| aptara | LANG | lang_534 | **Dispatch:** 9-9-2009 | **CE:** N/A |
|---|---|---|---|---|
| | **Journal** | **MSP No.** | **No. of pages:** 20 | **PE:** Matthew |

**Language Learning ISSN 0023-8333**

# A Usage-Based Account of Constituency and Reanalysis

Clay Beckner
University of New Mexico

Joan Bybee
University of New Mexico

Constituent structure is considered to be the very foundation of linguistic competence and often considered to be innate, yet we show here that it is derivable from the domain-general processes of chunking and categorization. Using modern and diachronic corpus data, we show that the facts support a view of constituent structure as gradient (as would follow from its source in chunking and categorization) and subject to gradual changes over time. Usage factors (i.e., repetition) and semantic factors both influence chunking and categorization and, therefore, influence constituent structure. We take as our example the complex prepositions of English, for instance, *on top of*, *in back of*, and *in spite of*, whose internal constituent structure has been much debated. From observing strong (but not absolute) usage trends in the corpus data, we find that these complex preposition sequences display varying degrees of emerging constituency. We conclude that constituent reanalysis, like language change generally, proceeds gradually.

## Introduction

Most theories of language take the categories of grammar and their hierarchical relations (i.e., constituent structures) as givens. Constituent structure, such as might be formalized with phrase structure rules or syntactic trees, typically takes a prominent place as a substantive universal in theories of Universal Grammar and is thus held to be innate and domain-specific (Chomsky, 1965; Jackendoff, 2002). We take the view, in contrast, that no part of grammar needs to be given a priori (Hopper, 1987); rather we follow Lindblom, MacNeilage,

Correspondence concerning this article should be addressed to Joan Bybee, University of New Mexico. Internet: jbybee@unm.edu Q1

and Studdert-Kennedy (1984) in their plea for explanations for linguistic structures and universals. They specifically incite us to "DERIVE LANGUAGE FROM NON-LANGUAGE!" (emphasis in the original). Thus, in this article we propose to derive constituent structure from the domain-general processes of chunking and categorization within the storage network for language. Because language is a dynamic system, an important part of our argument will rest on the idea that constituent structure, like all of grammar, is constantly undergoing gradual change. Thus, structural reanalysis, as often discussed in the context of grammaticalization, will be pivotal to our argument and exposition.

We mean by *structural reanalysis* a change in constituent structure, as when *to* as an earlier allative or infinitive marker with a verb as its complement fuses with *going* in the future expression *be going to* (*going* [*to see*] > [*going to*] *see*). Indicators of reanalysis include changes in distribution, such as the fact that selectional restrictions in a clause with *be going to* are determined by what is now the main verb, and phonological changes, such as the reduction of *going to* to *gonna*.

Are such changes abrupt or gradual? In generative models of syntax (see, e.g., Lightfoot, 1979; Roberts & Roussou, 2003), structural reanalysis is necessarily abrupt, because it is held that a sequence of words has a unique, discrete constituent analysis.[1] In this view, constituents are clearly defined and do not overlap; in a sequence such as *going to VERB*, *to* must be grouped either with the following verb, or with *going*, with no intermediate stages. The only way for discrete constituent boundaries to shift is via abrupt means—specifically, via the mechanism of language acquisition, when children misinterpret the constituents they hear in adult language and assign a different structural analysis than the previous generation.

However, because most linguistic change appears to be quite gradual, with slowly changing meanings and distributions and overlapping stages, a problem arises for a theory with discrete constituent structure. Evidence from the gradualness of change has led some researchers to doubt discrete categories and structures (Haspelmath, 1998; Hoffmann, 2005; Quirk, Greenbaum, Leech, & Svartvik, 1985).

Continuing from Bybee and Scheibman (1999), we join these researchers in proposing that constituent structure can change gradually. We take the view that it is altogether common even for an individual speaker to have nondiscrete syntactic representations for the same word sequence. Taking a complex systems-based perspective, we hold that syntactic structure is in fact much richer than the discrete constituency view would indicate. There are multiple overlapping and, at times, competing influences on the shape of units in the grammar, and these

multiple factors have an ongoing effect on each speaker's synchronic representations of syntactic structure. Specifically, syntactic constituents are subject to ongoing influence from general, abstract patterns in language, in addition to more localized, item-specific usage patterns. The foregoing perspective makes it possible that the same word sequence may be characterized by multiple constituent structures and that these structures have gradient strengths rather than discrete boundaries. Our position in this article is thus that constituency may change in a gradual fashion via usage, rather than via acquisition, and that structural reanalysis need not be abrupt.

As a case study of shifting constituent boundaries, we focus on the semantic and syntactic analysis of English complex prepositions (i.e., multiword sequences that function prepositionally, such as *on top of* or *in spite of*). Complex prepositions often may be replaced by a single word (a preposition), such as *The car is in back of the house/The car is behind the house*. This replaceability hints that certain complex sequences have formed (or have started to form) into new prepositions (see Quirk et al., 1985; Quirk & Mulholland, 1964). Complex prepositions are also quite often unpredictable in meaning and, as such, are taught as noncompositional units to second language learners. For example, the English Preposition-Preposition sequence *out of* has multiple meanings that cannot be predicted from the component words, even accounting for the fact that the component prepositions are themselves polysemous: *We are out of milk again*; *I've been out of town*; *The storm came out of the west*; *They made a decision out of desperation*.

