Disentangling the phraseological web

Sylviane Granger & Magali Paquot

[DRAFT]

Abstract
Although phraseology has recently begun to establish itself as a field in its own right, this process is being hindered by two main factors: the highly variable and wide-ranging scope of the field and the vast and confusing terminology associated with it. This chapter tackles these two issues successively in an attempt to disentangle the 'phraseological web'. We first draw up a clear distinction between two major approaches to the study of multi-word units, i.e. the phraseological approach and the distributional or frequency-based approach, which have set quite different boundaries to the field. We then argue that the variations in scope that characterize the field of phraseology are a direct result of its fuzzy borders with four neighbouring disciplines: semantics, morphology, syntax and discourse. We describe some of the most influential typologies of word combinations within the phraseological approach and present a categorization of multi-word units emerging from the distributional approach. We conclude with suggestions regarding both the scope of the field and the terminology used.

1. Introduction

Although the study of multi-word units has a long history, with Bally distinguishing between the fully fixed 'unités phraséologiques' and the looser 'séries phraséologiques' as early as 1909, phraseology has only recently begun to establish itself as a field in its own right. This process is being hindered by two main factors however: the highly variable and wide-ranging scope of the field on the one hand and on the other, the vast and confusing terminology associated with it.

Phraseology can be loosely defined as "the study of the structure, meaning and use of word combinations" (Cowie 1994: 3168). As word combinations come in many different shapes and forms, the scope of the field is a function of the criteria used by linguists to distinguish phraseological units from non-phraseological ones. While the East European tradition (see Section 2) has tended to favour fairly fixed combinations like idioms or proverbs, the more recent corpus-based approaches have adopted a much wider perspective and included many word combinations that would traditionally be considered to fall outside the scope of phraseology. Even neighbouring countries like Great-Britain and France have quite different traditions: the notion of fixedness ('figement') lies at the heart of the French tradition (Gross 1996), while the Anglo-Saxon tradition has from the start attached great importance to the less fixed category of collocation (Palmer 1933 in Cowie 1998a: 210-212). This diversity is a source of richness but it also hinders communication between linguists and generally increases the impression of fuzziness in the field. This impression is amplified by the unwieldy terminology employed, with different terms covering the same units and the same terms used to denote quite different

units. This situation is deplored by Cowie (1998a: 210) who refers to phraseology as "a field bedevilled by the proliferation of terms and by the conflicting uses of the same term". As pointed out by Wray & Perkins (2000: 3), one of the pernicious effects of the loose terminology is that it makes it "extremely difficult to be sure when like is being compared with like".

The purpose of this chapter is to give a brief overview of the field of phraseology, introduce some major typologies and make suggestions regarding both the scope of the field and the terminology used. In Section 2 we briefly present the two major approaches to phraseology. Section 3 argues that the variations in scope that characterize the field of phraseology are a direct result of its fuzzy borders with four neighbouring disciplines: semantics, morphology, syntax and discourse. In Section 4 some influential typologies are presented and in Section 5 we make suggestions as to how current (perceptions of) confusion in the field might be addressed. Section 6 offers some conclusions and suggestions for new ways forward.

2. Two major approaches to phraseology

The traditional approach to phraseology is greatly indebted to scholars from the former Soviet Union and other countries of Eastern Europe (cf. Cowie 1998b: 1). Russian scholars like Vinogradov and Amosova are at the foundations of a view of phraseology that restricts the scope of the field to a specific subset of linguistically defined multi-word units and sees phraseology as a continuum along which word combinations are situated, with the most opaque and fixed ones at one end and the most transparent and variable ones at the other. Cowie's (1981) continuum, which goes from free combinations to pure idioms through restricted collocations and figurative idioms, is a direct descendent of these early Russian schemes (see Section 4.1). One of the main preoccupations of linguists working within that tradition has been to find linguistic criteria for distinguishing one type of phraseological unit from another and especially for distinguishing the most variable and transparent multi-word units from free combinations, which have only syntactic and semantic restrictions and are therefore considered as falling outside the realm of phraseology (Cowie 1998b: 6). In this tradition, the most idiomatic units, whose meanings cannot be derived from the meanings of the constituents, are often presented as the most 'core'. This appears clearly from the following statement by Gläser (1998: 126): "Idioms form the majority and may be regarded as the prototype of the phraseological unit". This tradition deserves much of the credit for having established phraseology as a discipline in its own right, created a terminology for the field and provided linguists with a set of discrete criteria which can be used to categorize and analyze phraseological units. To refer to this tradition, we adopt Nesselhauf's (2004) term 'phraseological approach'.

A more recent approach to phraseology, which originated with Sinclair's pioneering lexicographic work, has literally turned phraseology on its head. Instead of adopting a top-down approach which identifies phraseological units on the basis of linguistic criteria, it uses a bottom-up corpus-driven approach to identify lexical co-occurrences (Sinclair 1987). This inductive approach, which is referred to as the distributional (Evert 2004) or frequency-based (Nesselhauf 2004) approach, generates a wide range of word combinations, which do not all fit predefined linguistic categories. It

has opened up a "huge area of syntagmatic prospection" (Sinclair 2004: 19) encompassing sequences like frames, collocational frameworks, colligations and largely compositional recurrent phrases (see Section 4.2). All these sequences illustrate Sinclair's (1991) idiom principle, a principle that views language as essentially made up of strings of co-selected words that constitute single choices. This new approach has "pushed the boundary that roughly demarcates the 'phraseological' more and more into the zone previously thought of as free" (Cowie 1998b: 20). Many of the units that were traditionally considered as peripheral or falling outside the limits of phraseology have now become central as they have revealed themselves to be pervasive in language, while many of the most restricted units (idioms, proverbs) have proved to be highly infrequent (Moon 1998). Unlike proponents of the classical approach to phraseology, Sinclair and his followers are much less preoccupied with distinguishing between different linguistic categories and subcategories of word combinations or more generally setting clear boundaries to phraseology. In Sinclair's model of language, phraseology is central: phraseological items, whatever their nature, take precedence over single words. This radical view has been criticized. Gaatone (1997: 168), for instance, welcomes the growing importance attached to multi-word units but warns against considering everything as phraseological.

3. The fuzzy borders of phraseology

The two approaches to phraseology set quite different boundaries to the field. In Figure 1 phraseology is represented as a field that has fuzzy borders – hence the dotted lines – with four other major fields: semantics, morphology, syntax and discourse.[1] The territory covered by the frequency-based approach (represented by the light grey circle) is much wider than that of the traditional view (in dark grey). Figure 1 highlights the inherently multidisciplinary nature of the field. As pointed out by Mel'čuk (1995: 227), phraseology has to deal with everything, which makes it "so difficult, but so appealing!"

Insert Figure 1 around here

3.1 Phraseology and semantics

The field with which phraseology has arguably the strongest - and at the same time fuzziest - links is semantics. Recourse to semantics is essential to distinguish between different types of lexical affinity. Allerton (1984) sets out semantic co-occurrence restrictions, which can be logically predicted from the lexical meaning and semantic traits of a given lexeme (e.g. in its literal meaning the adjective *pregnant* can only be used to describe female beings), from locutional co-occurrence restrictions, which cannot be generalized and should be described for every single lexeme. Only usage can explain why we say *strong coffee* and not **powerful coffee* or why we prefer to speak of *a chestnut horse* rather than of *a brown horse*. This distinction lies at the heart of the traditional approach. It makes a sharp distinction between free combinations like *spend a day/year* or *spend money/two pounds*, which are only governed by semantic co-occurrence restrictions and are thus considered as falling outside the realm of phraseology, and other

multi-word units whose co-occurrence cannot be accounted for by semantics and qualify as phraseological units or phrasemes.

Another semantic notion that lies at the heart of phraseology is non-compositionality. A lexical item is said to be non-compositional if its global meaning is different from the sum of its individual parts (for a thorough discussion, cf. Svensson this volume). Non-compositionality is considered by some linguists as the defining criterion of phraseological units while others view it as a secondary feature that characterizes some, but definitely not all units. Mel'čuk (1998: 24) clearly holds the former view: "the main substantive property of a phraseme is its non-compositionality". However, all linguists, including Mel'čuk, recognize that non-compositionality is a cline, ranging from fully compositional to fully non-compositional with several intermediate categories: "there is no clear dividing-line between idioms and non-idioms: they form the end-points of a continuum" (Cowie et al. 1983: xiv). This notwithstanding, full compositionality is sometimes used as a factor of exclusion from phraseology in the traditional approach. For example, Cowie (2005) includes *face pack* in phraseology but excludes *face flannel* on the grounds that in *face flannel* the meaning is entirely compositional.

