

# **Phonological and Grammatical Variation in Exemplar Models**

Joan Bybee & Rena Torres Cacoullos  
University of New Mexico

## Abstract

Phonological and grammatical structure is shaped by usage patterns, as demonstrated by the effects of context and frequency on variation and change. We argue for an exemplar model of lexical representations, in which tokens of use are registered in memory, including phonetic detail as well as linguistic and social contextual information. Since variation is omnipresent in the input, it comes to be represented directly in cognitive representations, which are a record of speakers' experience with language. Frequency of use and other lexical effects in sound change, which is gradual both phonetically and lexically, are built into exemplar models as the strengthening of exemplars by use and the clustering of exemplars based on phonetic and semantic similarities. The effects of particular lexical items and collocational discourse routines in morpho-syntactic variation and change, including the interaction of the particular and the general in grammaticization, are similarly modeled by the representation of specific instances of constructions and the gradient associations among related forms. Since variation in language use is pervasive and highly conditioned by context, exemplar models are particularly well-suited to account for variation and change.

## **1. Exemplar models for phonetics and phonology**

Exemplar models were originally proposed for phonetics and phonology as a result of attempts to understand how speakers recognize familiar voices. Given the general principle applied to phonological representations that all predictable and non-linguistic information is excluded, these representations would have to be normalized versions of actual speech tokens since such tokens are always spoken by a particular individual with his or her own set of acoustic properties. However, Goldinger (1996) demonstrated that when subjects were asked to listen to lists of words and then say if the word was 'new' or had occurred previously in such a list, words spoken by the same speaker were more likely to be correctly identified as previously heard, even when twenty four hours had elapsed. This demonstrated that in listening, the particular phonetic details of a perceived token remain in memory.

Other arguments for phonetic detail in lexical representation have come from the lexical diffusion of phonetically gradual sound changes, where it is often found

that particular words have their own range of phonetic detail (Bybee 2000, 2001, 2002, Hooper 1981). Given ongoing change, it is also often the case that high frequency words change earlier than low frequency words. This fact is a natural outcome of exemplar representation, as we shall see below.

In an exemplar model, every token of use is registered in memory; if an input token is the same as an existing exemplar, it is mapped onto that exemplar, strengthening it. If it is not similar enough for a mapping to an existing exemplar, a new exemplar is established, positioned in a metaphorical space close to similar exemplars (Bybee 2001, Pierrehumbert 2001). Thus for every word or phrase in a speaker's lexicon, there is a cloud or cluster of phonetic exemplars. The meaning of the word or phrase is also represented by a cluster of exemplars which represent the context and meaning for each token of a word. It is proposed that memory for linguistic objects is the same as for non-linguistic objects, which means that memories can also decay. Particular exemplars that are marginal and not reinforced may be lost, keeping word (and other) categories more centered (Pierrehumbert 2003, Wedel 2007).

In an exemplar model, then, variation can be represented quite directly as a record of the experience a user has with his or her language. Since variation is inherent to experience, it is also inherent to the exemplar representation. In this paper we discuss two dimensions of phonological variation that are handled quite naturally in exemplar models and then go on to demonstrate how construction cum exemplars helps us understand morphosyntactic variation as well.

## 2. Frequency effects in phonological/phonetic variation

Quite a number of cases have now been studied in which high frequency words have been shown to undergo sound changes at a faster rate than low frequency words where both contain the phonetic conditioning environment (Bybee 2000, 2002, Phillips 1984, 2001). In addition, in cases where phonetic variation occurs but may be relatively stable, it is also the case that high frequency words show more variation and/or greater reduction (Bybee & Scheibman 1999, Hooper 1976, Phillips 1984). We will examine one clear case from Spanish dialects, the case of eth-deletion in New Mexican Spanish.

Bybee (2001) reports on a study of 751 tokens of medial orthographic *d* from a corpus of contemporary New Mexican Spanish (Bills & Vigil 1999). Instances of medial *d* were transcribed as either present or absent. It was found that word frequency had a significant effect on the likelihood of deletion. Table 1 shows examples of words of lesser and greater frequency, their frequency in the corpus and their rate of deletion. The inflected forms of *quedar* and *poder* are grouped together.

