# Phraseology and linguistic theory

## A brief survey

Stefan Th. Gries

This chapter has three objectives. First, it argues in favor of more rigorous definitions of the term 'phraseologism' on the basis of six dimensions and exemplifies these dimensions for several different kinds of phraseologism. Second, it reviews the ways in which phraseologisms as defined here have figured in three different linguistic approaches: generative linguistics, cognitive linguistics, and corpus linguistics. Finally, it discusses some shortcomings in the identification of phraseologisms and points to relevant work to overcome these shortcomings.

### 1. Introduction

Interest in phraseology has grown considerably over the last twenty years or so. While the general linguists' view of phraseology before that time can probably be caricatured as 'idiom researchers and lexicographers classifying and researching various kinds of fairly frozen idiomatic expressions', this view has thankfully changed. Nowadays, the issues of identifying and classifying phraseologisms as well as integrating them into theoretical research and practical application has a much more profound influence on researchers and their agendas in many different sub-disciplines of linguistics as well as in language learning, acquisition, and teaching, natural language processing, etc. However, this influence is often not fully recognized or acknowledged, or reflected terminologically. This is undesirable, not only because it is often not easy to recognize the domains where research on phraseology has left its marks, but also because it renders the overlap of assumptions, concepts, and findings less transparent than is desirable.

This chapter attempts to take a modest step in this direction. I will try to identify and make explicit six crucial dimensions, or defining parameters, of phraseologisms. I think these actually underlie most phraseological work – if only implicitly – but I would like phraseologists to always be maximally explicit about which parameter settings are adopted in order to (i) render their definitions maximally precise and (ii) allow researchers from other frameworks to more easily recognize potential areas of overlap, or indeed conflict. In Section 3 I will then use the suggested parameters to discuss the role phraseologisms have played in different linguistic frameworks, viz. transformational-generative grammar, cognitive linguistics and

Construction Grammar, as well as corpus linguistics. While most scholars do not view corpus linguistics as a linguistic theory but rather as a methodology, it has given rise to many theoretical assumptions that, I believe, warrant its inclusion here. In Section 4, I will briefly, but critically, evaluate the methods practitioners in these three approaches have used to identify phraseologisms. Section 5 will conclude.

## 2. The notion of phraseology

While the notion of *phraseology* is very widespread, just as with other linguistic concepts, different authors have defined it differently, sometimes not providing a clear-cut definition, or conflating several terms that many scholars prefer to distinguish.[1] However, a closer comparative look at the vast majority of studies that exist allows us to identify a set of parameters that are typically implicated in phraseological research. I believe a rigorous definition of co-occurrence phenomena in general, and phraseology in particular, needs to take a stand regarding at least the following six parameters (cf. Howarth 1998: 25 for a similar critique of the absence of defining criteria and an alternative proposal).

i. the *nature* of the elements involved in a phraseologism;
ii. the *number* of elements involved in a phraseologism;
iii. the *number of times* an expression must be observed before it counts as a phraseologism;
iv. the permissible *distance* between the elements involved in a phraseologism;
v. the degree of *lexical and syntactic flexibility* of the elements involved;
vi. the role that *semantic unity* and *semantic non-compositionality / non-predictability* play in the definition.[2]

As to the first criterion, the definition of a phraseologism I will adopt is among the broadest conceivable. I consider a phraseologism to be the co-occurrence of a form or a lemma of a lexical item and any other kind of linguistic element, which can be, for example,

- another (form of a) lexical item (*kith and kin* is a very frequently cited example of a nearly deterministic co-occurrence of two lexical items, as is *strong tea*);
- a grammatical pattern (as opposed to, say, a grammatical relation), i.e. when a particular lexical item tends to occur in/co-occur with a particular grammatical construction (the fact that the verb *hem* is mostly used in the passive is a frequently cited case in point).

Note that this definition does not distinguish between lexical items and grammatical patterns that co-occur with a lexical item. Also, note that the definition does not commit to a particular level of granularity regarding the lexical elements involved: both can involve either all forms of a lemma or just particular morphological forms (cf. Rice & Newman 2005 and Gries, to appear, for conflicting points of view on this matter).

As to the second criterion, it is important to decide whether, for example, phraseologisms can consist of only two elements (such as word pairs) or can include a larger number of elements. I will assume that phraseologisms can contain more than two elements (as in, say, *to eke out a living*, which contains a verb slot filled with some form of *to eke*, a direct object slot filled with DET *living*, and a slot for the particle *out* complementing the transitive phrasal verb).

As to the third criterion, it is probably fair to say that there is little work which has defined phraseologisms solely on the basis of some quantitative criterion based on their frequency of occurrence (and/or additional frequency information). True, some scholars have used a threshold of absolute frequency of occurrence (usually defined arbitrarily or not at all; cf. Hunston & Francis 2000: 37, for example). Others, most notably British and Scandinavian scholars from the Sinclairian/Cobuild tradition, have argued that observed frequencies must exceed frequencies expected on the basis of chance (significantly or just at all), but most previous work has restricted itself to reporting frequencies or percentages of occurrence of phraseologisms. In order to avoid an inflation of what could be considered phraseological, I will consider an expression a phraseologism if its observed frequency of occurrence is larger than its expected one.

As to the fourth criterion, some work (especially *n*-gram-based studies in natural language processing) concerns itself only with immediately adjacent elements, but I will adopt the more widespread broader perspective which also recognizes discontinuous phraseologisms.

As to the fifth criterion, studies that are only concerned with completely inflexible patterns such as the standardly quoted example of *by and large* can be distinguished from studies that include relatively flexible patterns such as *kick the bucket* (which allows different tenses but no passivization), studies (also) involving partially lexically-filled patterns such as the *into*-causative ([$_{\text{VP}}$ V DO *into* V-*ing*]), and finally studies (also) including completely lexically unspecified and thus maximally flexible

[1] A case in point is Stubbs (2001). According to the index, the term *phraseology* and the cross-referenced *extended lexical unit* are first mentioned on p. 59 and p. 31f. respectively. However, no explicit definition of *phraseology* is provided on these pages nor on the page where *phraseology* is first mentioned (p. 24). Another example is Hunston (2002: 137f.), who first discusses "some examples" she categorizes as "[c]ollocation", "[p]hrases and variation", "[t]he tendency of certain verbs to occur in the passive rather than the active, or in the negative rather than the positive" (i.e., what is usually referred to as colligation), and "[t]he occurrence of complementation patterns", but then merely states that "[t]hese and the other consequences of sequence preference together might be called 'phraseology'". As will become apparent below, I largely agree with Hunston's inclusion of these examples as phraseologisms, but the definition as such is not as explicitly delineated as it could be and leaves much to inference processes on the part of the reader.