A large number of English sequences (among them, *because of, according to, by dint of, due to*) exhibit the above traits and/or other syntactic characteristics that imply that they are constituents (Quirk et al., 1985, pp. 669–673). Despite such evidence, the syntactic status of complex prepositions has been the matter of some debate. We enter this debate in this articler from the viewpoint that constituent structure is gradient, mutable, and emergent from domain-general processes. In the next section, we describe the way chunking and categorization together provide constituency analyses of phrases and utterances for speakers. In the third section, we describe how the model proposed in the second section accounts for reanalysis, using the case of the complex preposition *in spite of* whose development in terms of both meaning and syntax is discussed. In the fourth section, we respond to objections by critics who argue against assigning constituent status to complex prepositions, based on the discrete constituency view. We argue that our view, which references meaning as well as gradual change in cohesiveness and autonomy, provides a better explanation for the problems raised by the analysis of complex prepositions.

## Constituent Structure as Emergent From Chunking, Categorization, and Generalization

Bybee (2002, in press) discusses the nature of sequential learning and chunking as it applies to the formation of constituents. Because members of the same constituent appear in a linear sequence with some frequency, these items are subject to chunking, by which sequences of repeated behavior come to be stored and processed as a single unit. Ellis (1996) gave the following quote from Newell (1990), which emphasizes the domain-general application of chunking:

> A chunk is a unit of memory organization, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit. Chunking implies the ability to build up such structures recursively, thus leading to a hierarchical organization of memory. Chunking appears to be a ubiquitous feature of human memory.[2] (p. 7)

Chunking occurs automatically as behaviors are repeated in the same order, whether they are motor activities such as driving a car or cognitive tasks such as memorizing a list. Repetition is the factor that leads to chunking, and chunking is the response that allows repeated behaviors to be accessed more quickly and produced more efficiently (Haiman, 1994). Chunking has been shown to be subject to The Power Law of Practice (Anderson, 1993), which stipulates that performance improves with practice, but the amount of improvement decreases as a function of increasing practice or frequency. Thus, once chunking occurs after several repetitions, further benefits or effects of repetition accrue much more slowly.

Chunked elements in language are oft-repeated sequences such as determiner plus noun, preposition plus noun phrase, verb plus object, and so on. Specific lexemes that are used together, as in formulas or prefabs (e.g, *dark night, salt and pepper,* or *take a break*), also constitute chunks. The formation of chunks produces hierarchical structure in language, as smaller chunks will be more frequent, will have undergone more practice, and will therefore be more cohesive than larger ones. As smaller chunks appear within larger ones, a nested structure emerges.

Chunking is also responsible for the fact that some sequences of linguistic units show formal cohesion in the absence of semantic cohesion. Bybee (2002) gave as an example auxiliary contraction in English. Whereas most chunks have some semantic coherence, the English auxiliary is chunked with the subject, usually a pronoun (e.g., *I'm*), resulting in a formal unit that crosses a traditional constituent boundary (between NP and VP) and that does not

result in a semantically coherent unit. Another example is the common fusion of prepositions with definite articles in French, Spanish, German, and other European languages. However, because elements that are semantically related tend to occur together, most chunks are also semantically coherent and therefore considered to be constituents in most theories of grammar.

The second domain-general process that contributes to the formation of constituent structure is categorization. We propose conceiving of cognitive representations as a network of exemplars that undergoes change as language is used. An incoming token of linguistic experience, such as a word, is mapped onto an identical or similar stored exemplar, strengthening it. For the purposes of this articler we will assume familiarity with exemplar models and not provide the evidence and arguments for them here (but see Bybee, 2001, 2006, in press; Pierrehumbert, 2001). Instead, we will concentrate on the relationship between chunking and categorization.

Mapping experienced tokens onto stored exemplars is an act of categorization. For instance, deciding that *pull* in the idiom *pull strings* is the same verb as that occurring in other expressions (e.g., *pull the trigger*, *pull someone's leg*) is an act of categorization. It is based on phonetic and semantic similarity as well as morpho-syntactic distribution. In the network model first proposed in Bybee (1985), the categorization by similarity would be represented as in Figure 1.

In Figure 1, the sequence *pull strings* is represented as a consecutive string because the two words have been used together enough to constitute a chunk. It might also be argued on the basis of idiomaticity alone that *pull strings* has unitary status, but note that even semantically compositional sequences can become chunks as a result of usage. For example, compositional sequences such as *for some reason* and *dark night* represent the conventional way of expressing certain notions, in contrast with semantically plausible (but unlikely)

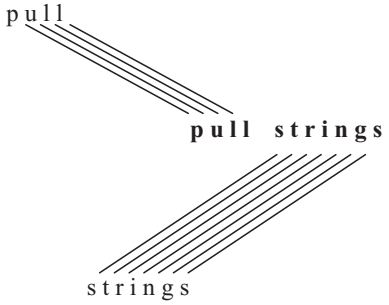


**Figure 1** The connections between an idiom and its component words (Bybee, 1998).

sequences like *from some reason* or *black night.* A fully realized exemplar model proposes that each of the conventionalized sequences has an independent mental representation: The conventionality of *pull strings*, *for some reason*, and *dark night* arises because our cognitive systems track the usage of these specific sequences in language.

