To appear, *The Oxford Handbook of Grammaticalization*, Bernd Heine & Heiko Narrog (eds.)

Collocations in Grammaticalization and Variation

Rena Torres Cacoullos and James A. Walker

1. Introduction: Grammaticalization, collocations, and autonomy

Grammaticalization is the set of gradual processes, both semantic and structural, by which constructions involving particular lexical items are used with increasing frequency and become new grammatical constructions, following cross-linguistic evolutionary paths (e.g., Bybee, Perkins, and Pagliuca 1994; Givón 1979; Heine and Kuteva 2002; Traugott 2003). Grammaticalization may involve not only individual lexical items, but also collocations of items, or "conventionalized word sequences" (Bybee 2006:713), including "prefabs" (Erman & Warren 2000), "reusable fragments" (Thompson 2002:141), or "formulaic language" (Corrigan, Moravcsik, Ouali, & Wheatley 2009). Given this gradualness, collocations undergoing grammaticalization will vary in analyzability (or, conversely, what Bybee (2003) calls autonomy), which raises two questions:

> First, to what degree do the sub-units composing the collocation retain individual independent associations with their cognates in other constructions?
> Second, to what degree do collocations retain an association with the (diachronically) related more general constructions?

For example, the phonetic reduction of future *(BE) going to* to *gonna* indicates decreased analyzability (or increased autonomy): *go* is absorbed into a new fused single unit which is autonomous from other instances of the verb *go* and from the general purposive construction from which this future arose (Bybee 2006:719-20). If collocations are ostensibly fixed, are they subject to variation? Do they play a role in grammaticalization and change more generally? If so, how?

In this chapter, we focus on collocations in grammaticalization. Using examples of grammaticalization in Spanish and English, we illustrate how patterns of distribution and co-occurrence can be used to demonstrate the variability and gradience of constituency. Furthermore, our quantitative analysis provides evidence that collocations may be viewed as particular instances of constructions that interact with their more general counterparts to shape grammatical structure.

We begin with two measures of collocation status, indices of unithood and relative frequency, in Section 2. In Section 3, we use multivariate models of variation between alternate forms to show that, despite their status as autonomous units, collocations retain the grammatical patterning of the more general construction with which they are associated. Finally, in Section 4, we examine the role of collocations in contributing to apparent semantic effects in grammatical variation.

2. Measures of collocations: distributional analysis and relative frequency

2.2 Indices of unithood from distributional analysis

Distributional analysis examines the patterning of collocations across different contexts, that is, the proportion of occurrences (tokens) of that collocation in different contexts. For several grammaticalizing constructions, one measure of unithood, or degree of fusion, is adjacency. Over time, a decreasing proportion of tokens occur with other material intervening between subunits of the collocation. For example, the Spanish construction [*ESTAR*$_{\text{Present}}$ + Verb-*ndo*]$_{\text{= present progressive}}$ is on its way to becoming an obligatory aspectual expression. In Old Spanish (12th – 15th centuries), it began as a particular instance of a general gerund construction in which finite forms of spatial (locative, postural, or motion) verbs combined with another verb in gerund (*-ndo*) form to mean '*be*/*go* Verb-*ing*'. In evolving from a locative to a progressive, the construction underwent a change in constituency from a sequence of two independent parts (main verb *estar* 'be (located)' with a gerund complement) to a periphrastic unit, with an auxiliary (*estar*) and a main verb (the gerund). In the absence of observable phonetic reduction, what evidence can be assembled for such a change?

In addition to adjacency, that is, the absence of elements intervening between *estar* and the gerund (such as *ambos* 'both of them' in (1)), another measure of unithood emerging from changes in distribution is association between the finite form of *estar* and a single co-occurring gerund, as opposed to the co-occurrence of two or more gerund complements (e.g. *comiendo* 'eating' and *solazándo* 'sunning' in (1)). A third measure is fusion, the placement of object clitic pronouns before the emerging unit rather than attached to the gerund, or "clitic climbing" (cf. Myhill 1988) (as with reflexive marker *se* 'themselves' in (1)) (Torres Cacoullos 2000:33-55, 71-88).