[2] Additional or alternative criteria one might wish to invoke are a possible separation of lexical flexibility and syntactic flexibility (or commutability/substitutability) of the elements involved in potential phraseologisms and/or the distinction between encoding and decoding idioms.

expressions, such as the English ditransitive pattern $[_{VP}$ V OBJ$_1$ OBJ$_2]$ (cf. Section 3.2 for references). My definition of phraseologisms excludes only the last of these because they do not involve at least one lexically specified element (as required by the first parameter).

As to the final, and for many researchers probably most important, criterion, the elements of a phraseologism – however they are distributed across a clause or sentence – are usually assumed to function as a semantic unit, i.e. to have a sense just like a single morpheme or word. However, one can distinguish between studies in which the sense of a phraseologism is by definition non-compositional (cf. Fraser's 1976: v definition of an idiom as "a single constituent or series of constituents, whose semantic interpretation is independent of the formatives which compose it") from studies where non-compositional semantics is not a necessary condition for phraseologisms (cf. Nunberg, Sag, & Wasow 1994: 499ff. as well as Wulff to appear and below for further discussion). For something to count as a phraseologism, I will require semantic unity, but not non-compositional semantics.

In sum, a phraseologism is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance.

While this definition is maximally explicit with respect to the above-mentioned parameters, it also follows that, as in Hunston's (2002) approach, the range of phenomena regarded as phraseologisms is very large. An example from the inflexible end of the continuum of phraseologisms is the largely fixed expression *to run amok*, which can be analyzed with respect to the six above criteria as follows:

- nature of the elements: words;
- number of elements: two;
- frequency of occurrence: the two parts of the expression co-occur more often than expected by chance: in the British National Corpus World Edition (BNC WE), any form of *to run* and *amok* occur in 38,088 and 43 of all 6,051,206 sentence units (lines beginning with "<s n=") respectively; thus, one would expect 0.27 within-sentence unit co-occurrences, but one actually obtains 40;[3]

- distance of elements: the two parts of the phraseologism usually co-occur adjacently (in all but one case, where *dangerously* intervenes);
- flexibility of the elements: *to run* can occur in various morphological forms, but *amok* can apparently not be preposed (judging from the hits in the BNC WE, that is);
- semantics: *to run amok* functions as one semantic unit, meaning roughly 'to behave violently and uncontrollably'.

Another group of examples are transitive phrasal verbs such as *to pick up*, *to give up*, or the concrete example of *to eke out* (usually *a living* or *an existence*), etc.:

- nature of the elements: words and phrases in a transitive phrasal verb frame (the direct object can be an NP or a clause);
- number of elements: two lexical elements and one syntactic slot to be filled;
- frequency of occurrence: in the BNC WE, any form of the verb *to eke* ("<w (VV.|VV.-...)>ek(e[sd]?|ing)\\W") and *out* tagged as an adverbial particle ("<w AVP>out\\W") occur in 78 and 140,975 of all 6,051,206 sentence units respectively (with a case-insensitive search), which is why one would expect 1.8 co-occurrences, but one actually obtains 70;
- distance of elements: the verb and the particle can occur right next to each other or with intervening material (the maximum length of direct objects in verb-particle constructions in general reported by Gries 2003 is 21);
- flexibility of the elements: the verb, the direct object, and the particle allow for constituent order variation in that they need not be adjacent, allow passivization, ...;
- semantics: transitive phrasal verbs function as one semantic unit, which is evidenced by (i) the well-known fact that many have a one-word near synonym (*to pick up*: *to lift/elevate*; *to give back*: *to return*, *to put down*: *to deposit*) and (ii) by the fact that many have non-compositional readings (or even a compositional and a non-compositional reading such as *to hold up* or *to throw up).*

A final group of examples are patterns which (i) are lexically partially filled, (ii) require the insertion of additional lexical material, and (iii) allow for syntactic variation, such

[3] All corpus data have been retrieved with regular expression searches performed with R for Windows 2.4 (cf. R Development Core Team 2006). The data discussed here are based on a retrieval of all case-insensitive matches for ">amok\\W" and "<w (VV.|VV.-...)>(ran|run(s|ning)?)\\W" in all lines with sentence units from the BNC WE. This, like the additional corpus data reported below, is of course only an approximation because it is only most, but not all within-sentence-unit co-occurrences that instantiate the construction in question. The expected frequency has been computed as is customary in nearly all measures of collocational strength or chi-square tests by multiplying the observed marginal totals of the two individual items in question and dividing by the corpus size. One might suspect that the frequency of co-occurrence criterion could be problematic for cases where the potential phraseologism involves one or more high-frequency items such as *to break the ice*. However, even in such cases the number of observed co-occurrences exceeds the expected value, as can be seen by retrieving all case-insensitive matches for "<w (VV.|VV.-...)>(break(s|ing)?)|(broken?)\\W" and "<w (N..|N..-...)>ice\\W" in all sentence unit lines of the BNC WE. These searches yielded 22,256 and 4,392 matches respectively, so 16.2 co-occurrences would be expected, but in fact 125 sentence units with both search strings were observed, approximately half of which instantiated the idiom in question.

as the comparative clause construction (i.e., [$_{XP}$ *the* Adj$_{comparative}$, *the* Adj$_{comparative}$]) or the *into*-causative (i.e., [$_{VP}$ V DO *into* V*ing*]).[4]

Given the broad definition proposed above, it might seem as if now everything in language is phraseological and *phraseologism* is a futile catch-all term devoid of empirical content and unworthy of empirical study. However, this is not the case. On the one hand, the definition does not include highly frequent co-occurring expressions such as *of the* or *in the*, because these do not function as a semantic unit. Also, the definition does not include completely lexically unspecified patterns such as those that Construction Grammarians refer to as argument structure constructions (cf. below Section 3.2; examples include the ditransitive construction [$_{VP}$ V OBJ$_1$ OBJ$_2$] or the caused-motion construction [$_{VP}$ V DO OBL]), which bridge the gap to the patterns posited in Pattern Grammar. As such and in other words, the definition of phraseologism proposed above serves as a convenient cover term for co-occurrence phenomena at the *syntax-lexis interface* (since at least one lexical element must be specified) as opposed to the *syntax-semantics interface*, to which lexically unspecified patterns from Pattern Grammar or argument structure constructions from Construction Grammar would be associated. On the other hand, the present definition does cover particular words' significant attraction to argument structure constructions (cf. Stefanowitsch & Gries's (2003) collostructions) and completely lexically filled and frozen expressions which, although diachronically derived from collocations, are synchronically single lexemes (e.g. *of course*, *at least*).

Not all researchers would of course subscribe to the parameter settings I have proposed and/or would prefer to exclude some of these and/or include additional ones (see Note 2). If, for example, non-compositionality were taken as a necessary condition for something to count as a phraseologism, many highly frequent but fully compositional elements would no longer count as phraseologisms anymore. Similarly, if the requirement for at least one specified lexical element were dropped, argument structure constructions would belong to the realm of phraseologisms.