By contrast, all types of word combinations are part and parcel of the distributional approach which does not use semantic criteria to identify multi-word units. This is not to say that meaning plays no part. Rather it is a different view of meaning that prevails, the Firthian contextual theory of meaning, according to which "the formalisation of contextual patterning of a given word or expression is assumed to be relevant to the identification of the meaning of that word or expression" (Tognini-Bonelli 2001: 4). In this framework meaning extends well beyond the limits of the word. The relationship between a lexical item and a lexical set of semantically related words is what Sinclair (1996, 1998) and Partington (2004) refer to as semantic preference. For example, Partington (2004:148) observes that collocates of the maximizers *utterly, totally, completely* and *entirely* share the semantic preference of 'absence/change of state', e.g. *totally uneducated* and *completely lacking*. The "proximity of a consistent series of collocates" (Louw 2000: 57) may establish yet another form of meaning, i.e. semantic prosody, whose primary function is "the expression of the attitude of its speaker or writer towards some pragmatic situation" (ibid.) (see also Louw 1993). Partington (2004: 150-151) illustrates the interdependence of semantic preference and semantic prosody using the example of the verb *undergo*, which collocates with, and thus shows semantic preference for, items from the lexical sets of 'change' (e.g. *dramatic changes, a historic transformation*), 'medicine' (e.g. *treatment, brain surgery*), 'testing' (e.g. *examinations*) and 'involuntariness' (e.g. *must, forced to, required to*). All these semantic preferences imbue the item *undergo* with a very strong unfavourable semantic prosody. Other often cited examples of words with negative semantic prosody include *happen*, *set in* (cf. Sinclair 1991) and *cause* (cf. Stubbs 1995). The systematicity of these relationships between a word and its environment has led Sinclair and his colleagues to postulate the existence of an extended unit of meaning "where collocational and colligational patterning (lexical and grammatical choices respectively) are intertwined to build up a multi-word unit with a specific semantic preference, associating the formal patterning with a semantic field, and an identifiable semantic prosody, performing an attitudinal and pragmatic function in the discourse" (Tognini-Bonelli 2002: 79).

3.2 Phraseology and morphology

The definition of phraseology as the study of word combinations entails that phraseological units are made up of at least two words. Polylexicality is generally described as one of the first necessary conditions for inclusion in the phraseological spectrum (cf. Gross 1996; Mejri 2005; Montoro del Arco 2006).[2] However, in view of the ambiguity surrounding the definition of the concept of word, this definition is not as helpful as may seem at first sight. The scope of phraseology varies according to whether 'word' is taken in the meaning of orthographic word (separated by blanks on either side) or in the meaning of "unit characterized by internal stability and uninterruptability" (Lyons 1968: 202). In the first meaning, *of course* or *letter box* are considered to be made up of two words and hence part of phraseology, in the second they are monolexemic and hence fall outside its scope. The situation is complicated by the fact that compounds can be written in three different ways (solid as in *bookstore*, hyphenated as in *father-in-law* and open as in *high school*) and regularly have more than one spelling form (*good will, good-will, goodwill*). Although crucial, this issue is rarely tackled explicitly in phraseological studies. One regularly has to scan through the examples given by the authors to find out whether or not (solid, hyphenated and/or open) compounds are included in the range of phraseological units covered. The traditional view either excludes compounds from phraseology altogether (Barkema 1996: 133) or only keeps units that meet some well-defined criteria (stress, meaning, etc). Others seem to exclude compounds written as one word, viz. solid compounds, but include open and hyphenated compounds (e.g. Mel'čuk 1995; Gläser 1998). In the distributional approach, all sequences made up of two or more graphic words are extracted if they meet some recurrence or co-occurrence threshold. As a result, a wide range of phraseological units are extracted, including open compounds (and possibly hyphenated ones) but excluding solid compounds.

It is important to note that compounds are not the only category to pose problems because of their uncertain status as single or multi-word units. Equally problematic and less often referred to are the categories of complex prepositions (*due to*), adverbs (*in fact*) and conjunctions (*even if*), which are generally either totally disregarded or regarded as minor categories (see Section 4.1).[3] Linguists often make quite arbitrary decisions as to what they include and exclude. For instance, Moon (1998: 79) excludes "for practical reasons" compound nouns, adjectives, and verbs such as *civil servant, self-raising*, and *freeze-dry*, but includes units such as *at last* and *in fact,* which she calls 'grammatical collocations'. The reason she invokes to exclude compounds is that "[t]he interest in compound words seems to me to rest largely in morphology" (ibid.: 3). However, her study of fixed expressions and idioms turns out to include units such as *ivory tower*, *trump card* or *full stop*, which would normally be classified as compounds. The issue of compounding is thus undoubtedly a major factor in blurring the line between phraseology and morphology.

3.3 Phraseology and syntax

As phraseology and syntax both deal with syntagmatic relations, it is normal that a clear demarcation line between the two fields should be difficult to draw. The whole debate

centres on the looser, less idiomatic phraseological units, in particular collocations. While collocations are usually defined as arbitrarily restricted combinations of lexical words like *strong tea* or *dispel fear,* some linguists, like Benson et al. (1986), subdivide collocations into lexical collocations, which contain two lexical words, and grammatical collocations, which are made up of a lexical word and a grammatical word (*aim at, afraid that*) or structure (*avoid* + -ing, *necessary* + infinitive).[4] This view is shared by several linguists (cf. Gries this volume; Hunston 2002) but criticized by others who, like Heid (2002), consider that Benson's grammatical collocations in fact pertain to syntax. The fuzzy area that is at stake here is that of word grammar, i.e. the syntactic constraints on the use of lexis or, to use Woolard's (2000: 45) term, the "grammatical signatures" of words. This area is very close to that of valency patterns, which describe words in terms of the obligatory and optional arguments they accept (cf. Herbst et al. 2004). The whole debate should also be put in relation to Hoey's (2005) notion of 'lexical priming', which posits that words are primed to favour particular collocates, grammatical roles and positions, semantic associations, etc.

Another fuzzy area between phraseology and syntax is that between compounds and syntactic phrases. In principle, the situation is clear: productive and regular phrases belong to syntax and compounds originate in the lexicon. However, Giegerich (2004, 2005) demonstrates that some constructions, for example Adj + N constructions like *dental building, mental hospital* or *financial advisor*, straddle the lexicon-syntax divide. The general issue at play is that of syntactic flexibility, a feature regularly presented as a determinant of phraseological status and more particularly of idiom status. It involves determining the extent to which word combinations are allowed to undergo syntactic variation (e.g. passivisation, insertion, deletion, pronominalisation) without losing their phraseological status. Recent corpus-based studies, notably that of Moon (1998), have shown up "the fallacy of the notion of fixedness of form" (ibid.: 47) and degrees of inflexibility now tend to be considered as an indication rather than a criterion of phrasemes (Svensson 2002 and this volume).

Generally speaking, the traditional approach tends to adopt a stricter attitude, clearly distinguishing between phraseology and syntax, while the distributional approach gives more emphasis to the lexico-grammar interface. Stefanowitsch and Gries (2003), for example, have recently proposed applying collocational analysis within a constructional view of language, viz. collostructional analysis (a blend of 'construction' and 'collocational analysis'), with the aim of providing "an objective approach to identifying the meaning of a grammatical construction and of determining the degree to which particular slots in a grammatical structure prefer, or are restricted to, a particular set or semantic class of lexical items" (ibid.: 211). Collostructional analysis can be performed on single words in specific constructions (e.g. *cause* in transitive constructions) and what the authors call 'variable idioms' (e.g. the [X *think nothing of* V_{gerund}] construction), partially filled and unfilled argument structure constructions (e.g. the *into*-causative, the ditransitive), and tense, aspect and mood (e.g. lexemes attracted by the progressive form, the imperative or past tense).

3.4 Phraseology and discourse

Phraseology has close links with discourse, a field which centres on the organisation of language above the sentence or above the clause, and therefore studies larger linguistic units, such as conversational exchanges or written texts (Stubbs 1983: 1).