(1) que tu marido está en la ribera de la mar et que ha por amigo un ximio; et **están** anbos **comiendo** et solazándose. (13th c., Calila e Dimna)
'that your husband is at the seaside and that he has befriended a monkey; and they **are (there)** both of them **eating** and sunning themselves'

Table 1, based on Spanish texts spanning six centuries, shows increasing unithood on all three measures. The proportion of occurrences without intervening material changes significantly across time, beginning with 36% of all tokens in the 13th century and reaching 78% in the 19th. Occurrences of single as opposed to multiple gerunds increase significantly, from 80% in the 13th century to 92% in the 19th. Finally, the rate of proclisis rises from 57% in the combined 13th -15th centuries to 76% in the combined 17th-19th centuries.

Table 1. Increasing unithood [*estar* + Verb-*ndo*] progressive*
(from Bybee & Torres Cacoullos 2009, Table 4)

	13th c.	15th c.	17th c.	19th c.
Adjacency (lack of intervening material)	36% (104)	50% (134)	67% (217)	78% (217)

Chi-square:		4.950998521;
		p = 0.0261
			9.799123895;
			p = 0.0017
				9.323668501;
				p = 0.0023

Association (absence of multiple gerunds)	80% (104)	86% (134)	88% (217)	92% (217)
Chi-square:	6.634394904; p = 0.0100 (13th vs. 19th c.)			
Fusion (“clitic climbing”)	63% (24)	50% (22)	82% (74)	70% (77)
	57% (46)		76% (151)	
Chi-square:	6.682716664, p = 0.0097			

*All tense occurrences. Numbers within ( ) are Ns (tokens); % is proportion of tokens.

A different set of unithood measures emerge from Spanish *a pesar de X* ‘in spite of X’, which developed from a meaning of opposition by another person into a concessive, where the element ‘X’ is an NP, an infinitive, or a finite clause (Torres Cacoullos & Schwenter 2005). The strongest evidence for analyzability appears in coordinated adnominal NPs, where repetition of *de* ‘of’ for each adnominal NP shows the relative independence of this component from the other subunit(s) of the collocation, *a* + *pesar*. In (2a), *de* is repeated with the coordinated NPs, but in (2b), one *de* suffices, or has scope over, both NPs. In fact, repetition of *de* in coordinated adnominal NPs declines from an average of 86% (N = 23) in the 17th and 18th centuries to 60% (N = 30) in the 19th and 20th centuries (Chi-square = 4.298, *p* = 0.038) (Torres Cacoullos 2006: 42).

(2a) algo de atrevido y varonil en todo el ademán, **a pesar de**l recogimiento **y de** la mansedumbre clericales (19th c., Pepita Jiménez)
‘something bold and virile in his whole look, **in spite of** the withdrawal **and of** the tameness of the cleric’

(2b) olía a lavanda y espliego, pero por debajo del perfume olía como yo, la fisiología nos igualaba **a pesar de** los potingues **y** las abluciones diarias (20th c., La tempestad)
‘she smelled of lavendar, but underneath the perfume she smelled like me, our physiologies making us equals **in spite of** the concoctions **and** the daily ablutions’

As erstwhile independent lexical item, *pesar* ‘sorrow’ is absorbed into the fused unit *a pesar de*, it undergoes what Hopper (1991:22) calls decategorialization, shedding its nominal trappings. This is measured by the loss of plural marking (from 6% of all *pesar* tokens in the 12th – 15th centuries, to 1% in the 17th – 20th centuries), a drop in determiners (from 67% to 7%) and adjectival modification (from 10% to 1%), and a decline in coordination with other nouns (as in *mucho* ***pesar*** *y tristeza* ‘much sorrow and sadness’ (15th c., La Celestina), from 30% to 5%) (Torres Cacoullos 2006:38-9). Thus, in both cases, the distributional analyses underscore the gradualness of decategorialization and the gradience of analyzability.

2.2 Collocations and relative frequency

Increased frequency of use is integral in grammaticalization. Although text-based studies have profitably paid attention to token frequency (e.g., papers in Bybee and Hopper 2001), relative frequency measures, which consider occurrences of subunits outside the collocation, may provide another gauge of analyzability.

