Exploring Sequentiality as the Basis of Constituent Structure

Joshua D. Tanner

Brigham Young University

1. INTRODUCTION. Bybee (2007) explains "The existence of constituent structure and the hierarchical organization resulting from it has always been taken by linguists as prime evidence that linguistic behavior does not consist merely of linear strings of elements. It is further believed that the hierarchical organization of sentences is one of the most basic aspects of language, indeed, a defining feature of human language" (p. 313). The question remains, however, as to why this hierarchical organization exists. This question has gone ignored by most, but Bybee attempts to show that "form is emergent from substance" (p. 313). In other words, hierarchical organization exists because we as speakers make it so by the way we speak. In simple terms, this is a chicken-or-egg type question. Is language born from form, or does form exist because of the language? In Chapter 15 of her book (Bybee 2007), she tests the hypothesis that "constituent structure emerges from sequentiality because elements that are frequently used together bind together into constituents" (p. 314). She shows, using English corpus data, that the most frequent words preceding nouns are articles and determiners and claims that the constituent structure for a noun phrase comes as a result of that. In her words, "Constituents of the type proposed for generative grammar that are described by phrase structure trees do not exist. Instead, units of language (words or morphemes) are combined into chunks as a result of frequent repetition" (p. 332). The present study looks to see if similar findings in Spanish can give further credence to Bybee's claims.

2. BACKGROUND. Constituents, as defined by Cipollone, Keiser, and Vasishth (1998) are semantically coherent groups of words found within sentences. To know whether a group of words is a constituent, students in syntax classes are taught three tests that can be used for identification:

1. Constituents can be used alone to answer questions logically. (*What did you run over? The cat.*)
2. Constituents can be replaced by pro-forms. (E.g., *I ran it over*. Where *it* replaces *the cat*.)
3. Constituents can appear in more than one place in a sentence. An example of this is that a noun phrase can be the subject of a verb or the object of a verb or preposition.

One possible explanation for the existence of a constituent based language organization is, as Langacker (1987) explains, that "hierarchy is fundamental to human cognition" (p. 310). Language forms part of human cognition and, as such, it is easy to see that hierarchy could be applied to it just as it is to other cognitive functions. The fact that constituents can hold semantic weight seems to validate the notion that language is dependent on a constituency organization.

But what comes first? Is language hierarchically based, and what is spoken is a result of that system? Or is that hierarchy created because of the way in which language is spoken? As mentioned before, Bybee (2007) claims that it is frequency that determines constituent structure. The basis for this claim comes from looking at the 11 most frequently appearing nouns in an English corpus and analyzing what words most

frequently appear both before and after each noun. She found that the most frequent items preceding the nouns were articles, possessives and other determiners. Prepositions, adjectives and conjunctions also were found, but with a much lower frequency. These findings lead her to the following proposal:

> Constituents of the type proposed for generative grammar that are described by phrase structure trees do not exist. Instead, units of language (words or morphemes) are combined into chunks as a result of frequent repetition. Most of the time the units of these chunks bear a semantic and/or pragmatic relation to one another allowing them to fulfill the grammatical criteria for constituency: they can sensibly be used alone, as in the answers to questions; they can be replaced by a pro-form; and they can be used in various positions in a sentence … In such cases, where frequency of co-occurrence corresponds to semantic relevance, we have traditional constituents. Indeed the semantic coherence of such units may facilitate their establishment as chunks. However, other types of chunks also exist, as I have demonstrated in this chapter, showing that frequency of co-occurrence is an independent factor. Thus pronoun + auxiliary, preposition + determiner, and verb + preposition sequences can form chunks but are difficult to describe in traditional frameworks since they do not meet the criterion of semantic relevance. For this reason, too, they do not fulfill the grammatical criteria of occurring alone or being replaceable by a pro-form. Thus constituency in this view is the convergence of two other factors and is itself not a basic structure. It is an emergent property of language. (p. 332)

3. THE PRESENT STUDY. What follows is a two-part study. The first part is a duplication of Bybee's study looking at Spanish nouns instead of English nouns. This will serve to validate her findings. The results back up arguments both for sequentiality as the basis of constituent structure and constituency as the basis of language production. The second part of the study focuses on verb + preposition chunks that occur in Spanish. This gives further evidence of non-constituent chunks that arise due to frequency as Bybee explains.

3.1 SPANISH NOUN PHRASES. Replicating what Bybee did in the above-mentioned chapter of her book, I set out to see if similar results would appear in a frequency search in Spanish. I used a Spanish corpus created and maintained by Mark Davies at Brigham Young University[1]. Using this corpus I found the nine most frequent nouns:

Noun	Frequency
Años	37,380
Vez	36,676
Tiempo	36,265
Parte	34,675
Vida	33,918
Hombre	30,876
Casa	29,880
Día	29,328
Señor	26,685

Using the wildcard symbol *, I was able to find what words most frequently precede each noun. I then calculated the frequency and percentage of each type of word found to precede each noun. This data is given here:

[1] www.corpusdelespanol.org

Word Type	Frequency	%
Definite Article	99,982	43
Indefinite Article	26,851	12
Possessive	21,567	9
Other Determiners	57,797	25
Prepositions	23,410	10
Adjectives	3,425	1
Conjunctions	276	0.1

Notice that articles and other determiners combine for 80% of the words directly preceding the nouns. Constituent structure reflects these findings being that determiners are found in the specifier position of a noun phrase. Another 10% of words preceding the nouns are prepositions. This also fits in with constituent structure as nouns are found in the compliment position of a prepositional phrase.

