaries around their object of focus will be less expected.

But the tables turn when theoretical considerations are brought into play. The four areas surveyed above (language processing, first language acquisition, second language acquisition and communication) indicate that FSs bear theoretical significance that marks them out as different from other constructions. This in turn results in particular and distinct research questions being asked of FSs compared to other constructions and imbues them with special significance notwithstanding their somewhat artificial delineation as a phenomenon. From this point of view, FSs are different and need special questions asked of them. Furthermore, as outlined, these questions are important for a full understanding of the workings of language. None of this is news to those researching FSs, but with few exceptions, construction grammatical work has taken comparatively little interest in the particularities and implications of substantiveness, and therefore appears less than fully aware of how the degree of substantiveness affects constructions and their properties in theoretically relevant and distinctive ways.

The findings presented on the empirical boundedness of FSs are consequently not as irreconcilably opposed to existing theoretical claims about FSs as they might at first appear. The degree of substantivity of a construction, rather than the (impossible) categorical classification of a group of constructions as predominantly substantive, may well be responsible for the theoretically expressed effects of FSs, and none of the relevant claims specifically demand clear-cut boundaries – they readily allow, if not suggest, gradedness. Seen in this way, FSs are best understood as simultaneously part of the ocean of constructions and also something more significant.

This realisation, supported by the evidence supplied by this study, moves the field forward by enabling future research into FSs and other constructions to consider a number of points:

first, studies that span over related constructions at various levels of schematicity, including the more substantive levels, can draw more confidently on the essential unity of all constructions and the special contributions of substantivity to gain fuller insight into their subject. Although such studies are still very few and far between, Lionel Wee and Ying Ying Tan (2008) and Philip Durrant and Julie Mathews-Aydınlı (2011) demonstrate their potential. Second, work focusing on FSs, in acknowledging the gradualness within the set of all constructions as shown by this study, is able to tap into a theoretical framework that offers a theory of language at large, rather than theories of a corner of language, as fascinating as that corner may be. With this come opportunities to discover further-reaching implications, and to link findings to other areas of language and cognition. Third, the development of constructionist theory, by exploring the contribution of substantivity and of constructions with a high degree of substantivity in more detail, is offered the opportunity to rid itself of a blind spot and strengthen its explanatory power not only in the areas outlined above, but in many others too. Some work in this vein has already started in usage-based construction grammar (cf. Bybee, 2010: ch. 3) and will benefit from being taken up by more researchers. Fourth, if evidence presented is able to play a part in encouraging a closer coming together or indeed an integration of work in historically distinct traditions that nevertheless work on the same overall phenomenon, future work on FSs and other constructions is set to benefit greatly from the synergies so produced.

Finally, this study has looked at FSs and constructions from a point of view that recognises their nature as conventionalised pairings of form and meaning which I have argued does justice to both. While results have shown that, on these terms, an integration is not only possible but desirable in many respects, there remain other ways of looking at FSs and constructions, as we saw near the beginning: for example, in strictly formal terms (e.g. as form chunks or fragments occurring with a certain frequency in a corpus) or from a psycholinguistic point of view. While taking form as the primary or exclusive characteristic of linguistic structure is today advocated by few, it would be fascinating to investigate how FSs and constructions relate in terms of psycholinguistic representation and processing. At this current stage, it appears that not enough is known to facilitate such an undertaking, particularly in relation to the psycholinguistic properties of less than fully substantive constructions and the impact of differing levels of schematicity. Given the connections and the blurring between FSs and other constructions shown in this study, and the suggestion that differences in processing could be gradual as well, it would however seem surprising if the overall outcome of a psycholinguistically focused look at FSs and other constructions were to conclude that they are entirely separate phenomena.

Notes

1. https://books.google.com/ngrams/; query term: 'at the * expense of' on 27 April 2015; arguably there are two distinct patterns present – some of the modifications (like *small*) prefer the right-edge slot (marked *X*) to be filled with an amount (e.g. *at the small expense of £500*), while others (like *common*) are only possible when the right-edge slot denotes the entity who meets the expense (e.g. *at the common expense of all of the owners*).
2. Alternatively, relations between constructions at various levels of schematicity are frequently described as *inheritance* networks in which more substantive constructions inherit features of more schematic ones (e.g. Goldberg, 1995: 67).