	Low frequency			High frequency	
	Occurrences	Deletions		Occurrences	Deletions
cada	6	1	nada	31	10
grado	3	0	lado	7	3
queda	} 8	0	pueda	} 36	10
quedo					
quedan					
quedamos					
			puedas		
			puedes		

**Table 1. Relative frequency of some word pairs in the New Mexican corpus and the number of deletions**

Table 2 shows the rate of deletion by word frequency for the non-past participle forms. The past participle forms are excluded here because it appears that they have a higher rate of deletion (but see Bybee 2001, pp.148-153). The token frequency numbers are taken from the much larger COREC corpus (Marcos Marín 1992). It was decided somewhat arbitrarily to divide the words into those with a frequency of 100 and greater vs. those with a frequency of less than 100.

	Low frequency (99-)	High frequency (100+)	Total
Retention	243	287	530
Deletion	23 (8.6%)	78 (21.4%)	101 (16.0%)
Total	266	365	631

Chi-square = 17.3,  $p < 0.001$

**Table 2. Rate of deletion according to token frequency for all non-past participle tokens in the NM corpus using the COREC as a measure of frequency (Marcos Marín 1992)**

The figures show a significant effect of word frequency. It should be noted that the claim is not that word frequency causes deletion, but rather that the phonetic context (between two vowels) is the primary cause, and the deletion is hastened in high frequency words.

An exemplar model provides a natural way to treat both the phonetic variation, which is lexically-specific, as well as the effect of token frequency. As individual words each have their own exemplar clusters, words with more deletion will have more exemplars lacking the medial *d*; thus the lexical variation is given a direct representation. So the first point is that in an exemplar model each word can have its own range of variation.

The effects of frequency of use are built into the model in three ways: 1) Exemplars are strengthened with each use, making them more likely to be chosen for subsequent use, 2) high frequency words are likely to have a greater range of variation so that their exemplar clusters will be larger, and 3) each use of an exemplar in real time has the potential for phonetic effects to have an impact on it. It is the third point that provides the mechanism for the greater phonetic effects on high frequency words.

Given that phonetic processes, such as the weakening of medial *d*, apply in real time as a word is articulated, high frequency words have more opportunities to undergo phonetic processes than low frequency words (Bybee 2000, 2001, Moonwomon 1992). This process has been modeled by Pierrehumbert (2001). The model selects an exemplar for production from an existing cluster and in each production the articulation of the word is biased towards lenition. The affected token is then mapped back onto the exemplar cluster, causing the whole cluster to gradually move towards greater lenition.

### 3. Other lexical effects in sound change

Words are used in many different environments. If a sound change is affecting an initial or final segment of the word, the conditioning environment is variable since it depends upon preceding or following words. If a word more often occurs in the environment for the change, the change will proceed more quickly in that word. Bybee (2002) demonstrates this effect for final *t/d* deletion in English. That study shows that words ending in */t/* or */d/* that frequently occur before consonants (the conditioning environment for deletion) have higher deletion in all contexts, including those before vowels.

Another such example is found in Brown (2004), who studies the reduction to [h] of syllable-initial, including word-initial, */s/* in New Mexican Spanish. The phonetic contexts favoring reduction are preceding and following non-high vowels. The overall rate of word-initial reduction is only 16%. Given the phonetic conditions there is a variable likelihood of deletion as shown in (1) and (2).

- (1) Likely to reduce  
*no sabemos* ‘we didn’t know’  
*la señora* ‘the lady’
- (2) Unlikely to reduce  
*el señor* ‘the gentleman’  
*su suegra* ‘his/her mother-in-law’

For word-initial */s/*, the following phonetic environment is always the same, but the preceding one changes in context. Brown found a significant difference in the

reduction of word-initial /s/ when taking into account how often the word occurred in a favoring environment. For instance, compare *señor* and *señora*. The latter occurs frequently after a low vowel, which conditions the reduction to [h] since it occurs after the definite and indefinite articles, both of which end in *a* in the feminine: *la señora, una señora*. However, the masculine articles are *el* and *un*, both of which end in consonants and thus do not condition the reduction. Thus when *señor* occurs in a favoring environment, as in *no señor*, the reduction is much less likely to occur than if *señora* occurs in that environment. Thus speakers are much more likely to say *no heñora* than they are to say *no heñor*. See Table 3 from Brown (2004, p.103).