However, even when a sequence of words is chunked together, the components of the chunk may remain identifiable in both form and meaning (Nunberg, Sag, & Wasow, 1994). As shown in Figure 1, the representation for *pull strings* maintains connections to other instances of the component words *pull* and *strings*. In Langacker's (1987) terms, the chunk is analyzable. Note that the categorization of the parts of the chunk provides an internal constituent structure. In our view, categorization within a network architecture is the mechanism that creates abstract syntactic patterns—those regularities that would be represented via phrase structure rules in a generative model. Because certain words have similar distributions and may be categorized together, generalizations emerge across recurrent categories of items, resulting in abstract constituent patterns. For instance, *the charming dog* is a constituent (labeled NP by linguists) because it fits into a general pattern in which phrases may consist of Determiner plus Adjective plus Noun.[3]

In sum, we find that the constituency of a sequence of words is best characterized by appealing to both "local" (item-specific) and "global" (general, type-based) influences. Local influences can chunk a specific, recurrent word sequence into a constituent—a constituent that nevertheless maintains a limited internal structure due to the way component words are categorized following global patterns. On the other hand, constituency is often determined largely via global influences, as words in a sequence are each categorized following general patterns and chunked together according to a recurrent type-based generalization. A complete syntactic model will recognize that local and global influences may oppose one another and that in different cases they will affect constituency to varying degrees. Given these multiple factors, chunking and categorizability are *gradient* properties of sequences and they may change over time with usage.

## Changes in Constituent Structure

In grammaticalization it often happens that grammaticalizing expressions change their constituent structure. Thus, it is often said that grammaticalization is the reanalysis of a lexical item as a grammatical item. As Haspelmath (1998) pointed out, often the change can be thought of as a simple change in a category

status. Thus, a verb becomes an auxiliary; a serial verb becomes an adposition or a complementizer; a noun becomes a conjunction or adposition. In some cases, however, shifts in constituent boundaries do occur; in particular, it is common to lose some internal constituent boundaries. A prominent example of such a change involves complex prepositions. Many complex prepositions start out as a sequence of two prepositional phrases (e.g., *on top of* NP) but evolve into a kind of intermediate structure in some analyses—the complex preposition—and eventually they can even develop further into simple prepositions, as has occurred with *beside*, *behind, and among* (Hopper & Traugott, 2003; König & Kortmann, 1991; Svorou, 1994).

### *In Spite of*: from P NP P to Prepositional Unit

A typical example of a complex preposition in English, *in spite of,* was originally constituted of a preposition *in,* whose object was the noun phrase headed by *spite.* A traditional phrase structure analysis of the earlier, analyzable sequence has a nested structure such as the following:

(1) $[\text{in } [\text{spite } [\text{of } [\text{the king}]_{\text{NP}}]_{\text{PP}}]_{\text{NP}}]_{\text{PP}}$

Basically, the starting point for reanalysis is a hierarchical constituent structure in which *spite* is an ordinary noun meaning "defiance, contempt, scorn," and there are two prepositional phrases with *in* and *of* . Note, however, that the most frequently recurring part of the structure is *in spite of* , as the object of *of* is highly variable and the rest of the expression is fixed. This means that *in spite of* can become a chunk.

The hierarchical analysis, as in (1), will remain only as long as the phrase remains analyzable—that is, as long as *spite* within the phrase continues to be categorized as a noun and as the same item as the noun *spite* that occurs in other expressions and as long as the prepositions are associated with other instances of these same prepositions. Because not much phonetic change is observed in this phrase, two factors are important to the change in analyzability. One is the effect of frequency of use, which leads to the access of the phrase as a unit; as Hay (2001) pointed out, each time the sequence is processed as a unit that increases its sequential cohesion. The second factor in reducing analyzability is semantic change, which, of course, interacts with frequency of use; the semantic change weakens the association of the noun *spite* with its lexical counterparts, leading to a loss of analyzability and also categoriality (Hopper, 1991; Hopper & Traugott, 2003). As the noun *spite* within the phrase becomes disassociated from the independent noun, it loses its nounlike behavior—that is, it ceases to take determiners or modifiers.[4]