However, phraseologisms as defined above are worthy of empirical study because the present definition does not single out any particular level of granularity at which co-occurrences, and thus phraseologisms, may be observed. This has two interesting consequences. First, it means that phraseologists must carefully define the linguistic level(s) at which they observe a potential phraseologism. In the case of the phrasal verb *to eke out a living*, for example, one could recognize at least the following co-occurrences as potential phraseologisms:

- *to* [$_{VP}$ *eke out a living*];
- *to* [$_{VP}$ *eke out* DO];
- *to* [$_{VP}$ V *out* DO];

- *to* [$_{VP}$ V Particle DO]; or even
- *to* [$_{VP}$ V Particle DO$_{NP}$] (while the DO does not have to be an NP, it usually is and one may want to include this probabilistic information); ...[5]

The crucial question is to decide which level of resolution to focus on, an issue that will sometimes be decided on the basis of a particular researcher's interest but which can also be decided purely quantitatively by, say, measuring the level of granularity at which the attraction between the elements involved is highest.

As another example, if one retrieved from a corpus many instances of the ditransitive pattern [$_{VP}$ V OBJ$_1$ OBJ$_2$] and inspected the verbs occurring in them, one could draw many different probabilistic conclusions about co-occurrence preferences. One could concentrate on

- the strong positive correlation between the verb form *gave* and the ditransitive;
- the strong positive correlation between the verb lemma GIVE and the ditransitive (as in collexeme analysis; cf. Stefanowitsch & Gries 2003);
- the strong positive correlation between the semantic class of transfer verbs and the ditransitive;
- etc.

However, not all of these are theoretically revealing or relevant (cf. Gries 2006b & to appear for discussion and exemplification of differences between word-form specific and lemma-specific results as well as differences between speaking and writing). On the whole, I think it is fair to say that there is as yet little empirically rigorous work on this issue.

Phraseologists must also decide how many elements a phraseologism is supposed to comprise. The potential phraseologisms listed above, derived from the example of *to eke out a living*, all involved three elements, but on occasion this may not be the most revealing analysis. Similarly, if semantic unity were not required for something to count as a phraseologism, one could posit that *in spite* is a phraseologism: it involves two words (number and nature of elements) that co-occur more often than expected by chance,[6] are adjacent and inflexible. However, it is obvious that a more reasonable assumption would be that the 'real' phraseologism is *in spite of*, which is what statistically more sophisticated approaches would recognize (cf. Mason's work on lexical gravity and Kita et al.'s cost criterion mentioned below in Section 4.3).

[4] Given particular lexical material and some syntactic structure, such phraseologisms may well develop into completely frozen units, as exemplified by the proverbial instance of the comparative clause construction *the more, the merrier*.

[5] Of course, not all these examples qualify as a phraseologism according to my definition (some are not a single semantic unit and the last two do not involve at least one specific lexical item). However, they may be phraseologisms according to other scholars' definitions.

[6] This claim is based on retrievals of all case-insensitive matches for "<w PRP>in\\W", "<w PRP>in spite\\W", and "\\Wspite\\W" in all lines of corpus files from the BNC WE that begin with "^<s n=", yielding 1,361,163, 2,683, and 2,897 matches respectively. In other words, nearly all occurrences of *spite* are actually instances of *in spite*.

Thus, even though the above definition may seem overly powerful at first, it still delimits the possible space of phraseologisms effectively and leaves room for many issues to be discussed. The central point to be made here, however, is that phraseologists should formulate more definitions of this kind. By this I do not mean that phraseologists should necessarily adopt *my* definition – but that it is essential that we, who are interested in something as flexible as patterns of co-occurrence, always make our choice of parameter settings maximally explicit to facilitate both the understanding and communication of our work. Using the above six parameters, the following section will explore how phraseologisms figure – sometimes rather implicitly – in different linguistic approaches.

## 3. The role of phraseology in linguistic theory

The role phraseology has played in linguistic theory is quite varied. On the one hand, it is varied because theoretical frameworks or approaches in linguistics differ widely in terms of the importance attached to phraseologisms. On the other hand, the importance that phraseology can play in a framework also crucially depends, of course, on how phraseologisms are defined, which is why I devoted so much space to the question of definition in Section 2. Space does not allow for a comprehensive comparison of the role of phraseology in many different frameworks so I have to be selective. Section 3.1 looks at transformational-generative linguistics. Section 3.2 discusses phraseology in cognitive linguistics and Construction Grammar while Section 3.3 is concerned with phraseology in corpus linguistics.

### 3.1 Generative linguistics

It is probably fair to say that phraseology has generally played a rather limited role in the development of the various versions of generative grammar. Given a conception of the linguistic system which crucially involves only

- a grammar, i.e. a set of algorithmic rules that combines linguistic elements only with respect to their structural characteristics and irrespective of their meaning; and
- a lexicon, i.e. a repository of all non-compositional irregularities that must be rote-learned;

it comes as no surprise that, of the above six parameters, the only one which plays a role for generative linguistics is the last one, semantic unity and non-compositionality. In this conception, an expression such as *to bite the dust* is recognized as an idiom, a non-compositional semantic unit as defined by Fraser (1976: v), and is thus stored with its syntactic characteristics as a separate item in the lexicon. Note also that this conception of the linguistic system is somewhat at odds with my definition of phraseologisms which does not treat grammatical and lexical elements as different in kind.

This generative conception of phraseologisms comes with a few problems. On the one hand, it is much more difficult to draw a strict dividing line between what is idiomatic and what is not than one may initially assume; cf. Fraser (1966: 59, n. 3) for the difficulty of obtaining unanimous judgments as well as Cowie & Mackin (1993: ix) and Gibbs (1994: Ch. 5–6) for discussion. On the other hand, research has shown that phraseologisms/idioms vary considerably in terms of the syntactic operations they allow for (cf. the seminal paper by Nunberg et al 1994 for discussion), and since not all of these can be explained away by straightforward performance factors, one would have to postulate that the lexicon contains for each putative unit a list of what operations are licensed, an option that is particularly unattractive for an approach that otherwise eschews redundant representation.