Preceding Phonological Environment	FFC < 50		FFC > 50	
	N	%	N	%
Favorable	35/403	9	<b>267/741</b>	<b>36</b>
Unfavorable	33/686	5	19/344	6

Favorable: Chi-square = 100.4769,  $p < 0.0001$

Unfavorable: Chi-square = 0.2428,  $p = 0.6222$

**Table 3. Word-initial /s/ reduction rates for words with favorable and unfavorable preceding phonological environment with low and high Frequency in Favorable Context (FFC). (Brown 2004, p.103)**

This reduction, then, is sensitive to the speaker's prior experiences. Since *señora* occurs frequently in the reducing environment, the reduced exemplars for this word are much stronger than they are for *señor* and thus more likely to be chosen for production. Interestingly, it is not just the current phonetic context that affects the word's phonetic shape, but also the contexts in which the word usually appears.

Thinking in terms of phonological representations that contain phonetic detail associated with the various exemplars of a word allows us to represent phonetic change as gradual on both the phonetic and the lexical dimension. It also allows us to discover and describe lexical effects that have not been previously noted.

#### **4. Lexical and collocation effects in grammatical variation**

Exemplar models provide the formal representation for usage-based grammar, which views linguistic structure as emergent from language use. Usage leads to different degrees of conventionalization of discourse patterns, from reusable fragments (Thompson 2002, p.141) or prefabs (Erman & Warren 2000) to more generalized and schematic constructions. Since tokens of experience are registered in memory, specific instances of constructions are represented in exemplar models. The model of lexical associations proposed for morphologically complex words in

Bybee (1985, 1988, 2001) can be extended to multiword units as associations made among related forms are gradient and depend upon the degree of semantic and phonological similarity and the token frequency of the specific items. In exemplar models, the representations of constructions consist of categories that group together all the exemplars of a given construction, based on semantic and formal similarity. One of the main determinants of memory storage is frequency in experience; thus specific instances of constructions may occur as units in memory storage, even if their meaning and form is predictable from the more general construction. In this section and the following two we demonstrate the appropriateness of exemplar models for the treatment of morphosyntactic variation and change.

An accumulating body of variationist work reveals a greater role for specific units – lexical items and collocations – in grammatical variation than has generally been acknowledged. For example, Poplack's (1992) work on variation in Canadian French shows that one verb, *falloir* 'to have to', exerts an enormous effect on the choice of verbal morphology in subjunctive contexts. Similarly, work on African American English shows that the verb *come* accounts in large part for unmarked past tense in irregular verbs (Poplack & Tagliamonte 2001). We find similar effects with specific units at the phrasal level (cf. Jackendoff 1997), or chunks (Bybee 2001). A great deal of variable use of *that* in English can be attributed to frequent first person singular and present tense collocations such as *I think* and *I guess* (Torres Cacoullós & Walker forthcoming). In variation between future expressions *will* and *be going to*, a number of constructions of differing levels of abstraction and productivity, including fixed expressions (e.g. *What's gonna happen?*) and ones with open slots (e.g. *X'll never Y*), contribute to the general patterns of distribution (Torres Cacoullós & Walker 2008). In other words, grammatical domains are contoured by lexical units or stored routines.