In the case of gerund (*-ndo*) periphrases (including the [*estar* + Verb-*ndo*] progressive mentioned above), token frequency rises in tandem with an increase in the proportion of gerunds in construction with an (emerging) auxiliary (in addition to *estar* 'be (located)', motion verbs *ir* 'go', *andar* 'go around', *seguir* 'follow, continue'): that is, an increased rate of gerunds preceded by an auxiliary relative to those gerunds that stand alone (as adverbials or relatives). Token frequency doubles between the 16th and early 20th centuries, from 8 to 16 occurrences per 10, 000 words, as does relative frequency, from 14% to 24%. Such increased relative frequency means a greater probability that a gerund is tied to an auxiliary, leading to the growing identification of the auxiliary + Verb-*ndo* sequence as a unit (Torres Cacoullos 2000:55-60).

In the development of concessive *a pesar de* 'in spite of', increased token frequency is accompanied by increased frequency of the collocation with respect to occurrences of lexical item *pesar* outside the collocation, which swells from 2% to 96%, a spectacular reversal (Table 2). As the collocation rises, other uses of *pesar* 'sorrow' (object, subject or adverbial phrases) decline steeply, to the point that *pesar* today occurs virtually always flanked by *a* and *de*.[1] Table 2 shows that the grammaticalization measures of decategorialization of *pesar* and fusion of the subunits in *a pesar de* (reviewed in Section 2.1) shift in tandem with shifts in the relative frequency of the collocation.

Table 2 Grammaticalization and relative frequency: collocation *a pesar de* is increasing proportion of occurrence of *pesar* (from Torres Cacoullos 2006, Tables 8-10)

Century	Unithood *a pesar de*	Decategorialization *pesar*	Token frequency *a pesar de*	Relative frequency *a pesar de/pesar*
12th – 15th	--	20%	< 1	2% (4/199)
17th	73%	5%	12	72% (58/81)
19th	52%	2%		86% (169/196)
20th	30%	0		96% (167/174)

Unithood measure is repetition of *de* in coordinated adnominal NPs, decategorialization measure is use of definite article (see Section 2.1); token frequency is normalized per 100,000 words.

The cases reviewed here provide evidence that one measure of collocations is diachronically increasing relative frequency of the sequence of words with respect to the lexical component of the sequence. Such relative frequency may promote the absorption of the lexical constituent as the sub-units of the collocation fuse, as well as the autonomy of the collocation from its erstwhile lexical constituent (similar to the reduced compositionality and semantic

[1] We do not expect that a decrease in the relative frequency of lexical use (with respect to a collocation) will necessarily correspond with a decrease in the absolute token frequency (e.g. per 10,000 words).

transparency of morphologically complex words that are more frequent than their bases (Hay 2001)).[2]

3. **Collocations, from the particular to the general: evidence from the linguistic conditioning of variants**

We now move beyond simple measures of frequency to patterns of variation: that is, quantitative models of speaker choices between variant forms serving generally similar discourse functions. These patterns are observable in linguistic conditioning, probabilistic statements about the relative frequency of co-occurrence of linguistic forms and elements of the linguistic context (Labov 1969, Sankoff 1988). The methodological tool is a comparison of the linguistic conditioning of variants involving putative collocations and variants that do not (cf. Poplack & Tagliamonte 2001; Tagliamonte 2002). While different linguistic conditioning indicates collocational status, parallels in linguistic conditioning reveal shared grammatical patterning with other instances of more general constructions.

Consider the variable use of complementizer *that* to link two clauses, a widely-attested feature of all varieties of English, illustrated in (3). Certain frequent collocations of main-clause subjects and verbs, such as *I think* and *I guess*, have been proposed as discourse formulas that function more as epistemic adverbials than as main clauses (Thompson and Mulac 1991).

(3a) And I let it slip **that** Darth Vader was Luke's father. (071.468)
(3b) I can't even believe **Ø** I just said that. (059.1840)[3]

Examining a corpus of Canadian English (Poplack, Walker & Malcolmson 2006), we note a remarkably skewed distribution of main-clause verb types in the variable complementizer construction: *think, know*, and *say* account for 63% of all the data, while five additional verbs (*guess, tell, remember, find*, and *realize*) account for a further 19%. In other words, just eight lexical types make up 81% of the data. Nevertheless, there is no one-to-one correlation between token frequency and the rate of *that* (vs. zero complementizer), which averages 21% (N=2820) in the data examined. Among middle-frequency lexical types (50-200 tokens), the highest frequency verb (*guess*) has a rate of *that* at 3%, but the second-highest (*tell*) has a higher than average 43%. Among high-frequency types, *know* and *say* occupy roughly the same proportion of the data (9–10%), but *know* shows 34% *that* while *say* has 27%, both at rates higher than several much lower frequency predicates (Torres Cacoullos & Walker 2009:19-20).