Supplementary material

The supplementary material for this article is available online at http://dx.doi.org/10.1080/13825577.2015.1136158

Acknowledgements

I would like to thank Alison Wray for her helpful comments on an earlier draft of this article and Hannah Frank and Michael Willett for help with the rating of word sequences.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

The work was supported by the Swiss National Science Foundation under Grant P2BSP1_148623.

References

Allerton, David J (1984). 'Three (or Four) Levels of Word Cooccurrence Restriction.' *Lingua* 63.1: 17–40.

Altenberg, Bengt (1998). 'On the Phraseology of Spoken English.' *Phraseology: Theory, Analysis and Applications*. Ed. Anthony Cowie. Oxford: Clarendon. 101–122.

Altenberg, Bengt and Eeg-Olofsson, Mats (1990). 'Phraseology in Spoken English: Presentation of a Project.' *Theory and Practice in Corpus Linguistics*. Eds Jan Aarts and Willem Meijs. Amsterdam: Rodopi. 1–26.

anon (2001). *Word Lists*. 30 April 2015 <http://wortschatz.uni-leipzig.de/html/wliste.html>.

Arnon, Inbal and Snider, Neal (2010). 'More than Words: Frequency Effects for Multi-Word Phrases.' *Journal of Memory and Language* 62.1: 67–82.

Attardi, Giuseppe and Fuschetto, Antonio (2012). *WikiExtractor 2.2* [software]. 30 April 2015 <http://medialab.di.unipi.it/wiki/Wikipedia_Extractor>.

Bally, Charles (1909). *Traité de Stylistique Française, premier volume*. [*Treatise on the Stylistics of French, first volume*] Paris: Klincksieck.

Becker, Joseph (1975). 'The Phrasal Lexicon.' *Theoretical Issues in Natural Language Processing.* Eds BL Nash-Webber and Roger Schank. Cambridge, Mass: Bolt, Beranek & Newman. 60–63.

Boers, Frank, et al. (2006). 'Formulaic Sequences and Perceived Oral Proficiency.' *Language Teaching Research* 10.3: 245–261.

Brunner, Annelen and Steyer, Kathrin (2007). 'Corpus-Driven Study of Multi-Word Expressions Based on Collocations from a Very Large Corpus.' *Proceedings of CL2007, University of Birmingham, UK, 27-30 July 2007*. 29 March 2010 <http://corpus.bham.ac.uk/corplingproceedings07/paper/182_Paper.pdf>.

Buerki, Andreas (2011). *SubString* [computer software]. 17 March 2012 <http://buerki.github.com/SubString/>.

Buerki, Andreas (2012). 'Korpusgeleitete Extraktion von Mehrwortsequenzen aus (diachronen) Korpora.' ['Corpus-Led Extraction of Multi-Word Sequences out of (Diachronic) Corpora']. *Aspekte der historischen*

Phraseologie und Phraseographie. [*Aspects of Historical Phraseology and Phraseography*]. Eds Natalia Filatkina Anne Kleine-Engel Marcel Dräger and Harald Burger. Heidelberg: Universitätsverlag Winter. 263–292.

Buerki, Andreas (2013). *N-Gram Processor 0.4* [computer software]. 19 May 2015 <http://buerki.github.io/ngramprocessor/>.

Burger, Harald, Häcki Buhofer, Annelies and Sialm, Ambros (1982). *Handbuch der Phraseologie. [Handbook of Phraseology]*. Berlin: de Gruyter.

Bybee, Joan (2006). 'From Usage to Grammar: The Mind's Response to Repetition.' *Language* 82.4: 711–733.

Bybee, Joan (2010). *Language, Usage and Cognition*. Cambridge: Cambridge UP.

Chomsky, Noam (1957). *Syntactic Structures*. The Hague: Mouton.

Columbus, Georgie (2013). 'In Support of Multiword Unit Classifications.' *Yearbook of Phraseology*. 4. 23–44.

Coulmas, Florian (1979). 'On the Sociolinguistic Relevance of Routine Formulae.' *Journal of Pragmatics* 3. 239–266.

Coulmas, Florian (1981). 'Introduction: Conversational routine.' *Conversational Routine*. Ed Florian Coulmas. The Hague: Mouton. 1–18.