The Spanish intransitive motion verbs *subir* 'go up' and *bajar* 'go down' variably appear with the middle marker *se*. This variability is structured, as reflected in the relative frequency of co-occurrence of the variant forms and elements of the linguistic context. Operationalizing and testing aspectual and pragmatic hypotheses about middle marking with variable-rule analysis (Sankoff 1988, Sankoff, Tagliamonte & Smith 2005), Torres Cacoullós & Schwenter (forthcoming) show that factors in speakers' choice of the *se*-marked variant include a focus on the moment of change (operationalized by co-occurring locative preposition) and speaker involvement and expectations (grammatical person, tense-mood-aspect) (cf. Maldonado 1999).<sup>1</sup> However, as shown in Table 4, *se*-marking is most likely to occur in a particular 'enter-exit a vehicle' *subir-bajar* construction (3), where vehicles include automobiles, buses, trains, airplanes, and carriages, as in (4). The rate of *se*-marking in this specific construction is significantly higher than the average for the corresponding variety, from 26% in literary data to 83% in Mexican oral data. Table 4 shows the variable-rule analysis of the linguistic factors

contributing to the choice of the *se*-marked form.<sup>2</sup> The closer the weight to 1, the more likely the *se*-marked form is to occur in the environment of the given factor.

	Weight	% <i>se</i>	Total N	% data
<b>Construction</b>				
Enter-exit vehicle	.78	50%	125	19%
Other uses	.42	14%	521	81%
<b>Co-occurring locative preposition</b>				
<i>a</i> 'to', <i>de</i> 'from', <i>en</i> 'on'	.67	33%	172	28%
Other or none	.43	16%	439	72%
<b>Subject</b>				
1 <sup>st</sup> and 2 <sup>nd</sup> person	.58	31%	204	33%
3 <sup>rd</sup> and Non-specific	.46	16%	415	67%
<b>Tense</b>				
Past	.55	27%	208	40%
Non-past	.47	19%	309	60%
Total N = 646, p = 0.049, Input .17 (21% <i>se</i> -marked)				

**Table 4. Variable-rule analysis of linguistic factors contributing to the choice of *se*-marked *subir-bajar***

(3) *Subir-bajar* enter-exit vehicle construction

$$\left\{ \begin{array}{l} \textit{subir} \\ \textit{bajar} \end{array} \right\} + SE + \left\{ \begin{array}{l} \textit{a} \textit{'to'} / \textit{en} \textit{'in'} \\ \textit{de} \textit{'from'} / \textit{en} \textit{'in'} \end{array} \right\} + \textit{'Vehicle'}$$

(4) ME **bajaba** del camión (Lope Blanch 1976, p.48)  
'I SE would **get off** the bus'

A number of other particular constructions show evident *se*-marking tendencies, for example tree climbing (5) tends to be *se*-marked whereas stair climbing (6) is mostly non-*se*-marked.

- (5) Las higueras yo creo que son los árboles donde más fácil SE **sube** uno (COREC, CCON034A)  
 ‘Fig trees I think are the trees one most easily SE **climbs**’
- (6) Rodrigo **subió** los peldaños de azulejo hasta su pieza interior (Real Academia Española n.d., CORDE, Fuentes, La región más transparente)  
 ‘Rodrigo Ø **went up** the tiled steps to his interior room’

These examples show that as *se* spreads in its usage with the movement verbs, *subir* and *bajar*, at least part of its spread is to particular local environments; instead of seeing the spread as generalization along abstract dimensions, we see instead the movement towards very specific local attractors. In the next section we will see further evidence for the importance of recognizing more local and specific conditions of use.

## 5. The interaction of the particular and the general in grammaticization

Rather than viewing stored lexical or collocational units as something distinct from and perhaps peripheral to grammar in the traditional sense, specific units constitute important loci of grammatical development. Bybee & Torres Cacoullós (forthcoming) examine the role of conventionalized instantiations, or prefabs, in the grammaticization of *estar* ‘be (located)’ + *V-ndo* (Gerund), which has evolved from a locative in Old Spanish toward becoming an obligatory expression of progressive aspect in present-day varieties (Torres Cacoullós 2000, forthcoming). The evidence shows that prefabs maintain associations with the general construction, advancing formal and semantic change. First, prefabs are ahead of the general construction in unithood status in early stages and thus demote the independent lexical status of the emerging gram. Second, in their association with semantic classes of which they are the most frequent member, prefabs promote the productivity of the general construction.