However, if we adopt a relative-frequency view of collocations, examining main-clause subjects and verbs reveals a substantial difference between the three highest-frequency lexical types, *think*, *know,* and *say*, which present quite disparate rates of *that.* Unlike *know* and *say*, *think* is largely restricted to first-person singular present tense. As shown in the middle column in Table 3, seven subject-verb collocations make up an average of 80% of their respective lexical

[2] Relative frequency as the proportion of a lexical type in a particular construction has been shown to correlate with phonetic reduction (Alba 2008, Hollman and Sierwieska 2007). Other measures of associations between words are tested in Krug (1998) and Jurafsky et al. (2001).

[3] Examples are taken from the *Quebec English Corpus* (Poplack et al. 2006).

types: *I think, I guess, I remember, I find, I'm sure, I wish, I hope*. In contrast, all other subject-verb combinations that make up a substantial proportion of their respective verb types present an average of only 19%.

Table 3: Main-clause subject-verb collocations by proportion of lexical type and rate of complementizer *that*.

Collocations		N	% Lexical Type	% *that*
I think		734	61	5
I guess		163	99	3
I remember		90	96	4
I find		59	66	24
I'm sure		40	74	10
I wish		17	85	0
I hope		15	79	7
	Total	1118	80	8
Low-relative-frequency subject-verb sequences		216	19	31

This sharp difference in relative frequency clearly correlates with rates of *that*. As shown in the right-most column in Table 3, the average rate of *that* for the seven high-relative-frequency subject-verb sequences is a bare 8% (though the rate for individual collocations within this set ranges from 0% to 24%), in contrast to 31% for the infrequent sequences. Are these high-relative-frequency subject-verb sequences indeed discourse formulas, that is, autonomous collocations? Beyond the correlation with lower rates of *that*, we seek evidence for their status as collocations by examining the patterns of *that*-variation, by comparing the linguistic conditioning of *that* in the putative collocations and in the other instances of the construction that do not involve collocations.

For each token, we noted whether *that* was present or not (the dependent variable) and coded for a number of factor groups (independent variables) to operationalize syntactic, semantic, or discourse-pragmatic hypotheses about the choice of variants (*that* or zero) based on contextual features. All of the factor groups were considered in multivariate analysis using GoldVarb X (Sankoff, Tagliamonte & Smith 2005), to discover the set of factor groups that jointly account for the largest amount of variation in a statistically significant way.

We are interested in two lines of evidence from the multivariate analyses (cf. Tagliamonte 2006:235-245). First, the direction of effect (or "hierarchy of constraints" (Labov 1969:742)) is instantiated in the order of the factors within a factor group, from most to least favorable, as indicated by the probability or factor weight: the closer to 1, the more likely, the closer to 0, the less likely that the variant of interest (here, *that*) will be chosen in the given environment. Second, the relative magnitude of effect is indicated by the range, the difference between the highest and lowest factor weight in the group.

Table 4 compares two multivariate analyses of factors contributing to choice of *that* in the subject-verb collocations identified by the relative frequency measure (Table 3) and in tokens not involving the collocations. The input, which indicates the overall likelihood that *that* will occur, is much lower in the collocations (.05) than in the other occurrences (.33), as expected.