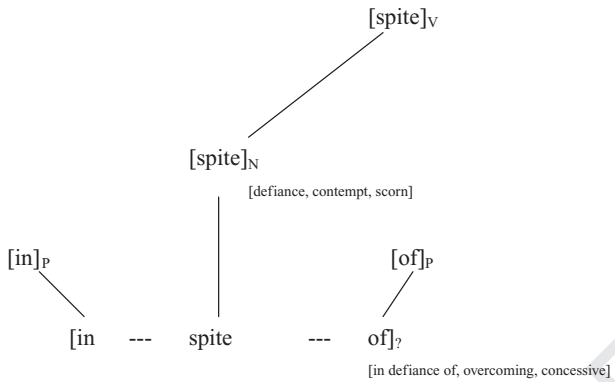


**Figure 2** Exemplar representation of *in spite of* and some of its lexical connections.

Figure 2 helps us visualize how gradual reanalysis can be modeled over time. A number of morphological models have been proposed in which morphologically complex words can be accessed in two ways: either directly, already composed, or from the component parts via a compositional mechanism. Some of these models (Baayen, 1993; Hay, 2001) propose to consider any accessing event as a combination of the two mechanisms in which one or the other may be prominent. As this is not an either-or situation, the extent to which the component parts are activated may vary. When each part is accessed and then combined, the connecting lines to the parts are strengthened. When the multiword sequence is accessed without activating the parts, the whole sequence is strengthened. Thus, given the network in Figure 2, over accessing events, the vertical connection lines (indicating categorization of the individual words) become relatively weaker while the sequential connections (indicating the formation of a multiword chunk) become relatively stronger.

The noun *spite* has a set of meanings and contexts of use; the phrase *in spite of* as an exemplar develops its own meanings and contexts of use. As the phrase becomes less associated with its component parts, it also becomes more autonomous pragmatically and semantically and begins to take on meanings inferred from the context, such as concessive meaning. Moreover, the increasing autonomy and fixedness of *in spite of* develop in tandem with a particular syntactic distribution, which is essentially the distribution of a preposition; that is, *in spite of* occurs as a chunk in similar environments as other English prepositions: We may say [*in spite of*] *resistance* or [*without*] *resistance* (see additional discussion in the fourth section). With respect to the formation of global, category-based generalizations, such a distributional pattern

would then gradually encourage classification of *in spite of* as a prepositional unit.

Given the above proposed model, Hay (2001) reasoned that if the complex unit is more frequent than its parts, it is more likely to be accessed as a unit, leading to the loss of analyzability that comes about through categorization. Applied to *in spite of*, we would predict that as the complex phrase becomes more frequent than the simple noun *spite*, it would also become more autonomous and less analyzable. In Shakespeare's comedies (written at the end of the 16th century) we find 20 occurrences of *spite*; only 6 of them are in the phrase *in spite of*. In Modern American English, over 90% of the occurrences of *spite* are in that phrase (see the Corpus of Contemporary American English [COCA], Davies, 2008).

Consider also some of the examples from Shakespeare's usage. In the next section we will discuss semantic change in more detail, but note here that in both (2) and (3) the meaning of *in spite of* invokes *spite* in its original meaning of defiance. Note also that in (2) Beatrice uses *spite* as a verb after using it in the phrase *in spite of*. This suggests analyzability of the phrase. In (3) Ulysses interrupts the phrase with a modifier, *very*, which is here used as an adjective meaning "true." The added modifier gives evidence that *spite* is being categorized as a noun, and the sequence is analyzable; such uses are very rare today.

(2) (*Much Ado About Nothing*):

> BENEDICK: Suffer love! a good epithet! I do suffer love
> indeed, for I love thee against my will.
>
> BEATRICE *In spite of* your heart, I think; alas, poor heart!
> If you spite it for my sake, I will spite it for
> yours; for I will never love that which my friend hates.

(3) (*Troilus & Cressida*):

> ULYSSES: Ajax hath lost a friend
> And foams at mouth, and he is arm'd and at it,
> Roaring for Troilus, who hath done to-day
> Mad and fantastic execution,
> Engaging and redeeming of himself
> With such a careless force and forceless care
> As if that luck, in very *spite of* cunning,
> Bade him win all.

In the next section, we look briefly at the gradual progression of the semantic change.

**The Semantic Development of *in Spite of***

Paralleling the morpho-syntactic and usage changes we have documented, developing complex prepositions undergo semantic change typical of grammaticalization. Hoffmann (2005) showed that each complex preposition follows its own trajectory and pointed out further that a strict chronology may not be possible, given the paucity of surviving early examples. Here, we focus on *in spite of* and sketch the general picture of the emergence of concessive meaning for that complex preposition.

*In spite of* appears to have been first used with a literal interpretation of the noun *spite*, which meant "scorn, contempt, or defiance." The most literal of uses are those that indicate an explicit defiance of an enemy, as in the following 15th century examples:

(4) c1400 *Destr. Troy* 1968. But for noy of my nobilte & my nome gret, I shuld..spede the to spille *in spite of* þi kynge.

If it were not for the risk to my nobility and my reputation, I would hasten to kill you in spite of your king. (Translation from Hoffmann, 2005)

(5) 1400–1482 *The Brut* The Erle en, with his pepill, drove ouer e havon of Gravenyng thaire pray of bestes, att lowe water, *in spite of* al e Flemmynges, and brought hem with al thaire prisoners to Caleis, and lost neuer a man; thonket be God!

Then the Earl, with his people, drove over the inlet at Gravening their herd of animals, at low water, in spite of the Flemish, and them with all their prisoners to Calais, and never lost a man; thanks be to God!