The traditional approach to phraseology has tended to favour units that reflect discourse as interaction. Cowie (1988) distinguishes a category of 'formulae', viz. pragmatically specialized units like *good morning* or *how do you do,* whose meanings "are largely a reflection of the way they function in discourse (as greetings, enquiries, invitations, etc.)" (ibid: 132). Another category of formulae is that of units such as *are you with me* or *would you mind repeating that,* which are used in "organizing turn-taking, indicating a speaker's attitude to other participants, and generally ensuring the smooth conduct of interaction" (ibid: 133). The same emphasis on interactional phrasemes is found in Mel'čuk (1998). These studies typically focus on the most fixed units typical of speech. It is revealing in this respect that Gramley & Pätzold's (1992: 58-61) section on what they call "pragmatic idioms" deals almost exclusively with spoken interaction.

With the distributional approach the focus moves from pragmatics to stylistics or rhetoric. While recognizing the importance of routine formulae, corpus analysts attribute equal – and in many cases even greater – importance to text structuring multi-word units. Using automatic methods Biber et al. (1999: 990ff) extract a wide range of prefabricated sequences, called 'lexical bundles', which they define as "simple sequences of word forms that commonly go together in natural discourse". These sequences, which mostly display syntactic and semantic regularity, fulfil a range of major discourse functions, such as hedging, organizing, etc. They are typically verbal and clausal units in speech (e.g. *I don't know, I think I might, what's the matter with*) and extended nominal phrases and prepositional phrases in writing (e.g. *the effect of, in the case of, the extent to which*). A similar study by Altenberg (1998) uncovers "a large stock of recurrent word-combinations that are seldom completely fixed but can be described as 'preferred' ways of saying things – more or less conventionalized building blocks that are used as convenient routines in language production. These building blocks come in all forms and sizes, from complete utterances to short snatches of words, and they display varying degrees of flexibility" (see also De Cock 2003 & 2004 for an analysis of recurrent word combinations in native and learner writing and speech). All these studies have highlighted the role of 'preferred ways of saying things' as key register markers and have contributed significantly to widening the scope of phraseology.

4. Categories of word combinations

As word combinations are highly heterogeneous, linguists have quite naturally felt the need to subcategorize them. Typologies abound in the literature: some are designed for lexicological or lexicographic purposes (Gläser 1986; Cowie 1988; Moon 1998), others are pedagogically-oriented (Nattinger & DeCarrico 1992; Lewis 1993) or take a psycholinguistic perspective (Wray & Perkins 2000; Wray 2002). Several 'ad hoc' descriptions have also been proposed within the field of natural language processing (cf. Sag et al. 2002; Tschichold 2000).

Differences between the typologies largely correspond to differences in the selection of the features used to categorize multi-word units and the prioritization of

selected features. Most classifications give prominence to one or more of five features of phrasemes: (1) internal structure (e.g. verb + noun or verb + preposition); (2) extent: phrase- vs. sentence-level; (3) degree of semantic (non-)compositionality; (4) degree of syntactic flexibility and collocability; (5) discourse function. Among the terms used to refer to subcategories of multi-word units, some appear to have acquired a relatively stable core meaning (e.g. idioms are usually defined as being non-compositional), others are much more confusing (e.g. collocations are used in a large number of different meanings).

The following sections are not intended as a comprehensive survey of the many typologies of phraseological units that have been proposed in different fields. Instead, Section 4.1 focuses on some influential traditional typologies which are deeply rooted in lexicology and lexicography. In Section 4.2, we propose a categorization of multi-word units emerging from the distributional approach.

4.1 Some influential typologies

One of the most influential typologies in English lexicology and lexicography is that of Cowie (e.g. 1988, 1994), which makes a primary distinction between composites, which function syntactically at or below the level of the sentence, and formulae, which function pragmatically as autonomous utterances.

Insert Figure 2 around here

Insert Figure 3 around here

As shown in Figure 2, composites are further subdivided into restricted collocations, figurative idioms and pure idioms, three categories which form a phraseological continuum, with the most transparent and variable at one hand and the most opaque and fixed at the other, as illustrated in Figure 3. The category of restricted collocations, often referred to simply as 'collocations', includes combinations such as *perform a task* or *heavy rain,* which are characterized by restricted collocability and figurative or specialized meaning of one of the elements. It includes verb-noun combinations with a delexical verb (e.g. *make a comment*). Figurative idioms have a figurative meaning but also preserve a literal interpretation (e.g. *do a U-turn*). They resist substitution of their components. Pure idioms such as *spill the beans* or *blow the gaff* are semantically non-compositional. The category of formulae includes 'sentence-like' units, "which function pragmatically as sayings, catchphrases, and conversational formulae" (Cowie 1998b: 4). Cowie (2001) later subdivides the category of formulae into routine formulae, like *good morning,* or *see you soon*, which perform speech-act functions, and speech formulae, which are used to organize messages and indicate speakers' or writers' attitudes (*you know what I mean, are you with me?*).

Another influential model is that proposed by Mel'čuk (1995, 1998) within the meaning-text theory. Although it uses a different terminology, it is very similar to Cowie's, notably in the primary distinction made between semantic phrasemes, which

roughly correspond to composites, and pragmatic phrasemes or pragmatemes, which are very close to Cowie's formulae.

Insert Figure 4 around here

One highly influential aspect of Mel'čuk's work is his treatment of collocations by means of lexical functions. When a native speaker of English wants to express the fact that somebody smokes a lot, he usually says that this person is a *heavy smoker* rather than a *big smoker*. By contrast, he will most probably speak of a *big eater* rather than a *heavy eater*. Mel'čuk (1995, 1998) attempts to describe these lexical preferences with lexical functions. A lexical function is "a very general and abstract meaning that can be expressed in a large variety of ways depending on the lexical unit to which this meaning applies" (Mel'čuk 1995: 186). Examples of lexical functions are:

- **Magn** which expresses the meaning of 'intense(ly)' or 'very' and functions as an intensifier, e.g. **Magn**(shave$_N$) = *close, clean*; **Magn**(easy) = *as pie, as 1-2-3*; **Magn**(to condemn) = *strongly*
- **Oper** which expresses the meaning of 'do/perform', e.g. **Oper$_1$**(cry) = *to let out* [ART~]
- **Real** which conveys the meaning of 'fulfil the requirement of X' or 'do with X what you are supposed to do with X', e.g. **Real$_1$**(car) = *to drive* [ART~]; **Real$_1$**(accusation) = *to prove* [ART~]

Unlike Cowie's and Mel'čuk's typologies, Burger's (1998) typology is primarily based on the function of phraseological units in discourse. As shown in Figure 5, the top subdivision distinguishes between the following three functional categories: referential units, communicative units and structural units. Referential phraseological units are divided into two sub-categories according to a syntactico-semantic criterion. First, nominative phraseological units are constituents of the sentence and refer to objects, phenomena or facts of life (e.g. *Schwarzes Brett* 'billboard' or *jemanden übers Ohr hauen* 'to rip somebody off'). This category broadly corresponds to Cowie's 'composites' and Gläser's (1998) 'nominations'. Following the Russian tradition and phraseologists such as Cowie and Mel'čuk, nominative phraseological units are sub-divided into idioms, partial idioms and collocations. Second, propositional phraseological units generally function at sentence level but a few propositional phraseological units function at text level; they refer to a statement or an utterance about these objects or phenomena (*Morgenstund hat Gold im Mund* 'the early bird catches the worm'). Propositional units include proverbs and idiomatic sentences, two broad categories that are classified as 'formulae' or 'pragmatic phrasemes' in models such as those put forward by Cowie and Mel'čuk that use both the criteria of function in discourse and function in the sentence. Communicative phraseological units or routine formulae fulfil an interactional function: they are typically used as text controllers to initiate, maintain and close a conversation or to signal the attitude of the addressor. Examples are *Guten Morgen* ('Good morning') and *Ich meine ...* ('Well, I mean…'). Unlike Cowie and Mel'čuk, Burger creates a third category of structural phraseological units which includes word combinations that establish grammatical relations, e.g. *in Bezug auf* ('concerning') and *sowohl ... als auch*

('as well … as …'). However, he regards structural phraseological units as the smallest and least interesting[5] category and does not go into any further detail.