Table 4: Two independent multivariate analyses of linguistic factors contributing to the occurrence of complementizer *that*: (a) excluding main-clause subject-verb collocations; (b) in main-clause subject-verb collocations (*I think, I guess, I remember, I find, I'm sure, I wish, I hope*)

	Non-Collocations	Collocations
Total N:	1552	1118
Input:	.33	.05
Complement-Clause Subject		
Noun Phrase (e.g., 3a)	**.65**	**.68**
Other Pronoun	.52	.48
I (3b)	.42	.42
it/there	.38	.43
Range:	*27*	*26*
Adjacency: Intervening Material		
Present (*I remember on Saturday, my mother used to)*	**.72**	**.72**
Absent	.45	.47
Range:	*27*	*25*
Main-Clause Subject		
Noun phrase (*the teacher realized that…*)	**.68**	N/A
Pronoun	.45	
Range:	*23*	
Main-Clause Adverbial		
Post-subject (*they still think that…)*	**.65**	**.83**
Phrasal (*at the beginning, we told the guy that…)*	**.59**	.52
None	.47	.48
Pre-subject (*so already they think…)*	.45	**.77**
Range:	*20*	35
Adjacency: Intervening Verbal Arguments		
Present (*we told the guy that…)*	**.60**	[ ]
Absent	.49	[ ]
Range:	*11*	
Main-Clause Verbal Morphology		
Non-finite (*you would tell…)*	**.58**	N/A
Finite (*they tell me that…*)	.47	
Range:	*11*	
Complement-Clause Transitivity		
Transitive (*…find that they had a job*)	**.56**	[ ]
Intransitive	.47	[ ]
Range:	*9*	

Factors not selected as significant:	Subject coreferentiality	X	X
	Harmony of polarity	X	
	Intervening verbal argument		X
	Comp-clause mood/morphology	X	
	Comp-clause transitivity		X
	Cotemporality	X	

How does the linguistic conditioning compare? In the non-collocations (instances of the general complementizer construction, excluding *I think* and the other subject-verb collocations identified in Table 3), shown in the left-hand column, the greatest effects (range = 27) are exerted by the increasing referentiality of the complement-clause subject (NP (3a) > other pronoun > *I* (3b) > *it/there*) and by the presence of intervening material between the two clauses. Also strong are the effects of main-clause NP subjects (range = 23) and of post-subject adverbials (range = 20). Complex main-clause verbal morphology and arguments in the main clause contribute lesser effects. This configuration of effects indicates that *that* serves to demarcate the boundary between two clauses which both have (lexical) content. In contrast, zero occurs when there is less semantic content and the two clauses behave more like a single proposition (cf. Fox & Thompson 2007).

If the collocations, such as *I think* and *I guess*, are fixed units, autonomous from other instances of the complementizer construction, the factors conditioning *that* should differ. As the right-hand column in Table 4 shows, the three factor groups with the greatest magnitude of effect for non-collocations are also significant for collocations: intervening material, complement-clause subjects, and main-clause adverbials. However, the relative magnitude of effect is not identical. The greatest contribution to *that* in the collocations is the main-clause adverbial (range = 35). This confirms their status as collocations, since the presence of a post- or pre-subject adverbial (4) nullifies the formulaic nature of the collocation.

(4a) I personally think that it is well worthwhile. (027.737)
(4b) Actually I- I think that those were the only two things they said in the entire skit. (071.737)

Nevertheless, it is important to note that the direction of effect is largely parallel. The choice of *that* is favored most by full NP (and least by *it/there* and *I*) complement-clause subjects and by the presence of intervening material. Even though the absence of *that* is near-categorical (input =.05), frequent collocations retain traces of grammatical conditioning. The parallelism in linguistic conditioning shows that, despite their status as fixed collocations, these units are not completely autonomous from other instances of the complement construction. Patterns of variation thus demonstrate that collocations may maintain associations with more general constructions.

4. Collocations shape grammatical variation

In the previous section, we compared the conditioning of linguistic variation by language-internal factors in collocations and in their associated general constructions. We provided evidence for formulaic or conventionalized collocations through differences in linguistic conditioning. In this section, we discuss a slightly more complicated case, in which the collocational effects do not occupy the same proportion of the data. Nevertheless, we will demonstrate that these collocational effects can still be powerful (cf. Walker 2007). Specifically, we will show that apparent semantic effects are derived from, and shaped by, the influence of frequent collocations.

We take the example of the future in English, an area of grammar in which a number of variants coexist. Here we focus on variation between the two most robust future forms, *will* and *going to*, as shown in (5).