An index of grammaticization is degree of unithood. As Bybee (2003) proposed, frequent multi-word sequences become automated as single processing units, gaining autonomy in two ways. First, the erstwhile individual constituents of the frequent collocation weaken their association with other instances of the same constituents. For example, as future expression (*be*) *going to* reduces to *gonna*, *going* loses its association with the uses of *go* as a verb of motion (Bybee 2003, p.618). Second, the collocation dissociates from other instances of the construction. For example, (*be*) *going to* weakens its association with the more general motion-purposive schema, as in *I am going/traveling/riding to see the king* (Bybee 2003, p.603).

Unithood for *estar* + *V-ndo* may be measured by the following (Torres Cacoullos 2000):

- a. Adjacency (absence vs. presence of material separating *estar* and the gerund);
- b. Association (a single gerund is uniquely associated with *estar* vs. multiple juxtaposed gerunds detract from such an association);
- c. Fusion (as indicated by clitics appearing as proclitics on *estar* (“clitic climbing”) vs. as enclitics on the gerund).

Table 5 shows a diachronic increase in adjacency, association and fusion for *estar* + *V-ndo*. The bottom row shows a cumulative “unithood index”.<sup>3</sup>

	13 <sup>th</sup> c.		15 <sup>th</sup> c.		17 <sup>th</sup> c.		19 <sup>th</sup> c.	
	N	%	N	%	N	%	N	%
Adjacency	37/104	36	67/134	50	145/217	67	169/217	78
Association	83/104	80	115/134	86	192/217	88	199/217	92
Fusion	15/24	63	11/22	50	61/74	82	54/77	70
Unithood Index	.60 (62.2/104)		.74 (99.33/134)		.79 (172.5/217)		.83 (180.75/217)	

Adjacency: 13<sup>th</sup> vs. 15<sup>th</sup> Chi-square = 4.951, p=0.0261; 15<sup>th</sup> vs. 17<sup>th</sup> Chi-square = 9.7991, p = 0.0017; 17<sup>th</sup> vs. 19<sup>th</sup> Chi-square = 6.6344, p = 0.01

Association: 13<sup>th</sup> vs. 19<sup>th</sup> Chi-square = 9.3237, p = 0.0023

Fusion: Combined 13<sup>th</sup>-15<sup>th</sup> vs. 17<sup>th</sup>-19<sup>th</sup> Chi-square = 6.6827, p = 0.0097

**Table 5. Grammaticization (unithood) measures for *estar* + *V-ndo*: Adjacency, Association, Fusion**

Prefabs are operationally defined as collocations making up 2% or more of the *estar* data and 50% or more of the gerund data. The latter relative frequency measure is important, since other location-postural or motion verbs, such as *andar* ‘go around’, *ir* ‘go’ and *quedar* ‘remain’, variably combine with the same gerund. For example, *estar hablando* ‘be talking’ makes up 5% (32/672) of *estar* data and 71% (32/45) of *hablando* data. Table 6 compares *estar hablando* and *estar esperando* to all tokens of *estar* + *V-ndo* in Old Spanish data, and shows that these two prefabs score higher than the general construction on the unithood measures. This result provides evidence that frequent collocations become automated as single processing units (Bybee 2003). Prefabs contribute to grammaticization precisely because they are accessed holistically, effacing the erstwhile individual components of the collocation. As part of a prefab, originally locative *estar* contributes less

meaning individually, which promotes the semantic bleaching of this emerging auxiliary.<sup>4</sup>

Century	Prefab <i>estar</i> +		General <i>estar</i> + V- <i>ndo</i> (from Table 5)
	<i>hablando</i>	<i>esperando</i>	
13 <sup>th</sup>	.67 (8/12)	.82 (4.9/6)	.60 (62.2/104)
15 <sup>th</sup>	.89 (8/9)	.74 (13.3/18)	.74 (99.33/134)

**Table 6. *Estar* + V-*ndo* prefabs in Old Spanish: Comparison of unithood (grammaticization) indices**

Given the relatively autonomous representation of high frequency collocations in their own exemplar clusters, how may these prefabs contribute to the productivity of a general grammatical construction? In exemplar models, the representations of general constructions are built up from all the exemplars registering specific instances via associations among them based on semantic and formal similarity. The grammaticization of *estar* + V-*ndo* illustrates how prefabs contribute to productivity via the semantic classes centered around them.