Using the same wildcard search, I found the most frequent words to follow each noun. The word type, frequency and percentage of total are given here:

Word Type	Frequency	%
Preposition	68,478	59
Conjunction	25,450	22
Adjective	6,034	5
Article	3,157	3
Se	3,176	3

In Bybee's findings in English, the most common unit occurring after the noun was the conjunction "and" which only occurred after 7% of the tokens. The Spanish data are quite different with prepositions occurring after the noun 59% of the time. These findings could give further credibility to her final claim that besides constituents, "other types of chunks also exist … showing that frequency of co-occurrence is an independent factor" (p. 332).

3.2 SPANISH VERB + PREPOSITION CHUNKING. Further evidence that constituent structure is born from sequentiality is found in the chunking of linguistic units frequently found together that do not form constituents. Bybee shows a "very robust example" in English auxiliary contraction "which occurs in *I'm, I've, I'd, I'll, he's, he'll, he'd*, and elsewhere" (p. 327). The reason for these contractions that cross constituent boundaries is frequency. Bybee claims that similar chunking occurs in Spanish verbs that are said to take a certain preposition. Her chapter does not include any frequency data to back this claim.

Most Spanish textbooks present verbs that take certain prepositions before infinitives as lists to be memorized. One textbook, *Conexiones* (Zayas-Bazán, Bacon, & García 2010), attempts to give an explanation, but only for verbs that take the preposition *a*: "The preposition **a** follows verbs of motion, of beginning, and of learning process, among others" (p. 229). After this explanation, there are merely lists of verbs that take *a*, verbs that take *de*, verbs take *con*, and verbs that take *en*. Frequency tests to check the validity of this "grammar rule" present interesting results.

In order to test Bybee's claims of frequency of verb + preposition leading to a linguistic chunk and later a grammar rule, I selected 10 verbs from the lists given in Conexiones (p. 230) and searched the aforementioned corpus for what prepositions are found between them and a following infinitive verb. The search string used to give these results was, for example, "[acabar] [E*] [VR*]". [E*] stands for any preposition and [VR*] stands for any infinitive verb. The brackets around the first verb allow any conjugation of the verb to be counted in the results. The results were then tallied, and are shown here:

Verb	Preposition	Frequency	%
Animar	a	59	81
	de	6	8
	para	5	7
	sin	1	1
	por	1	1
Aprender	a	18	60
	para	10	33
	sin	2	7
Invitar	a	142	95
	para	6	4
	sin	1	1
Acabar	de	128	84
	por	21	14
	sin	2	1
	para	2	1
Cesar	de	45	83
	a	5	9
	por	2	4
	sin	1	2
	hasta	1	2
Tratar	de	608	100
Contar	con	44	43
	para	41	40
	por	5	5
	a	7	7
	sin	3	3
	hasta	2	2
Soñar	con	81	73
	en	25	23
	a	4	4
	hasta	1	0
Insistir	en	110	97
	a	2	2
	para	1	1
Quedar	en	42	17
	a	85	34
	por	75	30
	sin	25	10
	para	13	5
	de	9	4

The first preposition given for each verb is the one listed in the textbook as the "correct" preposition. With every verb but one, *quedar*, the expected preposition is the most frequent. Interestingly, the most frequent preposition found with *quedar* has double the frequency of the expected preposition. Only one verb, *tratar*, had just one preposition, though in most cases the expected preposition was much more frequent than any others found. The only exceptions are the aforementioned *quedar* and *contar*, which had an almost even split between *con* and *para*.

4. CONCLUSION. Replicating Bybee's study using a Spanish corpus and further investigating the frequency of prepositions following Spanish verbs has given further credence to her claims. That is, it appears that constituents are born from sequentiality. The fact that there exists chunking outside the constituent realm, as evidenced by the occurrence of certain prepositions after verbs, indicates that language is not formed from constituency. Instead, constituency describes certain word combinations. I agree with Bybee's claim, cited earlier, that "[c]onstituents of the type proposed for generative grammar that are described by phrase structure trees do not exist" (p. 332).

References

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford, New York: Oxford University Press.

Cipollone, N., Keiser, S. H., and Vasishth, S. (1998). *Language Files: Materials for an introduction to language and linguistics*. 7th ed. Columbus, OH: Ohio State University Press.

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

Zayas-Bazán, E., Bacon, S. M., García, D. M. (2010). *Conexiones: Comunicación y cultura*. 4th ed. Saddle River, NJ: Prentice Hall.