It is only in more recent developments of the generative framework that the importance of phraseologisms has come to be recognized more openly. For example, Culicover (1999) insightfully discusses a variety of patterns that are usually classified as phraseologisms (examples include *had better*, *not*-topics, etc.) and points out that they pose serious challenges to a modular organization of language in terms of an algorithmic grammar and a lexicon because they appear to cut across this supposedly well-established boundary. A similar tack is taken in some recent work by Jackendoff. To name but one example, Jackendoff (1997) is concerned with a phraseological expression – the 'time' *away* construction exemplified by *We're twistin' the night away*, which, given its properties with respect to the above parameters, would certainly be recognized as a phraseologism by most phraseologists:

- nature of the elements: words and phrases in a transitive phrasal verb frame;
- number of elements: three: the V-slot must be filled with an intransitive verb; the DO slot must be filled with a time expression; the particle is *away*;
- frequency of occurrence: In the BNC WE, *night/nights* tagged as nouns ("<w NN[12]>night(s)?\\W") and *away* tagged as an adverbial particle ("<w AV0>away \\W") occur in 36,265, and 34,343 of all 6,051,206 sentence units respectively, which is why one would expect 206 co-occurrences, but one actually obtains 512;[7]
- distance of elements: the intransitive verb, the direct object, and *away* occur right next to each other;
- flexibility of the elements: just like regular transitive phrasal verbs, the intransitive verb, the direct object, and the particle can occur in the order [$_{VP}$ V DO Particle] or in the order [$_{VP}$ V Particle DO]; passivization and tough movement are possible, but rare;
- semantics: the pattern of transitive phrasal verbs with time expressions as direct object and *away* functions as a semantic unit, as is evidenced by the fact that this pattern forces a particular interpretation of the clause such that the referent of the

[7] Again, the resulting frequencies need of course not all be instances of the 'time' *away* construction but only serve to make the point that the observed co-occurrence frequency of *night* and *away* most likely exceeds chance levels considerably.

subject is understood to act volitionally; the verb must denote an activity, not a state, and the referent of the subject uses up the whole time denoted by the time expression (cf. Jackendoff 1997:534–7).

I am not aware that this has been recognized or even openly acknowledged by transformational-generative grammarians, but it is interesting to note that the notion of phraseologism, which has been rather on the fringe of transformational-generative grammar in particular and most of theoretical linguistics in general, is so crucial to the revision of the most dominant linguistic paradigm of the 20th century and, thus, to the way the linguistic system proper is viewed. More specifically, it is, among other things of course, the recognition of phraseologisms as theoretically relevant entities in their own right that begins (i) to undermine the modular organization of the linguistic system into a grammar and a lexicon and (ii) to make linguists aware of the way in which the analysis of phraseologisms in performance data reveals many subtle interdependencies on different levels of linguistic analysis. While this interpretation may be controversial, I believe it is supported by the fact that the next two frameworks or approaches to be discussed – cognitive linguistics / Construction Grammar and corpus linguistics – also rely heavily on the notion of phraseologism as I have defined it above, even though the term *phraseologism* is not always used. These parallels will be outlined in more detail in the following two sections.

### 3.2 Cognitive linguistics and Construction Grammar

As mentioned above, the discussions by Culicover and Jackendoff of what we have been referring to as phraseologisms have not made use of this term. However, the way their analyses are phrased makes the connections not only to the notion of phraseologism, but also to other theoretically related concepts rather obvious. Two related theoretical frameworks whose practitioners are currently very much concerned with phraseologisms are cognitive linguistics and Construction Grammar.

Cognitive linguistics as such is not so much a single theory as a set of related approaches that share several fundamental assumptions which set it apart from other competing frameworks. The same is true of Construction Grammar, where one may distinguish at least between, say, the version of Construction Grammar by Goldberg (1995, 2006), that of the Berkeley school (cf. the references to the works by Fillmore and Kay below), Croft's (2001) Radical Construction Grammar, and maybe others. My discussion of cognitive linguistics and Construction Grammar cannot encompass all the different approaches. Instead, for cognitive linguistics, I will focus on what I consider the most thoroughly developed approach, namely Langacker's Cognitive Grammar as outlined in Langacker (1987, 1991); my discussion of Construction Grammar will focus on Goldberg's version. As will become more apparent below, these two theories' equivalents of the notion of *phraseologism* are very similar, but until recently differed with respect to one of the above defining parameters of phraseologisms, viz. non-compositionality.

Cognitive Grammar as a discipline does not really have a theoretical notion that is a precise equivalent of *phraseologism*. Rather, it has a more general term, of which phraseologisms constitute a subset. As I did above, Cognitive Grammar does away with a strict separation between lexicon and grammar. The only kinds of element the linguistic system is said to contain are symbolic units. A unit is defined as

> a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement [...] he has no need to reflect on how to put it together. (Langacker 1987:57)

A symbolic unit in turn is a pairing of a form and a meaning/function, i.e. a conventionalized association of a phonological pole (i.e., a phonological structure) and a semantic/conceptual pole (i.e., a semantic/conceptual structure). The more often a speaker/hearer encounters a particular symbolic unit, the more entrenched this symbolic unit becomes in his or her linguistic system and the more automatically the unit is accessed. Thus, unit status correlates positively with a speaker/hearer not analyzing the internal structure of a unit. Crucially for our present purposes, the notion of symbolic unit is not restricted to morphemes or words, but comprises more abstract grammatical patterns such as transitive constructions, reference-point constructions (such as the *s*-genitive in English), idioms, etc. Using my defining parameters of a phraseologism, a symbolic unit can be defined as follows:

- nature of the elements: no restrictions as long as the forms in the expression are paired with some meaning;
- number of elements: no restrictions;
- frequency of occurrence: a symbolic unit must have occurred frequently enough for it to be entrenched in a speaker/hearer's linguistic system (I am, however, not aware of any rigorous operationalization of a sufficient frequency threshold);
- distance of elements: no restrictions as long as the speaker/hearer categorizes the parts as making up one symbolic unit;
- flexibility of the elements: no restrictions as long as the speaker/hearer can form one or more generalizations (a schema in Langacker's parlance) which sanction the concrete instances; for example, if a speaker recognizes that two expressions instantiate transitive constructions, it is unimportant that the two instances may contain different verbs in different tenses with different and/or differently long objects etc.;
- semantics: by definition, the symbolic unit must have a semantic pole or meaning/function, but non-compositionality is not required.

This definition is of course not only Langacker's; other scholars such as Bybee (1985) also subscribe to this kind of definition. This definition of a symbolic unit is nearly identical to that of a phraseologism given above: it is only somewhat broader, including as it does simple words/morphemes and also lexically unspecified patterns. However, given this definition, phraseologisms do not enjoy a special status within

Cognitive Grammar: they are just one kind of symbolic unit, requiring the same descriptive apparatus as the more specific categories of morphemes or words or the more general categories of argument structure constructions or clause patterns. In terms of what they consider the central units of analysis, Cognitive Grammar and phraseology research are, thus, nearly maximally compatible.