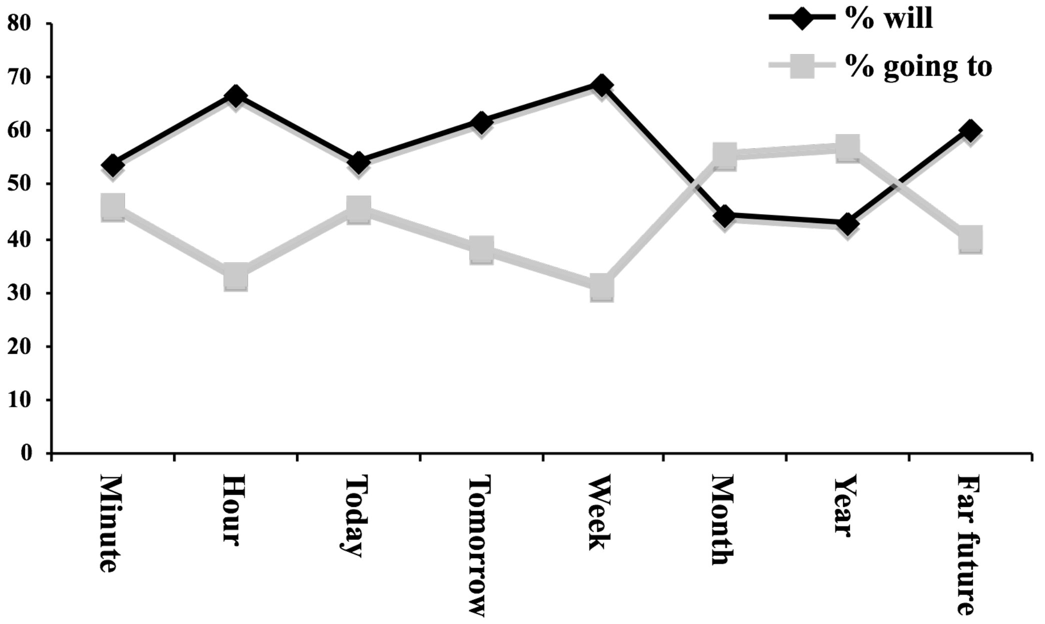
(5a) And he***'ll*** probably live 'til a hundred. (29:1480)
(5b) My doctor tells me I***'m going to*** live 'til a hundred. (29:341)

Table 5: Factors contributing to the choice of future *will* (vs. *going to*) in Quebec English.

	Total N:	2,807	
	Input:	.52	
Sentence Type			
Declarative		.54	
Negative		.47	
Yes/*no* question		.31	
Wh-question		.09	
	Range:		*45*
Clause Type			
Apodosis		.59	
Other main		.53	
Other		.36	
	Range:		*23*
Temporal Adverbial			
Nonspecific/indefinite (*never*)		.67	
Specific/definite		.48	
No adverbial		.48	
	Range:		*19*
Grammatical Person and Animacy of Subject			
1st person sg.		.56	
1st person pl.		.50	
3rd person, inanimate		.49	
3rd person, animate		.47	
2nd person		.38	
	Range:		*18*
Proximity			
Proximal (same day)		.49	
Distal		.51	
	Range:		*2*

Semantic differences attributed to the two variants include proximity, certainty or willingness (see Torres Cacoullos & Walker 2009b for an overview). Table 5 shows the results of a multivariate analysis of the factors contributing to the occurrence of *will*. There is no strong effect of proximity, but a more detailed breakdown (Figure 1) reveals that *will* tokens occur disproportionately in 'within a minute' and 'within an hour' contexts, most of which have first-person singular subjects (e.g. *I'll tell you*). Although we might interpret this as a persistence of willingness in the use of *will* for offers, first-person *will* (*'ll*) collocations (6) also make up substantial proportions of their corresponding lexical types. Taken together, these results suggest that perceptions of proximity and willingness may have more to do with fixed discourse formulas.

Figure 1: Distribution of future variants by temporal proximity.