Dąbrowska, Ewa (2014). 'Recycling Utterances: A Speaker's Guide to Sentence Processing.' *Cognitive Linguistics* 25. 4: 617–654.

Dąbrowska, Ewa, and Lieven, Elena (2005). 'Towards a Lexically Specific Grammar of Children's Question Constructions.' *Cognitive Linguistics* 16. 3: 437–474.

Dominey, Peter (2006). 'From Holophrases to Abstract Grammatical Constructions: Insights from Simulation Studies.' *Constructions in Acquisition*. Eds Eve Clark and Barbara P Kelly. Stanford: CSLI. 137–62.

Durrant, Philip, and Doherty, Alice (2010). 'Are High-frequency Collocations Psychologically Real?' *Corpus Linguistics and Linguistic Theory* 6.2: 125–155.

Durrant, Philip, and Mathews-Aydınlı, Julie (2011). 'A Function-First Approach to Identifying Formulaic Language in Academic Writing.' *English for Specific Purposes* 30.1: 58–72.

Dutton, Kelly (2009). *Exploring the Boundaries of Fomulaic Sequences: A Corpus-Based Study of Lexical Substitution and Insertion in Contemporary British English*. Saarbrücken: VDM.

Ellis, Nick, Frey, Eric and Jalkanen, Isaac (2009). 'The Psycholinguistic Reality of Collocation and Semantic Prosody: Lexical Access.' *Exploring the Lexis-Grammar Interface.* Eds Ute Römer and Rainer Schulze. Amsterdam: Benjamins. 89–114.

Erman, Britt (2007). 'Cognitive Processes as Evidence of the Idiom Principle.' *International Journal of Corpus Linguistics* 12.1: 25–53.

Erman, Britt, and Warren, Beatrice (2000). 'The Idiom Principle and the Open Choice Principle.' *Text* 20.1: 29–62.

Feilke, Helmuth (1994). *Common sense-Kompetenz: Überlegungen zu einer Theorie des "sympathischen" und "natürlichen" Meinens und Verstehens*. Frankfurt am Main: Suhrkamp.

Feilke, Helmuth (2003). 'Textroutine, Textsemantik und sprachliches Wissen.' ['Text routine, text semantics and linguistic knowledge.'] *Sprache und mehr. [Language and More]*. Eds Angelika Linke, Hanspeter Ortner and Paul Portmann-Tselikas. Tübingen: Niemeyer. 209–230.

Fillmore, Charles, Kay, Paul, and O'Connor, Mary C (1988). 'Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone.' *Language* 64.3: 501–538.

Fleischer, Wolfgang (1982). *Phraseologie der deutschen Gegenwartssprache. [The Phraseology of Contemporary German]*. Leipzig: Bibliographisches Institut.

Goldberg, Adele (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: U of Chicago P.

Goldberg, Adele (2003). 'Constructions: A New Theoretical Approach to Language.' *Trends in Cognitive Sciences* 7.5: 219–224.

Goldberg, Adele (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford UP.

Granger, Sylviane, and Paquot, Magali (2008). 'Disentangling the Phraseological Web.' *Phraseology: An Interdisciplinary Perspective*. Eds Sylviane Granger and Fanny Meunier. Amsterdam: Benjamins. 27–49.

Hausmann, Franz Josef (1991). 'Collocations in Monolingual and Bilingual English Dictionaries.' *Languages in Contact and Contrast: Essays in Contact Linguistics*. Eds Vladimir Ivir and Damir Kalogjera. Berlin: de Gruyter. 225–236.

Hilpert, Martin (2014). *Construction Grammar and its Application to English*. Cambridge: Cambridge UP.

Hoffmann, Sebastian, Fischer-Starcke, Bettina and Sand, Andrea (2013). 'Introduction.' *International Journal of Corpus Linguistics* 18.1: 1–6.

Hoffmann, Thomas and Trousdale, Graeme (2013). 'Construction Grammar: Introduction.' *The Oxford Handbook of Construction Grammar*. Eds Thomas Hoffmann and Graeme Trousdale. Oxford: Oxford UP. 1–13.

Howarth, Peter (1998). 'Phraseology and Second Language Proficiency.' *Applied Linguistics* 19.1: 24–44.

Jespersen, Jens Otto Harry (1904). *How to Teach a Foreign Language*. London: Allen & Unwin.