A striking difference between *estar hablando* and *estar esperando* is that the former participates in a large class of semantically related verbs of speech (e.g. *alabar* ‘praise’, *demandar* ‘request’, *explicar* ‘explain’, *gritar* ‘shout’, *murmurar* ‘murmur’, *razonar* ‘argue’, *rogar* ‘beg, pray’), whereas the class of verbs of ‘waiting’ is tiny, including only *aguardar* and *atender* besides *esperar*. Since *estar hablando* is not only of high frequency but is associated with a semantic class with many members appearing in the *estar* + V-*ndo* configuration, the prediction of an exemplar model, with associations between exemplars based on semantic and formal similarity, is that the contribution of this prefab to the development of a general *estar* + V-*ndo* construction will be greater than that of *estar esperando*. This is because participation in high type frequency categories, as in the case of *estar hablando*, contributes to a more general schema and thus greater productivity (Bybee & Eddington 2006, Torres Cacoullós 2000, p.130).

The exemplar model prediction is supported. First, no other auxiliary competes with *estar*’s association with *hablando*. In contrast, though *esperando* is still among the top ten or so gerunds combining with *estar* in present-day (COREC) data, its exclusive association has eroded: while in Old Spanish 100% (24/24) of *esperando* tokens co-occurred with *estar*, beginning with 17<sup>th</sup> century data *quedar* ‘remain’ variably combines with this gerund.<sup>5</sup> Second, verbs of speech make up a steady proportion (about one-sixth) of occurrences of *estar* + V-*ndo*, whereas the relative frequency of ‘waiting’ verbs declines from 11% (11/104) in the 13<sup>th</sup> century to 3% in 19<sup>th</sup> century (6/207) data.

Thus, prefabricated instances of constructions lead in the generalization of the construction as well as in manifesting structural indices of unithood. As predicted from their relative frequency of use, prefabs grammaticize earlier or at a faster rate than the general construction. At the same time, they serve as the centers of subclasses of the grammaticizing construction, attracting more lexical types and thereby contributing to the productivity of the construction. The interaction between the particular and the general in grammatical variation and change is modeled with exemplars and the associations among them.

## 6. Usage-based constructions and social meaning

In exemplar models, “grammar is the cognitive organization of [speakers’] experience with language” (Bybee 2006, p.711). Thus, constructions – form-meaning pairings – are usage-based; that is, they represent specific lexical items and their social and pragmatic contexts of use.

The most striking evidence of the association of social categories with linguistic ones has come from research into sound change beginning in the mid-1960s (Labov 1966) and more recently from sociophonetics, in which it is shown that both children and adults make strong associations of subphonemic variants with social categories, including gender and class (Foulkes & Docherty 2006). The example in (7) suggests a social value as applied to a morphosyntactic construction. *Andar* + *V-ndo* is in variation with *estar* + *V-ndo* in Mexican Spanish varieties as an expression of progressive and other imperfective aspects.

- (7) a. *estás hablando de una forma de vida* (Lope Blanch 1971, p.261)  
 ‘you are talking about a way of life’  
 b. *ando buscando unas tijeras* (Lope Blanch 1976, p.415)  
 ‘I am looking for scissors’

Speakers’ choice between the variant gerund constructions is both linguistically and socially conditioned, with each auxiliary favored by particular verbs and semantic classes of verbs and *andar* more likely to be used in popular varieties. In tracing the grammaticization of the two constructions, Torres Cacoullós (2001) shows that the social and linguistic associations are related.

Rather than aspectual differences, present-day variation reflects collocational routines. For example, variable-rule analysis shows that *estar* is more likely with *hablando* and other verbs of speech, as well as verbs denoting perceptible bodily activities (*llorando* ‘crying’) and mental activities (*pensando* ‘thinking’), while *andar buscando* is the conventional way to ‘be looking for’ something. Retention of the source-construction ‘going around’ meaning in *andar* + *V-ndo* is reflected in the tendency to occur with verbs denoting motion and physical activities, particularly outdoor activities, such as *dando la vuelta* ‘going around, strolling around, driving

around' and *trabajando* 'working' in the fields, since rural activities in large spaces are more compatible with the original meaning of *andar*. Torres Cacoullos (2001) suggests that the social stratification of *estar/andar* + V-ndo may originate as an indoor/outdoor, urban/rural difference. In an exemplar model, representations of general constructions are built up from exemplars which include real-world information about speakers and situations.