(6)	*I'll tell…*	58% of *tell* (40/69)
	I'll pay…	53% of *pay* (8/15)
	I'll ask…	44% of *ask* (8/18)
	I'll talk…	38% of *talk* (9/24)
	I'll call…	35% of *call* (11/31)
	I'll teach…	33% of *teach* (7/21)
	I'll give…	33% of *give* (21/64)
	I'll try…	29% of *try* (6/21)
	I'll speak…	26% of *speak* (6/23)
	I'll say…	20% of *say* (7/35)

The other effects shown in Table 5, although at first glance semantic, can in each case be shown to reflect (at least in part) the effects of collocations. The effect of indefinite adverbials is due in part to *never*, which constitutes a third (78/249) of these tokens and overwhelmingly (72%

(56/78)) occurs with *will*, largely in the collocation 'X *will/'ll never …*' (66% (37/56)). Many of the non-main clauses, which disfavor *will*, are preceded by a frequent collocation of cognition verb and first-person subject (7), such as *I think.* As we saw in the previous section, such collocations function more as epistemic phrases than as clauses. Rather than indicating degree of certainty, the effect may reflect early generalization of *go*-futures in contexts expressing speaker viewpoint (cf. Traugott 1995), and thus retention of earlier distribution patterns rather than source-construction meaning (cf. Torres Cacoullos 2001).

(7a) And I think she***'s gonna*** need uh- you know, she'll need that extra support. (48:865)

(7b) I'm sure today there***'ll*** be a lot of people at the movies. (67:1310)

The strongest effect, that of questions, can also be attributed to collocations. For example, *What am I going to do?, Is there gonna be…?* and *What's gonna happen?* make up substantial portions (approximately one-fifth) of their respective grammatical persons in questions. The significance of grammatical person in Table 5 is also due, at least in part, to the disfavoring effect of second person singular, which interacts with questions. Thus, general effects which at first glance appear to indicate semantic nuances may largely reflect particular constructions: nuances of proximity and willingness are due to formulas such as *I'll tell you*; the indefinite adverbial effect which operationalizes certainty is due to the *X'll never…* construction; rhetorical or formulaic questions contribute to the interrogative effect.

5. Conclusion

In this chapter, we have used various measures — token frequency, relative frequency, linguistic conditioning and comparative analysis — to examine the role of collocations in variation and grammaticalization. These measures demonstrate that analyzability in collocations (the converse of autonomy) is a gradient property. Collocations show retention of linguistic conditioning: that is, they retain patterns associated with their more general cognate constructions. This retention indicates that, rather than being completely autonomous units, collocations can be viewed as particular instances of constructions which, while formulaic, interact with their associated general constructions. The patterns of future forms reviewed here also provide evidence for an interaction of collocations with general, productive constructions. Diachronic studies show that collocations constitute an important locus of grammatical development, since they may lead in changes and constitute subclasses that contour the grammaticalization of more general constructions (Bybee & Torres Cacoullos 2009). Thus, rather than being a peripheral part of the linguistic system, collocations may be considered an integral part of grammar.

References

Alba, Matt (2008). 'Ratio frequency: Insights into usage effects on phonological structure from hiatus resolution in New Mexican Spanish', *Studies in Hispanic and Lusophone Linguistics* 1: 247-86.

Bybee, Joan (2003). 'Mechanisms of change in grammaticization: The role of frequency', in Richard Janda and Brian Joseph (eds.), *The Handbook of Historical Linguistics*. Oxford: Blackwell, 624-47.

Bybee, Joan (2006). 'From usage to grammar: The mind's response to repetition', *Language* 82: 529-51.

Bybee, Joan, Perkins, Revere, and Pagliuca, William. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Bybee, Joan, and Hopper, Paul J. (eds.) (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Bybee, Joan, and Torres Cacoullos, Rena (2009). 'The role of prefabs in grammaticization: How the particular and the general interact in language change', in Roberta L. Corrigan, Edith A. Moravcsik, Hamid Ouali, and Kathleen Wheatley (eds.), *Formulaic language, volume 1. Distribution and historical change*. Amsterdam: John Benjamins, 187-217.

Corrigan, Roberta L., Moravcsik, Edith A., Ouali, Hamid, and Wheatley, Kathleen (eds.) (2009). *Formulaic Language, Volume 1. Distribution and Historical Change*. Amsterdam: John Benjamins.

Erman, Britt, and Warren, Beatrice (2000). 'The idiom principle and the open choice principle'. *Text* 20: 29-62.

Fox, Barbara A., and Thompson, Sandra A. (2007). 'Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language* 31: 293- 326.

Givón, Talmy (1979). *On Understanding Grammar*. New York: Academic Press.

Hay, Jennifer (2001). 'Lexical frequency in morphology: Is everything relative?', *Linguistics* 39: 1041-1070.