Insert Figure 5 around here

4.2 Distributional categories

No categorization of phraseological units has emerged from studies rooted in the distributional approach to phraseology. It is, however, possible to draw up a typology of the types of units obtained by the different extraction procedures. As shown in Figure 6, a main subdivision can be made between two main extraction methods: n-gram analysis and co-occurrence analysis (cf. Stubbs 2002).

Insert Figure 6 here

N-gram analysis is a method which allows for the extraction of recurrent continuous sequences of two or more words, viz. "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al. 1999: 990). It has been used by a wide range of authors for a variety of purposes: terminology extraction, variation study, interlanguage study, information retrieval, etc. The extracted sequences are called n-grams (or the more specific terms bigrams or trigrams) (cf. Stubbs 2007a, 2007b), lexical bundles (Biber and Conrad 1999; Biber et al. 2003; Biber 2004), clusters (Scott and Tribble 2006), chains (Stubbs 2002; Stubbs and Barth 2003), recurrent sequences (De Cock 2003), recurrent word combinations (Altenberg 1998), etc. Examples of retrieved sequences are *I don't know what, I thought that was, can I have a, in the case of, on the other hand, the use of, the fact that.* A special category of recurrent sequences is that of collocational frameworks, which consist of sequences containing one or more free slots (Renouf & Sinclair 1991: 128).[6] Examples include '*a* + ? + of', '*an* + ? + *of*', '*be* + ? + *to*', and '*too* + ? + *to*'. Stubbs (2007a, 2007b) has recently referred to these multi-word sequences as 'phrase-frames'.

Co-occurrence analysis can be roughly defined as the statistical uncovering of significant word co-occurrences. To refer to the retrieved units, the term 'collocation' or 'collocate' are used (cf. Manning and Schütze 1999; Stubbs 2002). This type of analysis lay at the core of the COBUILD dictionary project, which relied on the following definition of collocates: "[t]he definition of regular or significant collocates was 'lexical items occurring within five words either way of the headword with a greater frequency than the law of averages would lead you to expect'. (...) Collocation was established only on the basis of corpus evidence" (Krishnamurthy 1987: 70). While collocation and collocates are the most widely used terms, some linguists (cf. Schmid 2003; Evert 2004) prefer to use the term 'co-occurrence' or 'co-occurrent' and, for reasons that will be made clear in Section 5, it is these terms which we have decided to include in our typology (cf. Figure 6).

These quantitative methods constitute fantastic heuristic devices that show their full potential in a program like the *Sketch Engine*, which provides lexicographers with "corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004: 105). Table 1 gives a sample of the word sketch for the noun *evidence* based on the British National Corpus (BNC).

Insert Table 1 around here

A word of caution is needed, however. In both types of analysis, the set of retrieved units depends on the settings adopted. N-gram analysis often relies on a relatively high frequency threshold. Biber et al. (2004: 376), for example, make use of a frequency cut-off of 40 times per million words to extract lexical bundles. A number of parameters may influence the outcome of a co-occurrence analysis. They include the size of the co-occurrence window or span used, the use of filters such as a minimum frequency threshold or a stopword list and, more importantly, the statistical measure used (e.g. mutual information, log-likelihood, t-test). For example, the association measure implemented in the *Sketch Engine* is the log-log. If other statistical measures such as the log-likelihood or mutual information (MI) were used, the word sketch for the noun *evidence* might look quite different (cf. McEnery et al. 2006: 208-226). The choice of an association measure clearly depends on the objectives of a co-occurrence analysis. As McEnery et al. (2006: 217) have suggested, word pairs that are significant when MI is used are generally interesting for lexicographical purposes while they are of secondary importance for pedagogical purposes. By contrast, they argue that word pairs highlighted by MI^3, a purely heuristic variant of MI, are probably "more useful for second language learners at beginning and intermediate levels." Other researchers have suggested that it is "difficult, if not impossible, to select one measure which provides the best assessment of the collocates" and that it is "probably better to use as much information as possible in exploring collocation, and to take advantage of the different perspectives provided by the use of more than one measure" (Barnbrook 1996: 101). Similarly, Bartsch (2004) uses three association measures to ensure identification of relevant co-occurrence data. She uses the MI score as the prime statistic for filtering what she calls 'collocation candidates' from the BNC word pairs and the t-test and chi-square scores for cross-checking purposes, as "these can support and sometimes supplement the data identified by MI" (ibid.: 112).

5. Reconciling the two approaches

The emergence of a new approach to phraseology is proving to be of immense value to the field. However, proponents of the two approaches are still too wide apart and both sides have a great deal to gain from a rapprochement. Many linguists working in the traditional framework seem to be largely unaware of the benefit they could derive from automatic corpus-based methods of extraction and analysis. Conversely, linguists working in the distributional framework seem not to appreciate how much they stand to benefit from the fine-grained linguistic analyses of the traditional approach. However, any rapprochement will only be fruitful if it is accompanied by some rigorous

clarification of the terminology. We suggest making a clear distinction between two typologies: one for automated extraction and one for linguistic analysis.

To refer to the results of automated extraction, we advocate the use of the terms in Figure 6. This means that in our view the term 'collocation' should not be used to refer to statistical word co-occurrences but instead kept in its traditional meaning of usage-based lexically restricted combination. We agree with Schmid (2003: 239) that "[i]t is not clear what is gained by calling co-occurrences of words 'collocations', when the term 'combination', or indeed 'co-occurrence' itself, covers the same range of phenomena."

As regards the linguistic classification, we think it is essential to integrate the new insights derived from the corpus-based approach. We propose an extended version of Burger's (1998) classification, as represented in Figure 7. Phraseological units are assigned to one of three major categories: referential phrasemes, textual phrasemes (an extension of Burger's category of 'structural phrasemes') and communicative phrasemes. Referential phrasemes are used to convey a content message: they refer to objects, phenomena or real-life facts. They include lexical and grammatical collocations, idioms, similes, irreversible bi- and trinomials, compounds and phrasal verbs. Textual phrasemes are typically used to structure and organize the content (i.e. referential information) of a text or any type of discourse; they include grammaticalized sequences such as complex prepositions and complex conjunctions, linking adverbials and textual sentence stems. Communicative phrasemes are used to express feelings or beliefs towards a propositional content or to explicitly address interlocutors, either to focus their attention, include them as discourse participants or influence them. They include speech act formulae, attitudinal formulae, commonplaces, proverbs and slogans.

Insert Figure 7 around here

Categories of multi-word units have received a wide range of definitions in the literature and a detailed survey would not be possible within the scope of this chapter. Tables 2-4 contain a set of working definitions, which draw heavily on the work of major phraseologists, notably Cowie, Mel'čuk and Burger. For in-depth discussion of each category, we refer the readers to publications by these authors and others in Cowie (1998), Allerton et al. (2004), Burger et al. (2007) and several articles in this volume.

Insert Table 2 around here
Insert Table 3 around here
Insert Table 4 around here

6. Conclusion

The major and rapid expansion of the field of phraseology in the last 25 years has resulted in the co-existence of two approaches – one linguistically-based, the other data-driven. While this development has undeniably further blurred the boundaries of a field whose inherent fuzziness has long been recognized, the resulting cross-fertilization

should be viewed as a unique opportunity to lead the field of phraseology into pastures new.

The unwieldy terminology used to refer to the different types of multi-word units is a direct reflection of the wide range of theoretical frameworks and fields in which phraseological studies are conducted and can be seen as a sign of the vitality of the field. To some extent, however, it impedes the process of cross-fertilization and hinders the smooth integration of phraseological insights into other fields, notably a field like language teaching where phraseology is taking on an increasingly important role. Some terminological order is clearly needed. However, we agree with Burger et al. (2007: 18) that "an international uniformity of terminology and classification is only possible and desirable to a certain degree." More than a unified terminology, what the field needs is some clear indication from researchers of the criteria used to identify multi-word units, as Gries (this volume) argues convincingly.

As regards the scope of the field, there can be little doubt that the new types of units uncovered by corpus-based approaches need to be fully incorporated into the mainstream of phraseology. Overemphasis on fixedness and semantic non-compositionality has tended to obscure the role played by a wide range of recurrent and co-occurrent units which are fully regular, both syntactically and semantically, and yet clearly belong to the field of phraseology. The crucial role played by these units in language is beginning to be recognized, as evidenced by a range of new publications (cf. Siepmann 2006; Gilquin et al. 2007; Pecman 2008). This remodelling of the field should go hand in hand with a better appropriation of the highly-developed analytical instruments provided by the traditional approach. Combining the best of the two worlds is the surest way of giving phraseology the place it deserves in linguistic theory and practice.