## 7. Conclusion

Usage-based grammar, coupled with the rich memory representations of exemplar models, allows a direct representation of variation and change in cognitive representations. Because it tracks frequency of co-occurrence, this model provides us with ways of explaining how new constructions arise gradually over time, how specific instances of constructions contribute to more general change and how specific instances take on special value, be it semantic or social. Applied at the phonological level, exemplars help us track the lexical diffusion of sound change and provide a direct representation of the effect that frequency has on reductive sound change.

These examples show clearly that variation in language use is pervasive, reaching the lexical, semantic, pragmatic and social domains, and thus cannot be tacked on to some posited invariant structure or representation in the form of rules or ordered constraints (see Pierrehumbert 1994 and many others). These examples also show how usage and context impact language structures. We conclude, then, that our models of language structure must be sensitive to instances of use in context with all their variation as well as the rich nuances they evoke on every level.

## Notes

- 1 Variable-rule analysis, the goal of which is to discover the set of factor groups that jointly account for the largest amount of variation in a statistically significant way, does not (necessarily) involve 'rules' and is independent of assumptions about underlying forms (Sankoff 1988, p.984).
- 2 Medium also contributes a significant effect, with the *se*-marked form more likely in oral (.71) than written (.34) varieties.
- 3 The Unithood index in Table 5 is calculated as a fraction (the point total for all tokens divided by the number of tokens). Points for each token are based on Adjacency (two points for no intervening material; one for an intervening subject, object, temporal or manner expression; zero for an intervening adjective, locative or more than one of the above), Association (one point for a single as opposed to multiple gerunds), and Fusion (one point for a proclitic as opposed to enclitic).

- 4 Over time, as the productivity of the general construction increases, *estar hablando* makes up a smaller portion of the data, from 12% (12/104) of all *estar* + V-*ndo* tokens in the 13<sup>th</sup> century to 2% (5/217) in the 19<sup>th</sup> century, and appears to follow general patterns.
- 5 *Quedar esperando* becomes a prefab in its own right, making up 12% (5/43) of all *quedar* + V-*ndo* tokens.

## References

- Bills, Garland & Neddy Vigil. 1999. Ashes to ashes: The historical basis for dialect variation in New Mexican Spanish. *Romance Philology* 53, 43-66.
- Brown, Esther L. 2004. Reduction of syllable-initial /s/ in the Spanish of New Mexico and Southern Colorado: A usage-based approach. Doctoral dissertation, University of New Mexico.
- Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.
- Bybee, Joan. 1988. Morphology as lexical organization. In Michael Hammond & Michael Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 119-141. New York: Academic Press.
- Bybee, Joan. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65-85. Stanford: CSLI.
- Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
- Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14, 261-290.
- Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian Joseph & Richard Janda (eds.), *Handbook of historical linguistics*, 602-623. Oxford: Blackwell.
- Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82, 711-733.
- Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82, 323-355.
- Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37, 575-596.
- Bybee, Joan & Rena Torres Cacoullous. Forthcoming. The role of prefabs in grammaticization: How the particular and the general interact in language change. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Proceedings of the 2007 University of Wisconsin-Milwaukee Symposium on Formulaic Language*. Amsterdam: John Benjamins.