Heine, Bernd, and Kuteva, Tania (2002). *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hollmann, Willem and Siewierska, Anna (2007). 'A construction grammar account of possessive constructions in Lancashire dialect: Some advantages and challenges'. *English Language and Linguistics* 11: 407-24.

Hopper, Paul J. (1991). 'On some principles of grammaticalization', in Elisabeth C. Traugott and Bernd Heine (eds.), *Approaches to Grammaticalization*. Amsterdam: John Benjamins, 17-35.

Jurafsky, Daniel, Bell, Alan, Gregory, Michelle, and Raymond, William (2001). 'Probabilistic relations between words: Evidence from reduction in lexical production', in J. Bybee and P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 229-54.

Krug, Manfield (1998). 'String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change'. *Journal of English Linguistics* 26: 286-320.

Labov, William (1969). 'Contraction, deletion, and inherent variability of the English copula', *Language* 45: 715-62.

Myhill, John (1988). 'The grammaticalization of auxiliaries: Spanish clitic climbing', *Berkeley Linguistics Society* 14: 352–63.

Poplack, Shana, and Tagliamonte, Sali (2001). *African American English in the Diaspora: Tense and Aspect*. Oxford: Blackwell.

Poplack, Shana, Walker James A., and Malcolmson, Rebecca (2006). 'An English "Like no Other"? Language contact and change in Quebec', *Canadian Journal of Linguistics* 51: 185-213.

Sankoff, David (1988). 'Sociolinguistics and syntactic variation', in F. J. Newmeyer (ed.), *Linguistics: The Cambridge Survey*. Cambridge: Cambridge University Press, 140-61.

Sankoff, David, Tagliamonte, Sali, and Smith, Eric (2005). *GoldVarb X: A Variable Rule Application for Macintosh and Windows*. Department of Mathematics, University of Ottawa, and Department of Linguistics, University of Toronto.

Tagliamonte, Sali A. (2002). 'Comparative sociolinguistics', in J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.), *Handbook of Language Variation and Change*. Oxford: Blackwell, 729-63.

Thompson, Sandra A. (2002). "Object complements" and conversation. *Studies in Language* 26: 125-64.

Thompson, Sandra A., and Mulac, Anthony (1991). 'A quantitative perspective on the grammaticalization of epistemic parentheticals in English', in Bernd Heine and Elizabeth C. Traugott (eds.), *Approaches to Grammaticalization, Volume II*. Amsterdam: John Benjamins., 313-29.

Torres Cacoullos, Rena (2000). *Grammaticization, Synchronic Variation, and Language Contact: A Study of Spanish Progressive* -ndo *Constructions*. Amsterdam: John Benjamins.

Torres Cacoullos, Rena (2001). 'From lexical to grammatical to social meaning', *Language in Society* 30: 443-78.

Torres Cacoullos, Rena (2006). Relative frequency in the grammaticization of collocations: Nominal to concessive *a pesar de*. in T. Face and C. Klee (eds.), *Selected Proceedings of the 8th Hispanic Linguistics Symposium.* Somerville, MA: Cascadilla Proceedings Project, 37-49.

Torres Cacoullos, Rena, and Schwenter, Scott A. (2005). Towards an operational notion of subjectification. *Berkeley Linguistics Society* 31: 347-58.

Torres Cacoullos, Rena, and Walker, James A. (2009a). 'On the Persistence of Grammar in Discourse: A variationist Study of *That*'. *Linguistics* 47: 1-43.

Torres Cacoullos, Rena, and Walker, James A. (2009b). 'The present of the English future: Grammatical variation and collocations in discourse', *Language* 85: 321-54.

Traugott, Elizabeth Closs (1995). 'Subjectification in grammaticalization', in Dieter Stein and Susan Wright (eds.), *Subjectivity and Subjectivisation*. Cambridge: Cambridge University Press, 31-54.

Traugott, Elizabeth Closs (2003). 'Constructions in grammaticalization', in Brian Joseph and Richard Janda (eds.), *The Handbook of Historical Linguistics*. Oxford: Blackwell, 624-47.

Walker, James A. (2007). "There's Bears Back There": Plural existentials and vernacular universals in (Quebec) English. *English World-Wide* 28: 147-66.