As will become obvious below, there is a similar degree of compatibility between Construction Grammar and phraseological research. Given the theoretical affinity of Cognitive Grammar and Construction Grammar and the parallel evolution of the two theories, this should not come as a big surprise, and the main difference between how Cognitive Grammar and Construction Grammar define their objects of study as compared to phraseological research is largely terminological. The central linguistic unit of Construction Grammar – the analogon to symbolic units in Cognitive Grammar – is the construction. A construction in the sense of Goldberg's (1995) Construction Grammar is defined as follows:

> C is a construction iff$_{\mathrm{def}}$ C is a form-meaning pair <F$_i$, S$_i$> such that some aspect of F$_i$ or some aspect of S$_i$ is not strictly predictable from C's component parts or from other previously established constructions. (Goldberg 1995:4)

The only major difference between this definition and those of symbolic units and phraseologisms discussed above is that a construction as defined here requires non-compositionality or, in Goldberg's terminology, non-predictability while this was not required of symbolic units or phraseologisms.[8] This difference certainly has implications for the nature of the linguistic system postulated but is not a major qualitative difference. Put differently, *symbolic unit* is a general notion, *construction* as defined above is slightly more specific by requiring one non-predictable aspect, and *phraseologism* as defined here is also more specific by not requiring non-predictability, but at least one lexically specified element. It remains obvious, though, that there is again a high degree of compatibility between phraseological research and construction grammarians. In a way, this is not even surprising given that it was prime examples of phraseologisms whose analysis 'gave rise' to Construction Grammar in the first place. As a particularly obvious example, some of the earlier publications that are now understood to have laid the foundations of at least one of the schools of Construction Grammar, viz. the Fillmore-Kay kind of Construction Grammar, took as their starting point elements that would uncontroversially be considered phraseologisms by most scholars; cf. Fillmore et al (1988) on *let alone* or Kay & Fillmore (1999) on the *What's X doing Y?* construction.

Finally, there is another aspect of both Cognitive Grammar and Construction Grammar that is worth pointing out here and will become more relevant below, viz. the importance both theories attach to actual frequencies of usage or occurrence. As mentioned above, Langacker's Cognitive Grammar is explicitly usage-based in the sense that (i) exposure to, and use of, symbolic units, i.e. performance, is assumed to shape the linguistic system of speakers and hearers and (ii) sufficient frequency of occurrence is a necessary condition for entrenchment and, in turn, unit status of a linguistic expression. In this respect, Goldberg's (2006) approach does not differ from Langacker's approach, and while non-compositionality was an additional necessary condition for constructionhood in Goldberg's (1995) Construction Grammar, sufficient frequency was of course also a necessary condition for construction status. Thus, many construction grammarians have made extensive use of the frequency distribution and behavior of constructions in authentic language data (i.e., natural language corpora) in the theoretical literature (cf., e.g., Brenier & Michaelis 2005) and in other domains such as first language acquisition (cf., e.g., Tomasello 2003), language change (cf., e.g., Israel 1996), etc.

By way of an interim summary, we can conclude that, unlike the transformational-generative paradigm, both Cognitive Grammar and Construction Grammar are highly compatible with phraseological research. True, terminologies differ and definitions are not completely identical, but it is easy to see that phraseologisms do not just have a marginal status in these two theories but are rather at the core of what they consider to be their fundamental entities. From this, it of course also follows in turn that phraseological research has a lot to offer to these theories in terms of descriptive work and exploration of the ontological status of phraseological elements. In the opposite direction, phraseological research can benefit from the elaborate theoretical apparatus and the cognitively plausible background provided by Cognitive Grammar and Construction Grammar.

The next section will be concerned with the approach that is probably most intimately connected to phraseological research, viz. corpus linguistics, and we shall see that there is again a high degree of both theoretical and practical overlap, testifying even more to the relevance of phraseological research.

## 3.3 Corpus linguistics

While the two previous sections have been concerned with different linguistic theories (from the opposite ends of virtually all conceivable dimensions), this section will look at the relation of phraseologisms to a methodological paradigm, that of corpus linguistics.

While much of 20th century linguistics was characterized by the strong methodological predominance of acceptability/grammaticality judgments, corpus linguistics as a method has constantly increased in importance in most fields of linguistics, and to my mind at least is currently the single most frequently used method employed in the study of phraseology. This predominance of corpus-linguistic methods within phraseological research is of course not accidental. Corpora as such can only provide

[8] In her most recent work, Goldberg has actually revised her approach such that now non-compositionality is not required anymore provided that the frequency of an expression is large enough for it to become entrenched and, thus, attain construction status (cf. Goldberg 2006:5). This move renders constructions and symbolic units even more similar.

frequency information – frequencies of occurrence and frequencies of co-occurrence.[9] From this, it is a relatively small conceptual leap to the above definition of phraseologism as a co-occurrence phenomenon. As a matter of fact, some of the most central notions in corpus linguistics can be straightforwardly compared to phraseologisms on the basis of the six criteria discussed above. The terms *word clusters*/*n-grams* and *collocations*, for example, refer to frequent co-occurrences of this kind:

- nature of the elements: words;
- number of elements: *n* (usually, that means 'two or more');
- frequency of occurrence: sufficiently frequent to be recognized as a combined element;
- distance of elements: for clusters/*n*-grams, the distance is usually 0 (i.e., the elements are immediately adjacent); for collocations, the distance between the elements involved can vary, but usually exhibits one or a few preferred distances;
- flexibility of the elements: for clusters/*n*-grams, there is usually no flexibility; for collocations, one usually allows for some flexibility: the collocation of *strong* and *tea* would be instantiated both by *strong tea* and *the tea is strong;*
- semantics: *n*-grams are usually retrieved for natural language processing purposes where the issue of non-compositional semantics is only sometimes relevant; for collocations, researchers differ as to whether they require some non-predictable behavior (*strong tea* is acceptable but *powerful tea* is not) or not.

Similarly, the notion of *colligation* is not nowadays usually used as it was originally defined by Firth (1968: 182) – as the co-occurrence of grammatical patterns – but also as a particular kind of phraseologism, namely one in which one or more words habitually co-occur with a grammatical pattern (cf. the example of *to hem*'s preference for passives mentioned at the beginning of this chapter).[10] From these brief remarks about the nature and the number of elements involved, it is clear that much work in corpus linguistics cuts across the boundary of syntax and lexis upheld in formal approaches

[9] By this I of course do not mean that corpus linguistics only provides frequency lists – quite the contrary. What I mean is that whatever information is gleaned from corpora is ultimately based on frequencies. Put differently, a corpus search does not output meaning, or pragmatic intention – it only yields frequency lists, concordances, or collocates, i.e., elements and their frequencies of occurrences or frequencies of co-occurrence of elements of various degrees of abstraction. All linguistic inferences in, say, the domains of morphology, syntax, semantics etc. are thus dependent on statistical information of some kind.

[10] Recently, Hoey (2004) has extended the notion of colligation to also include the co-occurrence of words with, say, particular grammatical slots (e.g., subject, object, complement etc.) and positions within sentences or paragraphs. Such co-occurrences would not qualify as phraseologisms as defined here.

to language,[11] and that there is a considerable overlap between the assumptions made by cognitive linguists, phraseologists, and, as we now see, also corpus linguists.