Later examples show a generalization of the object of *in spite of* to include obstacles of various sorts—for instance, authority figures, rules of law, or culture—as shown in the following examples taken from the *Oxford English Dictionary*, spanning the 16th to the 19th centuries.

(6) 1581 G. PETTIE tr. *Guazzo's Civ. Conv.* III. (1586) 129b, The wife *in spight of* the husband, gave halfe the meate . . . to a poore bodie.

(7) 1617 MORYSON *Itin.* I. 232 They . . said, that the Scripture must be beleeved, *in spite of* all Cosmographers and Philosophers.

(8) 1711 E. WARD *Quix.* I. 158 Who would *in Spite of* Wedlock Run To Cuddle with the Emp'rour's Son.

(9) 1853 KINGSLEY *Misc.* (1859) I. 15 The English are attacked treacherously *in spite of* solemn compacts.

At about the same time, examples appear in which the opposing force is the effort of someone, which, alas, is not successful. In some cases, the efforts are exerted by the same person who puts forth or undergoes the main action. Examples (10) and (11) show that such usage continues; the expression *in spite of oneself* is still in use today.

(10) 1765 *Museum Rust.* IV. 266 They grow poor, *in spite of* all possible industry.

(11) 1818 SCOTT *Br. Lamm.* xx, The tears, *in spite of* her, forced their way between her fingers.

Example (12) is also an instance in which the object of *in spite of* is someone's effort, but in this case, it is the effort of another actor.

(12) 1782 COWPER *Gilpin* xxii, That trot became a gallop soon *in spite of* curb and rein.

All of these examples carry a discourse-based inference of counterexpectation, which is the seed of the concessive meaning. The object of *in spite of* expresses an obstacle that is overcome or not overcome in the physical and social world, so it also sets up the expectation that the situation expressed in the clause is not to be expected. As uses with this inference of counterexpectation become more common, the concessive inference can become part of the meaning of the phrase (Traugott & Dasher, 2002). This leads to examples that are ambiguous between a reading in which the speaker/writer is describing counterforces in the real world and a reading in which the speaker/writer is expressing counterexpectation. Example (13) seems ambiguous, as does (14).

(13) 1859 *Bentley's Q. Rev.* No. 3. 26 *In spite of* this aimlessness the wealth and empire of England are constantly increasing.

(14) *In spite of* the rough conditions, travel advisories and the war on terrorism, scores of older Americans are uprooting their lives to help needy nations improve their living conditions. (*Time* Magazine Corpus, [Davies, 2007], 2003)

In the final development, tokens in which only the concessive meaning of counter expectation is discernible arise, as in (15) and (16):

(15) Yet *in spite of* music's remarkable influence on the human psyche, scientists have spent little time attempting to understand why it possesses such potency. (*Time* Corpus, 2000)

(16) The condition of accelerated puberty in girls is more of a hypothesis than a widely observed phenomenon—*in spite of* anecdotal reports. (*Time* Corpus, 2000)

Along with these purely concessive meanings—in which the *in spite of* phrase simply marks a counter-to-expectation condition—some of the older uses continue. Example (17) shows the older, more literal meaning of overcoming opposing forces.

(17) I saw the pictures of the Iraqi people walking to the polls to exercise their right to vote in the face of death threats, bombs and with entire families in jeopardy. To vote *in spite of* all that takes courage above and beyond what most Americans would show today. The Iraqis expressed the true spirit of democracy. (*Time* Corpus, 2005)

Although it is difficult to establish a reliable chronology, due to the paucity of early examples, the indications are that the true concessive use has only become common recently, perhaps in the last century and a half. It is thus important to note that the uses of *in spite of* do not change abruptly nor does one use replace another. A range of uses is maintained in the current language. However, we take the emergence of concessive meaning as a sure indicator of unithood for the phrase *in spite of* . As the phrase develops a weakened association with the semantics of *spite*, not coincidentally, the internal constituent structure of the phrase also weakens. We consider a study of the semantic change necessary for determining when a change in constituency occurs, as a sequence may begin to be extended to new semantic contexts only when loss of analyzability has occurred. In addition, we believe that the semantic changes are also necessary for understanding *why* reanalysis takes place, as no change occurs in isolation, but in a particular semantic-pragmatic context.

In the next section we turn to the traditional morpho-syntactic diagnostics of constituency, which we argue are epi-phenomena (Hopper, 1987), as the most basic determinants of constituency are usage and cognitive association of the phrase with its component parts in their other uses.

## Syntactic Diagnostics and Usage Data in Identifying Constituents

As we have noted, several traditionally oriented analysts have objected that it is incorrect to assign constituent status to *in spite of* and other complex prepositions that have been proposed for English (see Huddleston & Pullum, 2002; Pullum, 2006; Seppänen, Bowen, & Trotta, 1994). In this section, we

briefly characterize the nature of such objections, with a particular focus on *in spite of*.[5] We provide a response from our gradient constituency perspective and provide some usage data to contrast with introspective diagnostics.