- Erman, Britt, & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20, 29-62.
- Foulkes, Gerard & Paul Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34, 409-438.
- Goldinger, Stephen D. 1996. Word and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology* 22, 1166-1183.
- Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William Christie (ed.), *Current progress in historical linguistics*, 96-105. Amsterdam: North Holland.
- Hooper, Joan B. 1981. The empirical determination of phonological representations. In Terry Myers, John Laver & John Anderson (eds.), *The cognitive representation of speech*, 347-357. Amsterdam: North Holland.
- Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.
- Labov, William. 1966. *The social stratification of English in New York City*. Arlington, VA: Center for Applied Linguistics.
- Lope Blanch, Juan M. (ed.). 1971. *El habla de la ciudad de México: Materiales para su estudio*. Mexico City: Universidad Nacional Autónoma de México.
- Lope Blanch, Juan M. (ed.). 1976. *El habla popular de México: Materiales para su estudio*. Mexico City: Universidad Nacional Autónoma de México.
- Marcos Marín, Francisco (director). 1992. *Corpus oral de referencia de la lengua española contemporánea (COREC)*. Textual corpus. Madrid: Universidad Autónoma de Madrid. Available on-line:  
<http://www.lllf.uam.es/~fmarcos/informes/corpus/corpulee.html>
- Maldonado, Ricardo, 1999. *A media voz: Problemas conceptuales del clítico se*. Mexico City: Universidad Nacional Autónoma de México.
- Moonwomon, Birch. 1992. The mechanism of lexical diffusion. Paper presented at the Annual Meeting of the Linguistic Society of America, Philadelphia.
- Phillips, Betty S. 1984. Word frequency and the actuation of sound change. *Language* 60, 320-342.
- Phillips, Betty S. 2001. Lexical diffusion, lexical frequency, and lexical analysis. In Joan Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 123-136. Amsterdam: John Benjamins.
- Pierrehumbert, Janet. 1994. Knowledge of variation. In Katharine Beals, Jeannette Denton, Robert Knippen, Lynette Melnar, Hisami Suzuki & Erica Zeinfeld (eds.), *CLS 30, Vol. 2: Papers from the parasession on variation in linguistic theory*, 232-256. Chicago: Chicago Linguistic Society.
- Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137-58. Amsterdam: John Benjamins.

- Pierrehumbert, Janet. 2003. Phonectic diversity, statistical learning and acquisition of phonology. *Language and Speech* 46, 115-154.
- Poplack, Shana. 1992. The inherent variability of the French subjunctive. In Christiane Laeufer & Terrell A. Morgan (eds.), *Theoretical analyses in Romance linguistics*, 235-263. Amsterdam: John Benjamins.
- Poplack, Shana & Sali Tagliamonte. 2001. *African American English in the diaspora: Tense and aspect*. Oxford: Blackwell.
- Real Academia Española. n.d. Corpus diacrónico del español (CORDE). <http://corpus.rae.es/cordenet.html>
- Sankoff, David. 1988. Variable rules. In Ulrich Ammon, Norbert Dittmar & Klaus J. Mattheier (eds.), *Sociolinguistics: An international handbook of the science of language and society*, 984-997. Berlin: Mouton de Gruyter.
- Sankoff, David, Sali A. Tagliamonte & Eric Smith. 2005. Goldvarb X: A multivariate analysis application for Macintosh and Windows. [http://individual.utoronto.ca/tagliamonte/Goldvarb/GV\\_index.htm](http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm)
- Thompson, Sandra A. 2002. 'Object complements' and conversation. *Studies in Language* 26, 125-164.
- 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, 443-478.
- Torres Cacoullos, Rena. Forthcoming. La evolución de *estar* + *V-ndo* hacia expresión obligatoria del aspecto progresivo en español. In Concepción Company Company & José G. Moreno de Alba (eds.), *Actas del VII Congreso Internacional de Historia de la Lengua Española*. Madrid: Arco Libros.
- Torres Cacoullos, Rena & Scott A. Schwenter. Forthcoming. Constructions and pragmatics: Variable middle marking in Spanish *subir(se)* 'go up' and *bajar(se)* 'go down'. *Journal of Pragmatics*.
- Torres Cacoullos, Rena & James A. Walker. Forthcoming. On the persistence of grammar in discourse: A variationist study of *that*. *Linguistics*.
- Torres Cacoullos, Rena, and Walker, James A. 2008. The present of the future: Discourse variation and the bounds of grammaticization. Manuscript.
- Wedel, Andrew. 2007. Feedback and regularity in the lexicon. *Phonology* 24, 147-185.

Copyright of Studies in Hispanic & Lusophone Linguistics is the property of Studies in Hispanic & Lusophone Linguistics and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.