Another central notion in contemporary corpus linguistics, the *pattern*, involves additional parameters of the above set, viz. the parameter of non-compositionality/non-predictability. This is the definition of a pattern according to Hunston & Francis (2000: 37):

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it.

An expression that would therefore not count as a pattern according to this definition is the adjective *available* followed by spatial prepositions such as *at* or *from* (cf. Hunston & Francis 2000: 72f.) simply because (i) the information provided by the PP headed by these prepositions is straightforwardly and compositionally providing the place where something is available and (ii) the PPs are fairly freely movable within the clause. I would imagine that many, if not most, phraseologists would also not consider *CDs are available at the store* as an instance of a phraseologism. This definition of a pattern fulfils the above six defining parameters and is virtually the same as my definition of phraseologism as well as that of symbolic units in Cognitive Grammar and constructions in Construction Grammar. This testifies strongly to the fact that phraseology is one of the key concepts in both theoretical linguistics and in the method of corpus linguistics, although differing terminology may sometimes render this fact more opaque than is desirable.

In fact, the range of correspondences is even larger. For example, we have seen above that the notion of a (symbolic) unit in Cognitive Grammar involves a degree of automaticity in accessing a structure as well as the absence of the need to analyze the internal structure of a unit. Exactly these notions figure in the formulation of one of the most prominent principles in contemporary corpus linguistics, Sinclair's 'idiom principle'. This principle states that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair 1991: 110) and contrasts with the open-choice principle, which states that "[a]t each point where a unit is completed (a word or a phrase or a clause), a large number of choices opens up and the only restraint is grammaticalness" (Sinclair 1991: 109). In other words, phraseological research quickly leads to evidence for a claim by Pawley & Syder (1983: 213–215ff.), which Langacker (1987: 29–42) has discussed extensively under the heading

[11] In corpus linguistics, this assumption is often attributed to Sinclair (1991). Hunston & Francis (2000) are also often quoted in this connection. However, a wide range of earlier studies make the same assumption, as is obvious from Altenberg & Eeg-Olofsson's (1990: 1) introductory paragraph where several even earlier references building on the very same assumption are cited.

of "rule-list fallacy", that speakers' mental lexicons do contain much more than just lexical primitives, namely hundreds of thousands of prefabricated items that *could* be productively assembled but are, as a result of frequent encounter, redundantly stored and accessed. Thus, the analysis of phraseologisms does not only reveal patterns, and maybe peculiarities, of usage, but can also ultimately lead to more refined statements about matters of mental representation within the linguistic system.

A final parameter which is relevant in this connection and shared across disciplines is one that has just been mentioned, viz. the role that frequency plays for phraseologisms. As mentioned above, a phraseologism is characterized by a sufficiently high frequency of co-occurrence (even when a strict frequency threshold is not provided; cf. Hunston & Francis's definition of a pattern above). A corpus-based approach to phraseologisms and related phenomena such as collocations, colligations, and patterns provides just the frequency data that are needed. Similarly, in cognitive linguistics, the frequency with which a speaker/hearer encounters a particular symbolic unit is assumed to be positively correlated with the degree of cognitive entrenchment of that unit in the speaker's cognitive system, an assumption that has more recently been labeled the 'From-Corpus-to-Cognition Principle' (Schmid 2000: 39).

In sum, Sections 3.1 and 3.2 showed that phraseologisms are relevant not only because they have led some generative linguists to a critical re-evaluation of their framework, but also in the sense that phraseologisms constitute the core of theoretical frameworks that are currently becoming increasingly popular. Section 3.3 has now demonstrated that phraseology and cognitively-inspired linguistic theories overlap almost completely with much of the descriptive and theoretical apparatus of contemporary corpus linguistics. Section 4 will focus on this overlap from a methodological perspective, addressing the question of which strategies different frameworks employ to identify phraseologisms and how these may be improved or refined.

## 4. The identification of phraseologisms

As is obvious from the six parameters discussed in Section 2, there are many different ways of defining the notion of phraseologism. In addition, we have seen in Section 3 that the number of concepts in the semantic space around the words *phraseology* and *phraseologism* is huge. It, therefore, comes as no surprise that a considerable arsenal of empirical approaches has been used to identify phraseological units. It is useful to briefly review these, bearing in mind the role phraseology has played in different theoretical frameworks.

### 4.1 The identification of phraseologisms in transformational-generative linguistics

In transformational-generative linguistics, the identification of phraseologisms has been rather eclectic. Given (i) a linguistic system involving only perfectly productive rules and a lexicon as a grab bag of exceptions and (ii) the objective of developing a language-independent / universal grammar, there has never been a systematic identification of the inventory of phraseologisms in a language within transformational-generative grammar. Indeed, from this perspective, why should there be? Phraseologisms are by most accounts not productive, and thus can be relegated to the role of exceptions; moreover, they are by their very nature not universal and, thus, of little relevance to the core objective of the whole generative enterprise. The lack of a comprehensive identification procedure does therefore not come as a big surprise, and it is probably fair to say that the identification of phraseologisms has been largely based on recognizing that a particular semantic unit's behavior – be it a single- or multi-word unit – defies characterization in terms of the hard-and-fast rules of the grammar that are thought to be necessary on syntactic grounds alone.

### 4.2 The identification of phraseologisms in cognitive linguistics/Construction Grammar

In cognitive linguistics and Construction Grammar, the level of sophistication of the identification and analysis of phraseologism is certainly somewhat higher. First, since the architecture of these approaches does not simply allow their practitioners to simply relegate irregular elements to a peripheral component of the linguistic system, much more energy has been devoted to the identification of phraseological elements in the first place. Second, since both Cognitive Grammar and Construction Grammar are usage-based, their practitioners have relied on authentic corpus data much more than practitioners of many other linguistic theories. This has, of course, necessitated a higher degree of awareness and sophistication regarding the identification and analysis of potential phraseological expressions – as anybody who has ever compared the clarity of invented examples to the messiness of authentic data can confirm. However, it is also interesting to note that, in spite of this recent upsurge of usage-based approaches within cognitive linguistics, the notions of phraseologism, pattern, collocation, and colligation are not particularly frequent and the degree of methodological rigor may sometimes leave something to be desired. For example, Gries et al. (2005, to appear) exemplify this in their discussions of the so-called *as*-predicative – the phraseologism exemplified by *I never saw myself as a costume designer* or *He sees this as a problem* or *Politicians are regarded as being closer to actors* – and show that the most frequent kind of corpus-based quantitative data in Construction Grammar does not always correlate well with experimental results.

### 4.3 The identification of phraseologisms in corpus linguistics

The most comprehensive identification procedures for phraseologisms are certainly found in corpus linguistics. This is to be expected given that corpus linguistics is a methodology that is mostly concerned with lexical (co-)occurrences. Several levels of sophistication are discernible. As in cognitive linguistics and Construction Grammar, the most basic approaches are, it seems, also the most widely used ones.