First, we note that in traditional discussions of constituent status, there is a tendency to favor evidence based on introspective syntactic tests to the exclusion of any other types of evidence. Thus, in Seppänen et al. (1994, p. 4) and Huddleston and Pullum (2002, p. 621), semantics is explicitly rejected as a factor in determining constituency, on the assumption that syntax provides a more systematic and rigorous testing ground. Our view, however, is that the most thorough assessment of constituency will consider *all* evidence (semantic, pragmatic, morphosyntactic, and phonetic). As we will see in this section, even the syntactic criteria do not all uniformly point in the same direction.

Huddleston and Pullum (2002, p. 620) wrote that some multiword sequences have a "close semantic relation" to single-word prepositional items, such as *in front of/behind, on top of/underneath*, and *in spite of/despite*. These semantic similarities also have correspondences in (purportedly more rigorous) syntactic constituency tests. Syntactic tests such as the "Coordination" test hint that single-word and multiword prepositional forms in fact have similar syntactic distributions and seem to constitute similar types of units:

(18) Scorsese's strongest works are fictions of formation, in which a religious conviction comes *with* or *in spite of* a vocation. (COCA, 1991)

More importantly, it turns out that syntactic criteria give conflicting results. For instance, taking *in spite of,* it is found that *of* cannot be fronted in the constructed example **Of what obstacles did he say he would do it in spite?* (Seppänen et al., 1994). This would suggest that *in spite of* is a unit. Seppänen et al. argued, however, that the sequence is not a constituent because it fails tests for Coordination and Interpolation (Interruption).[6] With respect to Coordination, it is indeed the case that users of English sometimes coordinate *in spite of* in a way that indicates an awareness of internal structure for this sequence. In the 360-million-word COCA, we located 7 instances in which writers conjoined *in spite of* with other *of* sequences in the following pattern:

(19) The prime minister remains unable to reap the credit for economic success, which is perceived to have occurred *in spite*, not *because*, *of* his policies . . . (COCA, 1995)

(20) . . . a lesson in how Congress makes politically expedient decisions *at the expense* (or *in spite*) *of* the constitutional implications of their actions (COCA, 2002)

It is perhaps surprising that writers of English would conjoin *in spite of* in a way that reveals awareness of the individual status of *of*. Yet our position predicts that a word sequence may gradually form a unitary status even while component words are partially activated on each use. To return to Figure 2, note that even as *in spite of* strengthens in constituency, it does not instantaneously become fused into an indivisible unit. The sequence continues to maintain some connections to the separate words *in*, *spite*, and (most importantly here) *of*.

If we look at the full range of usage data, it is in fact unquestionable that *in spite of* has a mostly fixed status, and this fixedness must be acknowledged by a complete theory of constituency. Despite the occurrences of sentences like (20) and (21), it is far more common for English speakers to avoid splitting up *in spite of*, even when they could easily do so. In the COCA, we located 35 such instances. Two of these examples are as follows:

(21) . . . the dogma of self-expression says that the gifted child can flower *in the absence of* or *in spite of* art education. (COCA, 1995)

(22) . . . in this allegedly anti-American country Sarkozy would be elected (as early as the spring of 2007) either *because of* or *in spite of* the public perception that he is somehow "American." (COCA, 2005)

Even more striking are usage patterns with respect to multiple instances of *in spite of* that are conjoined. English speakers strongly prefer to present multiple instances of *in spite of* as an uninterrupted sequence; (23) is one characteristic example:

(23) *In spite of* motorbikes, *in spite of* karaoke music, *in spite of* the stink of gasoline fumes that seeps into each kitchen. (COCA, 2005)

There are 43 examples of this type in COCA. The corpus does contain two counterexamples in which only subparts of *in spite of* are conjoined, but neither instance occurs in very recent usage.[7] Given a traditional syntactic analysis, we might expect speakers to separate *in spite* and *of* in conjoined uses, assuming that a constituent boundary exists at that juncture. Instead, what we find is that speakers typically repeat this entire three-word sequence without interruption, providing evidence that *in spite of* is produced in a single, formulaic chunk (see Wray, 2006).

Of course, we must also consider the possibility that the unit status of a complex preposition can be questioned if interruptions are permitted by other words (for instance by hesitations or discourse markers). Seppänen et al. (1994, p. 22) pursues this line of thought also, arguing that *in spite of* retains an internal

constituent structure because it can be interpolated in speech. The constructed example they provided is *The morning air was clear and clean, in spite, one might add, of the traffic and crowds*.

In response, we note first that interpolation is not very reliable as a test of constituency, because discourse markers, hesitations, and parenthetical asides may be inserted into speech in many positions, including in the middle of traditional constituents (e.g., into a VP in *It is, however, a profitable company*; McCawley, 1982; see also Hoffmann, 2005, p. 34).