Acknowledgements

We gratefully acknowledge the support of the Communauté française de Belgique, which funded this research within the framework of the 'Action de recherche concertée' project entitled 'Foreign Language Learning: Phraseology and Discourse' (No. 03/08-301).

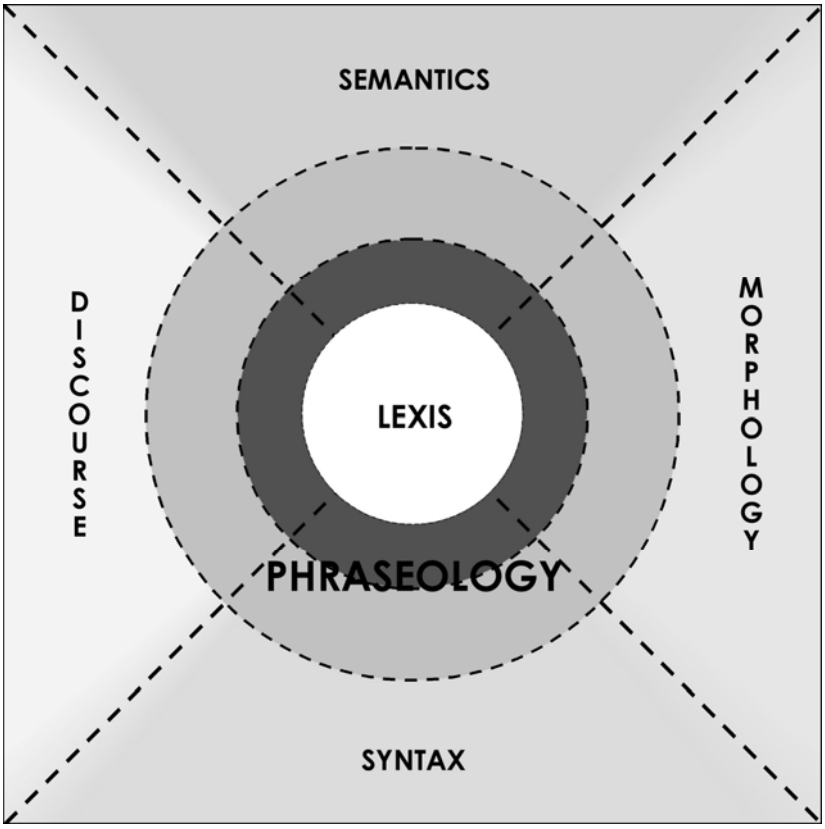

Figure 1: Phraseology wide and narrow

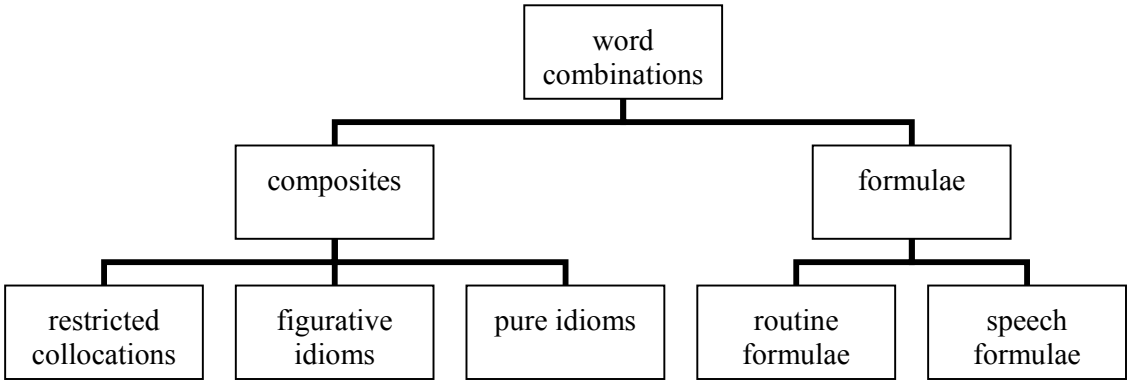

Figure 2: Cowie's (1988, 2001) classification of word combinations

free combination	restricted collocation	figurative idiom	pure idiom
blow a trumpet	*blow a fuse*	*blow your own trumpet*	*blow the gaff*