First, much work in this area, such as that by Stubbs and his colleagues (e.g. Stubbs 2001, 2002), involves the generation of frequency lists of *n*-grams, i.e. uninterrupted sequences of word forms; the upper limit of *n* is usually five.

Secondly, most studies involve the generation of concordances, although the ways in which concordances are generated or processed differ a lot. Some studies just generate a concordance of one or several word forms and interpret the data on the basis of various sorting styles to arrive at generalizations governing the preferred uses of the word forms in question (see, for example, Sinclair 1991:53ff., Hunston & Francis 2000). Other, slightly more complex, approaches generate concordances on the basis of uninterrupted sequences of part-of-speech tags (e.g., Justeson & Katz's (1995a, b) work involving the frames [Adj N] as well as [N N], [N P N], etc.).

Yet other work is based on a mixture of these two approaches. On the one hand, there is what Renouf & Sinclair (1991) call *collocational frameworks*, namely patterns matching [*a* N *of*] or [*be* Adj *to*]. On the other hand, there is work invoking the above notion of colligation. Like all methods employed so far, these approaches are also usually based on frequencies and percentages.

While these methods are no doubt the most widespread ones, they do have some methodological shortcomings. One of the most severe is their limited degree of quantitative sophistication. For example, Stubbs (2002) and Stubbs & Barth (2003) largely ignore the immensely interesting work that has been done on the automatic or semi-automatic identification of multi-word units (cf. below for a range of relevant references). Similarly, Hunston & Francis's (2000:37) above formulation that a combination of words needs to be "relatively frequent" to qualify as a pattern is so vague as to be practically vacuous. Hunston (2002:147) discusses the frequencies of *after a moment*, *after a few moments*, and *after a few moments of*, and then asks "[h]ow many examples of a three-, four, or five-word sequence are necessary for it to be considered a phrase?"

This is all the more regrettable because there is a very large body of research employing sophisticated tools for the identification of phraseologisms.[12] For example, there is a vast array of studies researching how and which collocational statistics improve on the predominant approach of just reporting observed frequencies. Church et al. (1991) is an example of an early study in this vein. More recent examples include Evert and colleagues (e.g. Evert & Krenn 2001) and Gries & Stefanowitsch's work on collostructional analysis, a family of methods concerned with measuring and interpreting the statistical association of words to constructions/patterns (cf. Stefanowitsch & Gries 2003, to appear and Gries & Stefanowitsch 2004) as well as Gries et al. (2005, to appear) for experimental confirmation. In addition, Mason (1997, 1999) and Cantos & Sánchez (2001) discuss a variety of issues concerning the overall validity of collocational studies.

Most of these studies are based on a particular search span or presuppose a particular length for the collocation being investigated. However, the definition of phraseologisms in Section 2 requires decisions about both the length of phraseologisms *and* the different levels of granularity at which co-occurrences can be observed. In addition, a top-down, or *a priori*, approach may not always be the most useful strategy. Sometimes it may be more revealing to let the data – rather than the preconceptions of any particular researcher – decide what the potentially most revealing pattern is. There is a large body of immensely interesting work in this area: for example, Kita et al.'s (1994) cost criterion serves to identify *in a bottom-up manner* the size of interesting uninterrupted multi-word units, which are prime candidates for phraseologisms. Similarly, Mason's (1997, 1999) notion of lexical gravity (cf. Sinclair & Jones 1974 for an earlier though simpler approach) helps to identify the range of collocates – the span – of a word that exhibits interesting distributional patterns. This notion has unfortunately never received the recognition it deserves. Also, the methods proposed by Dias et al (1999), Nagao & Mori (1994), Ikehara et al. (1996), to name but a few, all contain interesting concepts and methodological tools concerning the (semi-)automatic identification of phraseologisms that most corpus-linguistic, let alone cognitive-linguistic, work has not even begun to recognize or utilize to their fullest potential. I hope that the ideas developed in these and similar studies find their way into phraseological research soon and that this chapter, as well as the one specifically addressing this area (cf. Heid this volume), will help promote these approaches.

## 5. Concluding remarks

This chapter has provided a brief and necessarily selective discussion of the notion of phraseologism. I have suggested six parameters that I consider to be the characteristics that every definition of phraseologisms should include. While these parameters lay no claim to originality, I hope that they both underscore the linguistic dimensions that are relevant to phraseology and provide a unified framework of reference for definitions. By choosing different settings of these parameters, it is possible to define a variety of interrelated concepts from different frameworks including, but not limited to, idioms, word-clusters, *n*-grams, collocations, colligations, collostructions, constructions, patterns, fixed expressions and phraseologisms.

While space has precluded a more exhaustive discussion of how all these notions relate to phraseology, I have briefly discussed the role that phraseology plays both in different theoretical schools of thought within linguistics and in the methodological paradigm of corpus linguistics. I have pointed out how the notion of phraseologism –

[12] See. Gries (2006a) for a brief overview of several related problems.

even if labeled differently – plays a role for different linguistic frameworks that can hardly be overestimated, either by providing motivation to critically revise the assumptions of a framework or by integrating seamlessly into, and enriching, the framework. The overlap in terms of the relevant theoretical assumptions and parameters of Cognitive Grammar and Construction Grammar on the one hand and phraseology research and Pattern Grammar on the other is actually so enormous that it is amazing that up to now phraseologists and cognitively inspired linguists have worked on similar issues largely separately. They may not even have recognized or topicalized the overlap, although Schönefeld (1999) and Mukherjee (2004) have made laudable attempts to bridge the gap.

My hope is that the large conceptual overlap of especially corpus-based phraseologists and usage-based cognitive linguists will henceforth be recognized more explicitly and will stimulate a larger amount of cross-disciplinary work. Cognitive linguists as well as construction grammarians have often been relatively lenient and shown little rigor in their handling of frequency data. They, thus, have much to gain from looking at how natural language processing researchers interested in phraseologisms use frequencies and other more elaborated statistics to identify recurrent patterns that are prime candidates for phraseologisms and symbolic units and constructions within cognitive linguistics. In addition, especially some of the early work in Construction Grammar has been concerned with relatively marginal constructions. Phraseologists are probably in an ideal position to provide less specific patterns against which cognitive linguists and construction grammarians can test their analyses. On the other hand, many phraseologists – often with an applied linguistics or lexicographic background – have focused on rather descriptive work on phraseologisms (or, more narrowly, idioms) and have often not been concerned with integrating their accounts of phraseologisms in particular and other patterns more generally into a larger theory of the linguistic system. Given the cognitive linguist's cognitive commitment "to make one's account of human language accord with what is generally known about the mind and the brain, from other disciplines as well as our own" (Lakoff 1990: 40), it seems to me as if this theory would be the one which usage-based phraseologists could work with best. Thus, if only a few of these gaps between these different fields are bridged, this chapter has served its purpose.