Moreover, with respect to *in spite of* , notwithstanding the constructed example by Seppänen et al. (1994), it seems that speakers very seldom interpolate any material into this sequence. Our corpus search yielded only one attested example, which occurred in a Robert Ingersoll quote from 1877:

(24) The religionists of our time are occupying about the same ground occupied by heretics and infidels of one hundred years ago. The church has advanced *in spite*, as it were, *of* itself. (COCA, 1999)

Example (24) is striking because Ingersoll interrupts *in spite of* precisely for the purpose of calling attention to the component words in the sequence, as he intends to revive the original semantics of *spite.*

Thus, although we concede that it may be possible for speakers to add asides to *in spite of* , it is worth noting how truly rare such interruptions are. Of particular interest in *in spite of* is the transition between *spite* and *of,* because Seppänen et al. (1994) focused on the constituent boundary, which they maintained remains active in that juncture. In the COCA, we find 6254 tokens of *in spite*. Out of these instances, 6241 are *also* tokens of *in spite of* . What this means is that in the corpus data, the transitional probability between *in spite* and *of* (the likelihood of following *in spite* with *of* ) is 99.5. We find that this fact provides overwhelming evidence that *in spite of* constitutes a single constituent, with only a very weak association of *of* with this preposition elsewhere. We claim that it is difficult to maintain that there is an immutable constituent boundary before *of* , given that people quickly learn transitional patterns on the basis of very limited data (Saffran, Aslin, & Newport, 1996, and related studies), and the usage patterns for *in spite of* would encourage speakers to group *of* with *in spite*, rather than with the following noun phrase.

## Conclusion

We have taken stock here of the traditional, discrete constituency view that holds that a word sequence either has a holistic structure or a unique, nested

hierarchical structure. The accounts we have examined ultimately reject usage as an indicator of constituent structure—discarding evidence from semantics and any usage data that might be countered by partial evidence from introspective syntactic tests. Such a conservative approach rejects even the possibility of finding evidence that particular sequences may have reached an intermediate stage of constituency. Moreover, the discrete constituency view would seem to hold that grammar is driven only by abstract syntactic generalizations and is immune to any gradual effects from item-specific usage patterns.

In contrast, as we do not take constituent structure as given innately, we do not give priority to syntactic tests. Rather we consider data from usage, semantics, and language change. Indeed, we have shown that chunking and categorization have semantic effects and change incrementally over time.

Moreover, in keeping with the theory of complex adaptive systems, we consider constituent structure to be emergent from the domain-general processes of chunking and categorization. Human minds track multiple factors related to constituency, and this complex processing correlates with a rich and dynamic structural representation for word sequences. In our model, constituency is the result of interacting influences that are both local and global in nature. The *global* influences that help shape constituents correspond to general patterns in usage. On the other hand, constituency may also be shaped *locally* by item-specific forces over time. If a sequence is consistently used in a particular context (with complex prepositions like *in spite of* as a case in point), that sequence will gradually form into a unit, overriding general patterns elsewhere in usage. In this regard, we embrace Bolinger's early complex systems view of language as a "jerry-built" and heterogeneous structure that is also intricate and tightly organized (1976, p. 1). Rather than assuming that structure is given a priori via top-down blueprints, we agree with Bolinger (1976) and Hopper (1987) that structure emerges locally and is subject to ongoing revision, even while general patterns exhibit apparent stability.

Revised version accepted 9 June 2009

## Notes

1 We recognize that there are some generative syntacticians who have adopted nondiscrete or multifaceted models of constituent structure (for one review, see Carnie, 2007). Our views in the present article may be compatible with such approaches, although we would emphasize that proposals for particular constituent structures should be grounded in usage rather than being postulated ad hoc. Despite recent broadening in generative models, the discrete constituency view remains the

norm in much of linguistic theory, as reflected in descriptive grammars such as in the work by Huddleston and Pullum (2002).

2 One reviewer objects that Newell's quote would predict an infinite regress of chunking in cognition. It is indeed the case that multiword chunks consist of words that are themselves chunks, and these chunks are, in turn, made up of phonetic chunks. However, the human perceptual system is not infinitely fine-grained and thus the nested chunking "bottoms out" on just-noticeable-differences in acoustics (see Pierrehumbert, 2001, p. 141). Regarding an emergentist account of linguistic units, see also Bybee and Beckner (2009).

3 Of course, the constituency of *the charming dog* would also be reinforced by a functional unity for the sequence that arises from semantics.

4 Note that because *spite* is a mass noun, we cannot observe any neutralization of number developing for *in spite of* . However, in grammaticalizing complex prepositions, it is common for count nouns to lose plural markers as another indicator of decategorialization. For example, one would say [*on top of* ] *the houses*, rather than *on tops of the houses* (DeLancey, 1994). It is possible to say *on the tops of the houses*, but only if the speaker is using *tops* referentially, rather than relationally as part of a preposition.

5 Our discussion in this section is paralleled by a broader corpus study in Chapter 3 of Hoffmann (2005), which examines 30 frequent Preposition-Noun-Preposition sequences in British English. Hoffmann similarly found that in actual usage, complex prepositions are unlikely to undergo the syntactic modifications proposed by Seppänen et al. (1994). Further, Hoffmann (2005, pp. 45–46) found compelling evidence that complex prepositions are retrieved as uninterrupted wholes, based on the distribution of filled pauses in speech.