Figure 3: Cowie's (1981) phraseological continuum

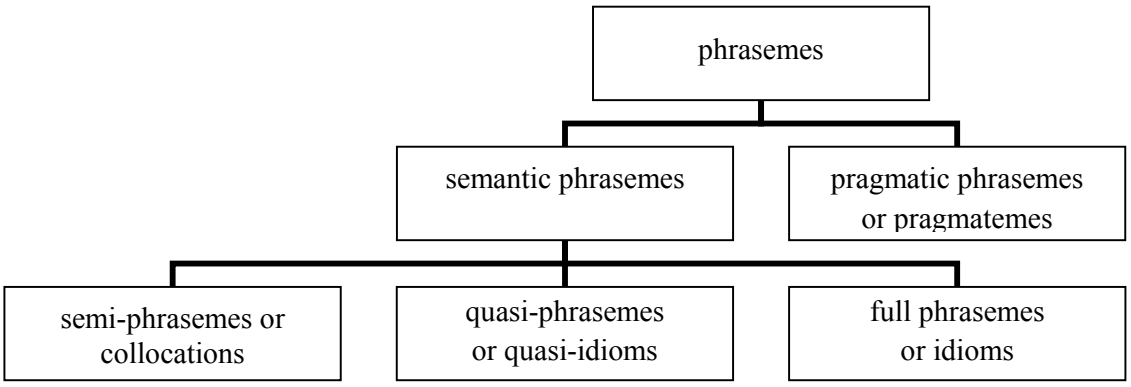

Figure 4: Mel'čuk's (1998) typology

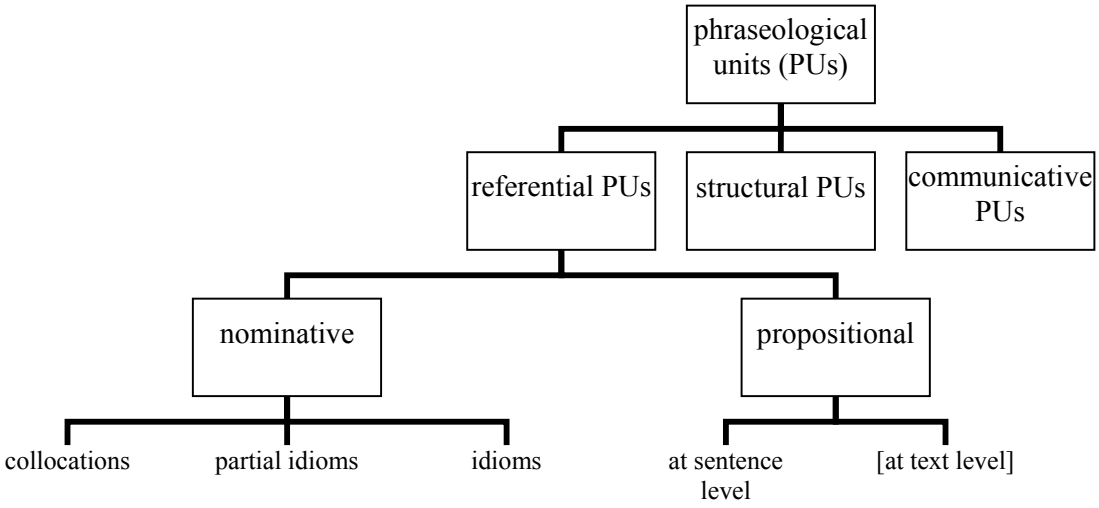

Figure 5: Burger's (1998) typology

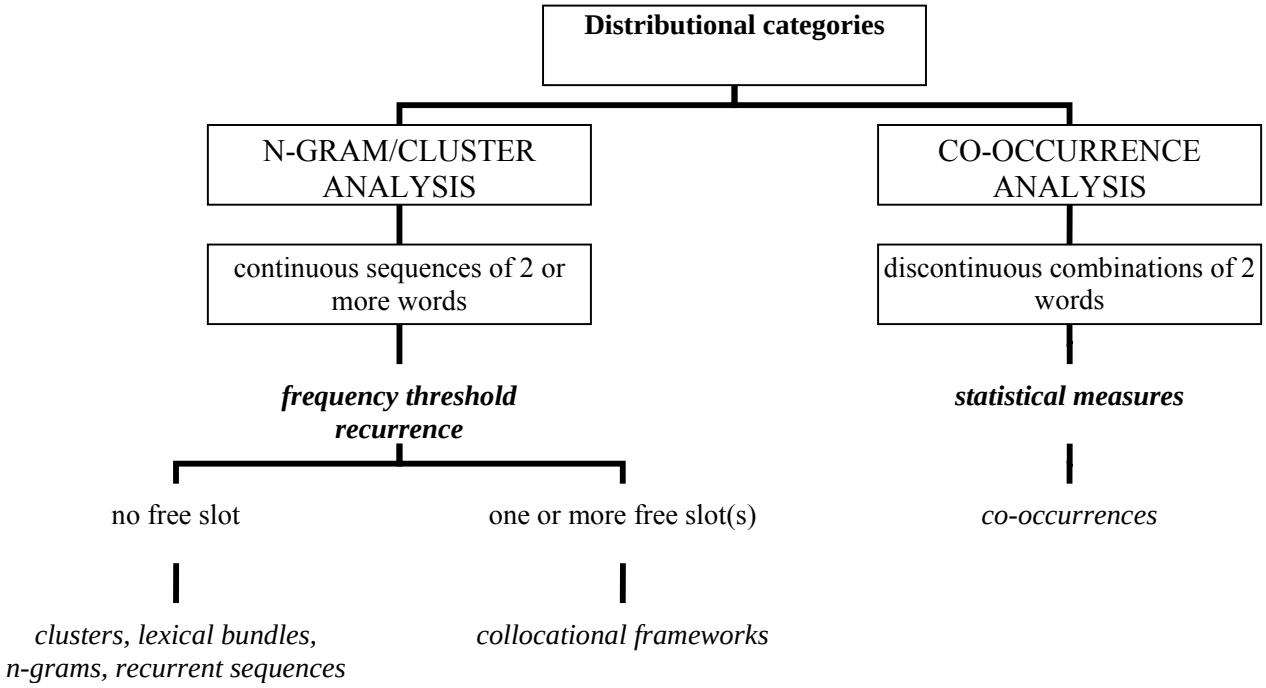

Figure 6: Distributional categories

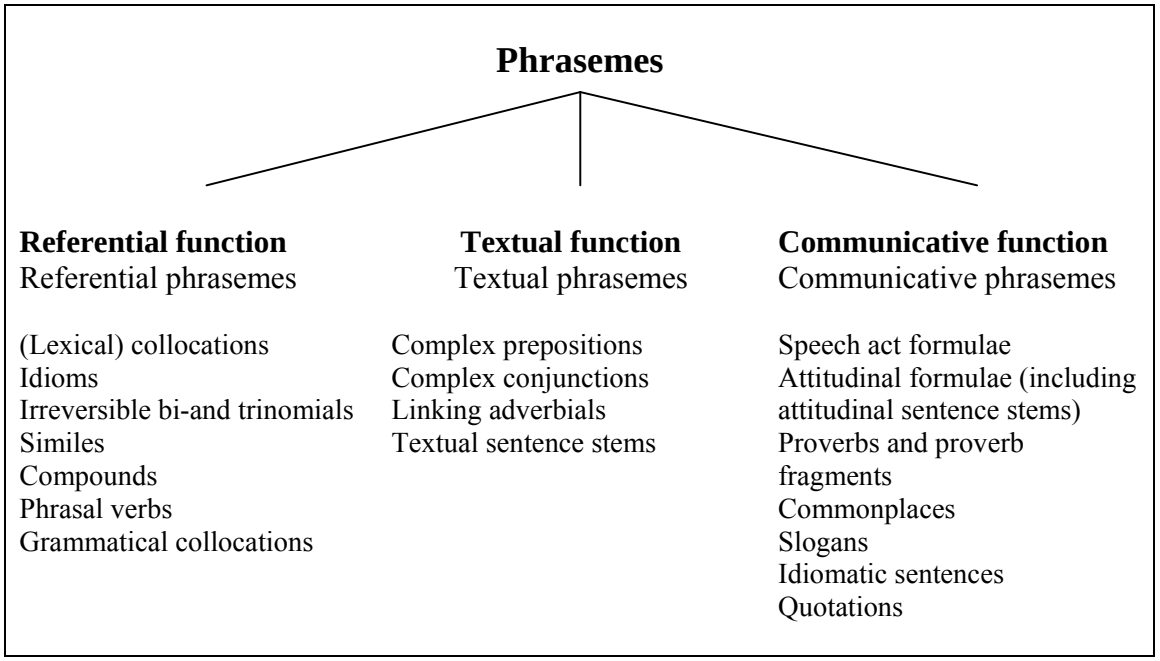

Figure 7: The phraseological spectrum

object of	**5522**	**2.7**	**subject of**	**1982**	**1.9**	**adj. modifier**	**6173**	**2.4**
adduce	64	46.31	suggest	412	51.95	circumstantial	83	54.25
provide	622	40.61	support	117	33.69	conclusive	94	51.82
give	941	39.04	indicate	82	32.04	empirical	163	50.61
obtain	130	29.55	point	59	29.77	anecdotal	67	50.26
gather	68	28.89	show	146	28.6	ample	91	45.51
produce	187	28.44	exist	43	26.81	archaeological	75	41.35
find	334	27.62	emerge	40	26.76	forensic	57	40.86
present	120	27.49	accumulate	20	26.33	further	283	40.76
hear	144	26.96	implicate	16	25.85	sufficient	148	39.12
collect	62	24.57	relate	52	24.82	supporting	67	38.98
n. modifier	**820**	**0.4**	**pp_of-p**	**3614**	**3.3**	**pp_on-p**	**282**	**1.4**
documentary	115	59.59	senses	24	23.66	oath	9	24.54
hearsay	30	47.97	efficacy	13	20.82	behalf	9	22.15
expert	62	36.45	infection	25	19.99	issue	10	13.54
affidavit	21	35.9	abuse	26	19.89	matter	7	11.52
dating	19	32.79	damage	31	18.96	subject	7	11.46
research	72	30.47	ischaemium	6	18.7	point	9	11.27
fossil	20	29.48	witness	20	18.6	ground	5	10.15
confession	14	26.24	nephropathy	5	18.08	nature	5	9.32
parole	5	25.92	competence	15	17.86	effect	6	9.1
video	21	22.34	disease	34	17.08	side	5	7.31
pp_in-p	**393**	**0.8**	**pp_obj_to-p**	**187**	**0.7**	**pp_obj_by-p**	**248**	**1.6**
case	49	26.39	relate	15	21.81	support	66	38.83
court	41	25.69	point	12	21.49	unsupported	10	34.21
trial	21	24.61	regard	7	20.73	substantiate	5	20.74
proceedings	14	22.43	listen	9	20.12	contradict	5	18.55
prosecution	7	17.04	refer	8	16.25	convince	6	17.31
favour	6	15.93	reference	6	13.86	justify	5	13.31
chief	6	13.41	apply	6	12.33	prove	6	12.27
form	9	10.16	add	6	11.43	confirm	5	11.99
action	6	8.57	give	8	8.15	establish	5	9.83
area	7	7.05	make	5	4.11	suggest	5	9.42