## Acknowledgements

I am grateful to Sylviane Granger and especially Stefanie Wulff for feedback and discussion. Two anonymous reviewers provided a wealth of comments that made me clarify many issues considerably. I also owe much to Anatol Stefanowitsch for extensive cooperation and thus much input and discussion in the overall subject area over the past few years. The usual disclaimers apply.

## References

Altenberg, B. & M. Eeg-Olofsson (1990). Phraseology in spoken English: Presentation of a project. In Aarts, J. & W. Meijs (eds.) *Theory and Practice in Corpus Linguistics*, 1–27. Amsterdam: Rodopi.

Brenier, J. M. & L. A. Michaelis (2005). Optimization via syntactic amalgam: Syntax-prosody mismatch and copula doubling. *Corpus Linguistics and Linguistic Theory* 1(1): 45–88.

Bybee, J. (1985). *Morphology: A Study into the Relation of Meaning and Form*. Amsterdam: John Benjamins.

Cantos, P. & A. Sánchez (2001). Lexical constellations: What collocates fail to tell. *International Journal of Corpus Linguistics* 6(2): 199–228.

Church, K. W., W. Gale, P. Hanks & D. Hindle (1991). Using statistics in lexical analysis. In Zernik, U. (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115–164. Hillsdale, NJ: Lawrence Erlbaum.

*Cobuild on Compact Disc* V1.2 (1995). HarperCollins.

Cowie, A. P. & R. Mackin (1993). *Oxford Dictionary of Phrasal Verbs*. Oxford: Oxford University Press.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Culicover, P. W. (1999). *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford: Oxford University Press.

Dias, G., J. G. Pereira Lopes & S. Guilloré (1999). Mutual expectation: A measure for multiword lexical unit extraction. In *VEXTA: Proceedings of the Conference, 22–24 November 1999, Università Ca' Foscari, San Servolo, VIU, Venezia, Italia*, 133–138. Padova: Unipress.

Evert, S. & B. Krenn (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188–195. Morristown, NJ: Association for Computational Linguistics.

Fillmore, C. J., P. Kay & M. C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3): 501–538.

Firth, J. R. (1968). *Selected Papers of J. R. Firth 1952–59*. F. R. Palmer (ed.). London: Longman.

Fraser, B. (1966). Some remarks on the VPC in English. In Dinneen, F. P. (ed.) *Problems in Semantics, History of Linguistics, Linguistics and English*, 45–61. Washington, DC: Georgetown University Press.

Fraser, B. (1976). *The Verb-Particle Combination in English*. New York, NY: Academic Press.

Gibbs, R. W. Jr. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge: Cambridge University Press.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: The University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, S. T. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum Press.

Gries, S. T. (2006a). Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 191–202.

Gries, S. T. (2006b). Exploring variability within and between corpora: Some methodological considerations. *Corpora* 1(2): 109–51.

Gries, S. T. (to appear). Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions?

Gries, S. T., B. Hampe & D. Schönefeld (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4): 635–676.

Gries, S. T., B. Hampe, & D. Schönefeld (to appear). Converging evidence II: More on the association of verbs and constructions. In Newman, J. & S. Rice (eds.) *Experimental and Empirical Methods in the Study of Conceptual Structure, Discourse, and Language*. Stanford, CA: CSLI.

Gries, S. T. & A. Stefanowitsch (2004). Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129.

Hoey, M. (2004). Textual colligation: A special kind of lexical priming. *Language and Computers* 49(1): 71–94.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics* 19(1): 24–44.

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

Hunston, S. & G. Francis (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Ikehara, S., S. Shirai, & H. Uchino (1996). A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of the 16th Conference on Computational Linguistics – Volume 1*, 574–579. Morristown, NJ: Association for Computational Linguistics.

Israel, M. (1996). The *way* constructions grow. In Goldberg, A. E. (ed.) *Conceptual Structure, Discourse, and Language*, 217–230. Stanford, CA: CSLI.

Jackendoff, R. S. (1997). Twistin' the night away. *Language* 73(3): 534–559.

Justeson, J. S. & S. M. Katz (1995a). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1: 9–27.

Justeson, J. S. & S. M. Katz (1995b). Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics* 21(1): 1–27.

Kay, P. & C. J. Fillmore (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language* 75(1): 1–33.

Kita, K., Y. Kato, T. Omoto & Y. Yano (1994). A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing* 1(1): 21–33.

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics* 1(1): 39–74.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar: Descriptive Application*. Stanford, CA: Stanford University Press.

Mason, O. (1997). The weight of words: An investigation of lexical gravity. In Lewandowska-Tomaszczyk, B. & P. J. Melia (eds.) *PALC'97: Practical Applications in Language Corpora*, 361–75. Lodz: Lodz University.

Mason, O. (1999). Parameters of collocation: The word in the centre of gravity. In Kirk, J. (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, 267–280. Amsterdam: Rodopi.

Mukherjee, J. (2004). Corpus data in a usage-based cognitive grammar. In Aijmer, K. & B. Altenberg (eds.) *The Theory and Use of Corpora: Papers from the 23rd ICAME Conference*, 85–100. Amsterdam: Rodopi.

Nagao, M. & S. Mori (1994). A new method of *n*-gram statistics for large number of *n* and automatic extraction of words and phrases from large text data of Japanese. *COLING* 14 Vol.1, 611–615. San Mateo, CA: Morgan Kaufmann.

Nunberg, G., I. A. Sag, & T. Wasow (1994). Idioms. *Language* 70(3): 491–538.

Pawley, A. & F.H. Syder (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Richards, J.C. & R. W. Schmidt (eds.) *Language and Communication*, 191–226. London: Longman.

R Development Core Team (2006). *R: A language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Renouf, A. & J. M. Sinclair (1991). Collocational frameworks in English. In Aijmer, K. & B. Altenberg (eds.) *English Corpus Linguistics*, 128–143. London: Longman.

Rice, S. & J. Newman (2005). Inflectional islands. Paper presented at the International Cognitive Linguistics Conference, Seoul, South Korea.

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Mouton de Gruyter.

Schönefeld, D. (1999). Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics* 4(1): 131–71.

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

Sinclair, J. M. & S. Jones (1974). English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24(2): 15–61.

Stefanowitsch, A. & S. T. Gries (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–43.

Stefanowitsch, A. & S.T. Gries (to appear). Corpora and grammar. In Lüdeling, A. & M. Kytö (eds.) *HSK Korpuslinguistik – Corpus Linguistics*. Berlin: Mouton de Gruyter.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2): 215–44.

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

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Wulff, S. (to appear). *Rethinking Idiomaticity: A Usage-Based Approach*. London: Continuum Press.