Jones, Susan, and Sinclair, John (1974). 'English Lexical Collocations: A Study in Computational Linguistics.' *Cahiers de lexicologie* 24.1: 15–61.

Kuiper, Konrad (2007). 'Cathy Wilcox Meets the Phrasal Lexicon.' *Lexical Creativity, Texts and Contexts*. Ed Judith Munat. Amsterdam: Benjamins. 93–112

Langacker, Ronald (2008). 'Cognitive Grammar as a Basis for Language Instruction.' *Handbook of Cognitive Linguistics and Second Language Acquisition*. Eds Peter Robinson and Nick Ellis. Abingdon: Routledge. 66–88.

Langlotz, Andreas (2006). *Idiomatic Creativity: A Cognitive-Linguistic Model of Idiom-Representation and Idiom-Variation in English*. Amsterdam: Benjamins.

Lieven, Elena and Brandt, Silke (2011). 'The Constructivist Approach.' *Infancia y Aprendizaje* 34.3: 281–296.

Manning, Christopher and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Martinez, Ron and Schmitt, Norbert (2012). 'A Phrasal Expressions List.' *Applied Linguistics* 33.3: 299–320.

Mieder, Wolfgang (2004). *Proverbs: A Handbook*. Westport, Conn: Greenwood.

Mittmann, Brigitta (2004). *Mehrwort-Cluster in der englischen Alltagskonversation*. Tübingen: Narr.

Moon, Rosamund (1995). 'Introduction.' *Collins COBUILD Dictionary of Idioms*. Ed. John Sinclair. London: HarperCollins. iv-vii.

Moon, Rosamund (1998). 'Frequencies and Forms of Phrasal Lexemes in English.' *Phraseology: Theory, Analysis and Applications*. Ed. Anthony Cowie. Oxford: Clarendon. 79–100.

Myles, Florence (2004). 'From Data to Theory: The Over-Representation of Linguistic Knowledge in SLA.' *Transactions of the Philosophical Society* 102.2: 139–168.

Nattinger, James and DeCarrico, Jeanette (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford UP.

Pawley, Andrew (2001). 'Phraseology, Linguistics and the Dictionary.' *International Journal of Lexicography* 14.2: 122–134.

Pawley, Andrew and Syder, Frances (1983). 'Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency.' *Language and Communication*. Eds Jack Richards and Richard Schmidt Harlow: Longman. 191–226.

Renouf, Antoniette and Sinclair, John (1991). 'Collocational Frameworks in English.' *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Eds Karin Aijmer and Bengt Altenberg. Harlow: Longman. 128–143.

Sag, Ivan et al. (2002). 'Computational Linguistics and Intelligent Text Processing.' *Multiword Expressions: A Pain in the Neck for NLP*. Ed. Alexander Gelbukh. Berlin: Springer. 189–206.

de Saussure, Ferdinand (1974) [1916]. *Course in General Linguistics*, , trans. Wade Baskin. London: Peter Owen.

Seidlhofer, Barbara (2009). 'Accommodation and the Idiom Principle in English as a Lingua Franca.' *Intercultural Pragmatics* 6.2: 195–215.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford UP.

Sorhus, Helen (1977). 'To Hear Ourselves – Implications for Teaching English as a Second Language.' *English Language Teaching Journal* XXXI.3: 211–221.

Tomasello, Michael (2005). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard UP.

Tremblay, Antoine, et al. (2011). 'Processing Advantages of Lexical Bundles.' *Language Learning* 61.2: 569–613.

Villavicencio, Aline et al. (2004). 'Lexical Encoding of MWEs.' *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*: 80–87. 28 September 2012 <http://dl.acm.org/citation.cfm?id=1613197>.

Wee, Lionel and Tan, Ying Ying (2008). 'That's So Last Year! Constructions in a Socio-Cultural Context.' *Journal of Pragmatics* 40.12: 2100–2113.

Wray, Alison (2002a). *Formulaic Language and the Lexicon*. Cambridge: Cambridge UP.

Wray, Alison (2002b). 'Formulaic Language in Computer-Supported Communication: Theory Meets Reality.' *Language Awareness* 11.2: 114–131.

Wray, Alison (2006). 'Formulaic Language.' *Encyclopedia of Language and Linguistics*. Ed Keith Brown. Boston: Elsevier. 590–597.