Table 1: A sample of the word sketch for the noun *evidence*

Category	Definition and illustration
(Lexical) collocations	(Lexical) collocations are usage-determined or preferred syntagmatic relations between two lexemes in a specific syntactic pattern. Both lexemes make an isolable semantic contribution to the word combination but they do not have the same status. Semantically autonomous, the 'base' of a collocation is selected first by a language user for its independent meaning. The second element, i.e. the 'collocate' or 'collocator', is selected by and semantically dependent on the 'base'. Examples: *heavy rain, closely linked, apologize profusely.*
Idioms	The category of idioms is restricted to phrasemes that are constructed around a verbal nucleus. Idioms are characterized by their semantic non-compositionality, which can be the result of a metaphorical process. Lack of flexibility and marked syntax are further indications of their idiomatic status. Examples: *to spill the beans, to let the cat out of the bag, to bark up the wrong tree*
Irreversible bi- and trinomials	Irreversible bi- and trinomials are fixed sequences of two or three word forms that belong to the same part-of-speech category and are linked by the conjunction 'and' or 'or'. Examples: *bed and breakfast, kith and kin, left, right and centre*.
Similes	Similes are sequences of words that function as stereotyped comparisons. They typically consist of sequences following the frames '*as* ADJ *as* (DET) NOUN' and 'VERB like a NOUN'. Examples: *as old as the hills, to swear like a trooper.*
Compounds	Compounds are morphologically made up of two elements which have independent status outside these word combinations. They can be written separately, with a hyphen or as one orthographic word. They resemble single words in that they carry meaning as a whole and are characterized by high degrees of inflexibility, viz. set order and non-interruptibility of their parts. Examples: *black hole*, *goldfish, blow-dry.*
Grammatical collocations	Grammatical collocations are restricted combinations of a lexical and a grammatical word, typically verb/noun/adjective + preposition, e.g. *depend on, cope with, a contribution to, afraid of, angry at, interested in*. The term 'grammatical collocation' is borrowed from Benson et al. (1997) but our definition is slightly more restricted as these authors also use the term to refer to other valency patterns, e.g. avoid + *-ing* form, which we do not consider to be part of the phraseological spectrum.
Phrasal verbs	Phrasal verbs are combinations of verbs and adverbial particles. Examples: *blow up, make out, crop up*.

Table 2: Categories of referential phrasemes

Category	Definition and illustration
Complex prepositions	Complex prepositions are grammaticalized combinations of two simple prepositions with an intervening noun, adverb or adjective. Examples: *with respect to, in addition to, apart from, irrespective of*
Complex conjunctions	Complex conjunctions are grammaticalized sequences that function as conjunctions. Examples: *so that, as if, even though, as soon as, given that*.
Linking adverbials	Linking adverbials include various types of phrasemes such as grammaticalized prepositional phrases, adjectival phrases, adverbial phrases, finite and non-finite clauses that play a conjunctive role in the text. Examples: *in other words, last but not least, more accurately, what is more, to conclude.*
Textual sentence stems	Textual sentence stems are routinized fragments of sentences that are used to serve specific textual or organizational functions. They consist of sequences of two or more clause constituents, and typically involve a subject and a verb. Examples: *the final point is …; another thing is …; it will be shown that …; I will discuss ….*

Table 3: Categories of textual phrasemes

Category	Definition and illustration
Speech act formulae	Speech act formulae (or routine formulae) are relatively inflexible phrasemes which are recognized by the members of a language community as preferred ways of performing certain functions such as greetings, compliments, invitations, etc. They display different degrees of compositionality. Examples: g*ood morning!, take care!, happy birthday!, you're welcome, how do you do*?
Attitudinal formulae	Attitudinal formulae are phrasemes used to signal speakers' attitudes towards their utterances and interlocutors. Examples: *in fact*, *to be honest, it is clear that*, *I think that*.
Commonplaces	Commonplaces are non-metaphorical complete sentences that express tautologies, truisms and sayings based on everyday experience. Examples: *Enough is enough, We only live once*, *it's a small world*).
Proverbs	Proverbs express general ideas by means of non-literal meaning (metaphors, metonymies, etc.). They are equivalent to complete sentences but are often abbreviated. Examples: *A bird in the hand is worth two in the bush, When in Rome*.
Slogans	Short directive phrases made popular by their repeated use in politics or advertising Example: *Make love, not war*.

Table 4: Categories of communicative phrasemes

Notes

1. There are other relevant fields, notably phonology/prosody, sociolinguistics and psycholinguistics (cf. Wray 2002).
2. Note, however, that some linguists include single words in phraseology. For example, Zuluaga (1980) includes *Salud* (En. 'cheers') and *Adiós* (En. 'bye bye') in the phraseological spectrum on the basis that these words display pragmatic fixedness (Sp. 'fijación pragmática').
3. Spanish and French phraseologists have tended to give more prominence to complex prepositions, complex conjunctions and complex adverbials and to classify them as phraseological units (cf. Gross 1996; Montoro del Arco 2006).
4. The term 'colligation' is often used as a near-synonym of grammatical collocation. For example, Stefanowitsch and Gries (2003: 210) define colligations as "linear co-occurrence preferences and restrictions holding between specific lexical items and the word-class of the items that precede or follow them". In their system, the word *involvement* is said to colligate with prepositions but to collocate with *in* and *with*.
5. "Von den drei Gruppen ist [strukturelle Phraseologismen] die kleinste und am wenigsten interessante." (Burger 1998: 37)
6. Although a collocational framework is defined by Renouf & Sinclair as "a discontinuous sequence of two words, positioned at one word remove from each other" (1991:128), we have classified them as recurrent sequences because they are not usually extracted by co-occurrence analysis but rather by means of n-gram analysis software tools (e.g. Fletcher's website: Phrases in English <http://pie.usna.edu/>).

References

Allerton, D.J. (1984). Three (or four) levels of word co-occurrence restriction. *Lingua* 63: 17–40.

Allerton, D.J., N. Nesselhauf & P. Skandera (eds.) (2004). *Phraseological Units: Basic Concepts and their Application.* Basel: Schwabe.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-Combinations. In Cowie, A.P. (ed.). *Phraseology. Theory, Analysis, and Applications*, 101–122. Oxford: Oxford University Press.

Bally, C. (1909). *Traité de stylistique française*. Paris: Klincksieck.

Barkema, H. (1996). Idiomaticity and terminology: a multi-dimensional descriptive model. *Studia Linguistica* 50(2):125-160.

Barnbrook, G. (1996). *Language and Computers*. Edinburgh: Edinburgh University Press.

Bartsch, S. (2004). *Structural and Functional Properties of Collocations in English*. Tübingen: Gunter Narr Verlag Tübingen.

Benson, M., E. Benson & R. Ilson (1986). *The Lexicographic Description of English*. Amsterdam & Philadelphia: John Benjamins.

Biber, D. (2004). Lexical bundles in academic speech and writing. In Lewandowska-Tomaszczyk, B. (ed.). *Practical Applications in Language and Computers (PALC 2003)*, 165–178. Frankfurt am Main: Peter Lang.

Biber, D. & S. Conrad (1999). Lexical bundles in conversation and academic prose. In Hasselgård, H. & S. Oksefjell (eds.). *Out of Corpora: Studies in Honour of Stig* Johansson, 181–190. Amsterdam: Rodopi.

Biber, D., S. Conrad & V. Cortes (2003). Lexical bundles in speech and writing: an initial taxonomy. In Wilson, A., P. Rayson & T. McEnery (eds.). *Corpus Linguistics by the Lune: a Festschrift for Geoffrey* Leech, 71–92. Frankfurt: Peter Lang.

Biber, D., S. Conrad & V. Cortes (2004). *If you look at* ….: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371– 405.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *Longman Grammar of Spoken and Written English.* Longman: Harlow.

Burger, H. (1998). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt Verlag.

Burger, H., D. Dobrovol'skij, P. Kühn & N.R. Norrick (eds.) (2007). *Phraseology: an International Handbook of Contemporary Research*. Berlin & New York: Mouton de Gruyter.

Burger H., D. Dobrovol'skij, P. Kühn & N.R. Norrick (2007). Phraseology: Subject area, terminology and research topics. In Burger, H., D. Dobrovol'skij, P. Kühn & N.R. Norrick (eds.). *Phraseology: an International Handbook of Contemporary Research*, 10–19. Berlin & New York: Mouton de Gruyter.

Cowie, A.P. (1981). The treatment of collocations and idioms in learners' dictionaries, *Applied Linguistics* 2(3): 223–235.

Cowie, A.P. (1988). Stable and creative aspects of vocabulary use. In Carter, R. & M.J. McCarthy (eds.). *Vocabulary and Language Teaching*, 126–137. London: Longman.

Cowie, A.P. (1994). Phraseology. In Asher, R.E. (ed.). *The Encyclopedia of Language and Linguistics*, 3168–3171. Oxford: Oxford University Press.

Cowie, A.P. (ed.) (1998). *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press.

Cowie, A.P. (1998a). Phraseological dictionaries: some East-West comparisons. In Cowie, A.P. (ed.). *Phraseology: Theory, Analysis and Applications*, 209–228. Oxford: Oxford University Press.