6 We do not discuss at length here an additional test Seppänen et al. (1994) mentioned, namely Ellipsis, which they illustrated with the following constructed example: *Speaker A: He did it in spite of John and the auditor. Speaker B: Of what auditor? I didn't know they had one in this firm* (p. 22). Such a usage strikes us as unacceptable, and it is unattested in the corpora we have reviewed. Similarly, after doing a search of the BNC involving 30 complex preposition sequences, Hoffmann (2005, pp. 48–49) found only one instance of ellipsis that crossed a complex preposition boundary (*with respect to*). Q3

7 Both counterexamples are quotes in academic prose from older sources. One quote is from Henry James (1903) and the other is from Emily Ruete (English translation, 1888): "In spite of her very small size, and of her plain exterior, she possessed an immense power . . . ."

## References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Baayen, H. (1993). On frequency, transparency and productivity. *Yearbook of morphology 1992* (pp. 181–208). Dordrecht: Kluwer Academic. **Q4**

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum, 1*(1), 1–14.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia: n Benjamins.

Bybee, J. (1998). The emergent lexicon. *CLS 34: The Panels* (pp. 421–435). University of Chicago: Chicago Linguistic Society. **Q5**

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givon & B. Malle (Eds.), *The evolution of language from pre-language* (pp. 109–132). Philadelphia: Benjamins.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language, 82*, 711–733.

Bybee, J. (in press). *Language, usage and cognition*. Cambridge: Cambridge University.

Bybee, J., & Beckner, C. (2009). Usage-based theory. In B. Heine & H. Narrog (Eds.), *The Oxford handbook of linguistic analysis* (pp. 915–950). Oxford: Oxford University Press.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degree of constituency: The reduction of *don't* in English. *Linguistics, 37*, 575–596.

Carnie, A. (2007). *Constituent structure*. Oxford: Oxford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Davies, M. (2007). *Time* Magazine Corpus (100 million words, 1920s–2000s). Retrieved July 1, 2008, from http://corpus.byu.edu/time

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 385 million words, 1990–present. Retrieved November 5, 2008, from http://www.americancorpus.org

DeLancey, S. (1994). Grammaticalization and linguistic theory. *Proceedings of the 1993 Mid-America linguistics conference and conference on Siouan/Caddoan languages* (pp. 1–22). Boulder: University of Colorado Department of Linguistics.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *SStudies in Second Language Acquistion, 18*, 91–126.

Haiman, J. (1994). Ritualization and the development of language. In W. Pagliuca (Ed.), *Perspectives on grammaticalization* (pp. 3–28). Amsterdam: Benjamins.

Haspelmath, M. (1998). Does grammaticalization need reanalysis? *Studies in Language, 22*, 315–351.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics, 39*(6), 1041–1070.

Hoffmann, S. (2005). *Grammaticalization and English complex prepositions: A corpus-based study* (Routledge Advances in Corpus Linguistics 7). London: Routledge.

Hopper, P. J. (1987). Emergent grammar. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the 13th annual meeting of the Berkeley Linguistic Society* (pp. 139–157). Berkeley: University of California at Berkeley.

Hopper, P. J. (1991). On some principles of grammaticization. In E. C. Traugott & B. Heine, (Eds.), *Approaches to grammaticalization* (Vol. 1, pp. 17–35). Amsterdam: Benjamins.

Hopper, P. J., & Traugott, E.C. (2003). *Grammaticalization* (2nd ed.). Cambridge: Cambridge University Press.

Huddleston, R. D., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

König, E., & Kortmann, B. (1991). On the reanalysis of verbs as prepositions. In G. Rauh (Ed.), *Approaches to prepositions* (pp. 109–125). Tübingen: Gunter Narr.

Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford, CA: Stanford University Press.

Lightfoot, D. (1979). *Principles of diachronic syntax*. Cambridge: Cambridge University Press.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & Ö. Dahl (Eds.), *Explanations for language universals* (pp. 181–203). New York: Mouton.

McCawley, J. (1982). Parentheticals and discontinuous constituents. *Linguistic Inquiry, 13*, 91–106.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

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

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: Benjamins.

Pullum, G. (2006). Phrasal prepositions in a civil tone. In M. Liberman & G. Pullum (Eds.), *Far from the madding gerund and other dispatches from Language Log* (pp. 41–45). Wilsonville, OR: William James & Company.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A concise grammar of contemporary English*. New York: Harcourt Brace Jovanovich.

Quirk, R., & Mulholland, J. (1964). Complex prepositions and related sequences. *English Studies, 45*, 64–73.

Roberts, I., & Roussou, A. (2003). *Syntactic change: A minimalist approach to grammaticalization*. Cambridge: Cambridge University Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Seppänen, A., Bowen, R., & Trotta, J. (1994). On the so-called complex prepositions. *Studia Anglia Posnaniensia, 29*, 3–29.

Svorou, S. (1994). *The grammar of space*. Amsterdam: Benjamins.

Traugott, E. C., & Dasher, R. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Wray, A. (2006). Formulaic language. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (Vol. 4, pp. 590–597). Oxford: Elsevier.

## Queries

**Q1** Author: Please provide the complete address for the author for correspondence.

**Q2** Author: Please update Bybee, in press, throughout the text and in the References.

**Q3** Author: Please spell out BNC.

**Q4** Author: If an edited book, provide the editor(s).

**Q5** Author: If an edited book, provide the editor(s).