Wray, Alison (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford UP.

Wray, Alison (2009). 'Identifying Formulaic Language: Persistent Challenges and New Opportunities.' *Formulaic Language: Distribution and Historical Change*. Eds Roberta Corrigan, Edith Moravcsik, Hamid Ouali and Kathleen Wheatley. Amsterdam: Benjamins. 27–52.

Wray, Alison and Perkins, Michael (2000). 'The Functions of Formulaic Language: An Integrated Model.' *Language and Communication* 20.1: 1–28.

Wulff, Stefanie (2013). 'Words and Idioms.' *Oxford Handbook of Construction Grammar*. Eds Thomas Hoffmann and Graeme Trousdale. Oxford: Oxford UP. 274–289.

Semantic Unity Rating Guidelines

You will be asked to decide whether a given word sequence forms a semantic unit or not. Here is an explanation of what is meant by semantic unit:

The type of unit of meaning which is characteristic of words or phrases

Examples:

(A) semantic unit:
at home / fact to face / even though / for example / of course / went to school

(B) **no** semantic unit:
know that it / of small / more to / with the smallest / did several

Something is missing from sequences in (B). Meaning is present in parts, but forms no unit. Here are a few more examples of semantic units:

(C) in this way

(D) if only

(E) such that

The last two are semantic units despite the fact that they are not complete structural units, so structural unity is no definite guide.

Grammatical case, tense, mood or voice should not be decisive in rating. All of (F) to (J) are semantic units, even though this seems clearer in (F) and (H) than (G) and (I). Note that a subject is not always necessary for the semantic unit to be complete even if there is a verb, as (H) and (I) show. (J) is a semantic unit because it parallels the semantic unity of the verb 'do' – it is simply a different form of the verb 'do'.

(F) the central committee (H) injure oneself (J) would have done

(G) of the central committee (I) injured himself

There is one additional point to consider: word sequences must be rated as semantic units if, by adding a single element to either their right or left edge, they can be made into semantic units. In the examples below, this element is marked with an 'X' (though in the examples you will be asked to rate, no Xs will be shown):

(K) a choice of X (N) fear of X

(L) the disagreement with X (O) X years old

(M) his former X (P) to worry that X

Note that X might be a single word or a unit larger than a word as in (P). The following restrictions on the use of Xs apply:

1) An X must only be added if we have a reasonable idea of what sort of thing X typically is. For example, in (L), X is restricted structurally to a noun phrase and semantically typically to a person with whom a difference in opinion exists. If it is completely open or unclear, such as in the sequences in (Q) or (R), the sequence is **not** a semantic unit. When deciding this, it is sometimes useful to think about where the main meaning, or the weight of meaning lies. If too much of it lies or depends on the X, and we have very little meaning without knowing X, we do **not** have a semantic unit.

(Q) a blue X

(R) which is used X

In (Q), X could be almost anything at all, including 'and bright yellow shirt' or 'collar worker' or any noun at all. Compare this with the structurally similar (M), where we nevertheless know that X is typically a

person (or organization) with which the 'he' used to have a relationship of sorts (e.g. his former boss/company/colleague/mate/residence, etc.). This is why (M) is a semantic unit, but (Q) is **not**.

2) Only a single X is allowed per word sequence. If more than one X is necessary to establish a semantic unit, such a sequence is **not** to be rated a semantic unit.

3) If the X is a particular word (rather than a **type** of word or other unit), it is **not** classed as a semantic unit. An example is given in (S), where the X has to be the particular word 'later' rather than a type of word or other unit as in (K) to (P).

(S) sooner or X

In general, rating word sequences for semantic unity requires the use your linguistic intuition. There are many clear cases, but there may also be some cases that are not clear cut and there may not be a correct answer in all instances: it is a judgment call. Do not think for too long when making decisions – no longer than about two seconds. There is no need to consult dictionaries or other resources, the decision should be made on the basis of the sequences as presented.

Finally, the following needs to be considered:

1) The word sequences do not contain commas – do not let this distract you.
   i.e. no not at all → no, not at all

2) NUM always stands for a number (cardinal or ordinal)
   i.e. NUM cups of tea → five cups of tea / NUM game → first game

3) MONTH stands for any month
   i.e. NUM MONTH NUM → 21 April 1974