Cowie, A.P. (1998b). Introduction. In Cowie A.P. (ed.). *Phraseology: Theory, Analysis and Applications,* 1–20. Oxford: Oxford University Press.

Cowie, A.P. (2001). Speech formulae in English: problems of analysis and dictionary treatment. In van der Meer, G. & A.G.B. ter Meulen (eds.). *Making Senses: From Lexeme to Discourse. In Honor of Werner Abraham*, 1–12. Groninger Arbeiten zur germanistishen Linguistik 44. Center for language and Cognition Groningen.

Cowie, A.P. (2005). Review of S. Nuccorini (ed.) Phrases and Phraseology – Data and Descriptions. *International Journal of Lexicography* 18(1): 103–106.

Cowie, A., R. Mackin & I.R. McCaig (1983). *Oxford Dictionary of Current Idiomatic English.* Oxford: Oxford University Press.

De Cock, S. (2003). *Recurrent Sequences of Words in Native Speaker and Advanced Learner Spoken and Written English*. Unpublished doctoral dissertation. Université catholique de Louvain.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2: 225–246.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis. Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Gaatone, D. (1997). La locution : analyse interne et analyse globale. In Martins-Baltar, M. (ed.). *La Locution entre langue et usages*, 165–177. Langages. Fontenay-Saint Cloud: ENS éditions.

Giegerich, H.Z. (2004). Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8(1): 1–24.

Giegerich, H.Z. (2005). Associative adjectives and the lexicon-syntax interface*. Journal of Linguistics* 41: 571–591.

Gilquin, G., S. Granger & M. Paquot (2007) Learner corpora : The missing link in EAP pedagogy. In Thompson, P. (ed.). *Corpus-based EAP Pedagogy*. Special issue of *Journal of English for Academic Purposes* 6(4): 319–335.

Gläser, R. (1986). *Phraseologie der englischen Sprache*. Tübingen: Max Niemeyer Verlag.

Gläser, R. (1998). The stylistic potential of phraseological units in the light of genre analysis. In Cowie A.P. (ed.). *Phraseology. Theory, Analysis, and Applications*, 125–143. Oxford University Press: Oxford.

Gramley, S. & K.-M. Pätzold (1992). *A Survey of Modern English.* London & New York: Routledge.

Gross, G. (1996). *Les expressions figées en Français: noms composés et autres locutions.* Paris: Ophrys.

Heid, U. (2002). Collocations in lexicography. Presentation given at Colloc02, workshop on computational approaches to collocations, 23 August 2002. Austria: Vienna. Retrieved October 2007 from http://www.ofai.at/~brigitte.krenn/colloc02/workshop_prog.html

Herbst, T., D. Heath, I. Roe & D. Götz (2004). *A Valency Dictionary of English: A Corpus-Based Anaysis of the Complementation Patterns of English Verbs, Nouns and Adjectives.* Berlin & New York: Mouton de Gruyter.

Hoey, M. (2004). The textual priming of lexis. In Aston, G., S. Bernardini & D. Stewart (eds.) *Corpora and Language Learners*, 21–41. Amsterdam & Philadelphia: Benjamins.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London & New-York: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kilgarriff, A., P. Rychly, P. Smrz & D. Tugwell (2004). The Sketch Engine. In Williams, G. & S. Vessier (eds.). *Proceedings of the Eleventh EURALEX International Congress*, 105–116. Lorient: Université de Bretagne-Sud.

Krishnamurthy, R. (1987). The Process of Compilation. In Sinclair, J. (ed.). *Looking Up. An Account of the COBUILD Project in Lexical Computing*, 62–85. London & Glasgow: Collins ELT.

Lewis, M. (1993). *The Lexical Approach: The State of ELT and a Way Forward.* Hove: Language Teaching Publications.

Louw, B. (1993). Irony in the text or insincerity in the writer? In Baker, M., G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair*, 157–176. Amsterdam: Benjamins.

Louw, B. (2000). Contextual prosodic theory: bringing semantic prosodies to life. In Heffer, C., H. Sauntson & G. Fox (eds.). *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Manning, C. & H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT press.

McEnery, A., R. Xiao & Y. Tono (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London & New-York: Routledge.

Mejri, S. (2005). Introduction: polysémie et polylexicalité. In Mejri, S. (ed.). *Polysémie et Polylexicalité. Syntaxe et Sémantique* 5 : 13–30.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. In Everaert, M., E.J. Van der Linden & A. Schenk (eds.). *Idioms: Structural and Psychological Perspectives*, 167–232. Hillsdale: Lawrence Erlbaum Associates.

Mel'čuk, I. (1998). Collocations and lexical functions. In Cowie, A.P. (ed.). *Phraseology. Theory, Analysis, and Applications*, 23–53. Oxford: Oxford University Press.

Montoro del Arco, E.T. (2006). *Teoría fraseológica de las locuciones particulares: las locuciones prepositivas, conjuntivas y marcadoras en español*. Frankfurt am main: Peter Lang.

Moon, R. (1998). Frequencies and forms of phrasal lexemes in English. In Cowie, A.P. (ed.). *Phraseology. Theory, Analysis, and Applications*, 79–100. Oxford: Oxford University Press.

Nattinger, J.R. & J.S. DeCarrico (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nesselhauf, N. (2004). What are collocations ? In Allerton, D.J., N. Nesselhauf & P. Skandera (eds.). *Phraseological Units: Basic Concepts and their Application,* 1–21. Basel: Schwabe Verlag.

Palmer, H.E. (1933). *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Partington, A. (2004). "Utterly content in each other's company": semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1): 131–156.

Pecman, M. (2008). Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In Meunier, F. & S. Granger (eds.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: Benjamins.

Renouf, A. & J. Sinclair (1991). Collocational frameworks in English. In Aijmer, K. & B. Altenberg (eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik,* 128–143. London & New York: Longman.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002). Multi-word 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, 1–15.

Schmid, H.-J. (2003). Collocation: hard to pin down, but bloody useful. *ZAA* 51(3): 235–258.

Scott, M. & C. Tribble (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: Benjamins.

Siepmann, D. (2006). Collocation, colligation and encoding dictionaries. Part II: lexicographical aspects. *International Journal of Lexicography* 19(1): 1–39.

Sinclair, J. (ed.) (1987). *Looking Up. An Account of the COBUILD Project in Lexical Computing.* London & Glasgow: Collins Cobuild.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996) The search for units of meaning. *TEXTUS* IX: 75–106.

Sinclair, J. (1998). The lexical item. In Weigand, E. (ed.) *Contrastive Lexical Semantics*, 1–24. Amsterdam: Benjamins.

Sinclair, J. (2004). Trust the text. In Sinclair, J. & R. Carter (eds.) *Trust the Text – Language, Corpus and Discourse*, 9–23. London: Routledge.

Stefanowitsch, A. & S. Gries (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language.* Oxford: Basil Blackwell.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23–55.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2): 215–244.

Stubbs, M. (2007a). Quantitative data on multi-word sequences in English: the case of the word 'world'. In Hoey, M., M. Malhberg, M. Stubbs & W. Teubert (eds.). *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum.

Stubbs, M. (2007b). An example of frequent English phraseology: distribution, structures and functions. In Facchinetti, R. (ed.). *Corpus Linguistics 25 Years on*, 89–105. Amsterdam & New York: Rodopi.

Stubbs, M. & I. Barth (2003). Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language* 10(1):61–104.

Svensson, M.H. (2002). Critères de figement et conditions nécessaires et suffisantes. *Romansk Forum* 16(2): 777–783.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam & Philadelphia: Benjamins.

Tognini-Bonelli, E. (2002). Functionally complete units of meaning across English and Italian: towards a corpus-driven approach. In Altenberg, B. & S. Granger (eds.) *Lexis in Contrast: Corpus-based Approaches*, 73–95. Amsterdam & Philadephia: Benjamins.

Tschichold, C. (2000). *Multi-word Units in a Lexicon for Natural Language Processing*. Olms: Hildesheim.

Woolard, G. (2000). Collocation – encouraging learner independence. In Lewis M. (ed.). *Teaching Collocation: Further Developments in the Lexical Approach*, 28–46. Hove: Language Teaching Publications.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. & M. Perkins (2000). The functions of formulaic language: an integrated model. *Language and Communication* 20: 1–28.

Zuluaga, A. (1980). *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter Lang.