Language Variation and Change, **24** (2012), 317–341.

doi:10.1017/S0954394512000154

Grammatical relation probability: How usage patterns shape analogy

ESTHER L. BROWN AND JAVIER RIVAS
University of Colorado at Boulder

ABSTRACT

It has been argued speakers' knowledge of the probabilities of certain phones, words, and syntactic structures affects language production (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Tily, Gahl, Arnon, Snider, Kothari, & Bresnan, 2009). This study provides evidence for effects of grammatical relation probabilities by identifying significant effects on verb morphology in the Spanish presentative [*haber* 'there (be)'+ NP] construction stemming from nouns with varying proportion of use in subject function. In addition to this novel type of probability (grammatical relation), we present calculations that are not context-dependent but cumulative, reflecting speakers' overall experience with these nouns in the grammar. We conduct variationist analyses on corpora of spoken Puerto Rican Spanish. Our results reveal that nouns with a high probability of subject function promote the analogical leveling of *haber* by increasing the likelihood of reanalysis of the object as subject of the construction. We interpret these results as suggesting speakers possess lexicalized knowledge of grammatical relation usage patterns.

Usage-based studies maintain that language use shapes language structure. Speakers possess fine-grained, detailed knowledge of forms, meanings, and contexts of use of linguistic units (Pierrehumbert, 2001), and such knowledge affects acquisition, usage patterns, and language change. Findings from research into patterns of use for words, as well as for combinations of words in constructions, allow us to view the lexicon and grammar not as two separate entities, but rather as highly intertwined (e.g., Beckner, Blythe, Bybee, Christiansen, Croft, Ellis, Holland, Ke, Larsen-Freeman, & Schoenamann, 2009; Bybee, 2001; Chang, Dell, & Bock, 2006; Goldberg, 1995; Langacker, 1987). Representations of words in memory, thus, include lexical information regarding phonological shape and semantic content, but also facts regarding contextual discourse patterns (Bybee, 2002), and this knowledge, in turn, affects use.

A linguistic factor known to play a determinative role in language variation and change is probabilistic linguistic knowledge. For instance, it has been shown that speakers' phonotactic knowledge reflects more than only licit and illicit phone combinations, but also exhibits emergent probabilistic knowledge of likely and unlikely phone-to-phone transitions or combinations within one's language (Frisch, Large, & Pisoni, 2000). This type of phone bigram probability affects language processing and language production (Raymond, Dautricourt, & Hume, 2006), where predictable forms reduce more readily. Similarly, transitional

probability effects are noted for words (Bush, 2001; Jurafsky, Bell, Gregory, & Raymond, 2001; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), where words with high predicted probability from their context exhibit increased phonological reduction. Such variation in phonological production suggests probabilistic knowledge may be stored lexically for phones and words.

Gahl and Garnsey (2004:748) demonstrated, however, that "word-to-word or sound-to-sound probabilities" are not the only examples of probabilistic knowledge to form part of grammar. These authors argued, based on an analysis of subcategorization-based probabilities in English *–t/d* deletion, that speakers possess knowledge of syntactic probabilities. Gahl and Garnsey (2004) found that rates of word final *–t/d* deletion in verbs reflect the probability that a certain syntactic structure (i.e., direct object versus sentential complement) will follow each verb. Verbs followed by syntactic structures of high probability are more likely to exhibit final *–t/d* deletion than are verbs followed by a syntactic structure of low probability. As such, in addition to frequency and probabilistic measures of words and phones, Gahl and Garnsey (2004) and subsequently Tily, Gahl, Arnon, Snider, Kothari, and Bresnan (2009) established that probabilistic knowledge of abstract grammatical structure and categories form part of speakers' grammar.

This evidence for the role of verb biases in pronunciation variation implicates the lexicon and supports the notion that lexical organization into exemplars and exemplar clusters reflects not just linguistic form and meaning, but also contexts of use (Beckner et al., 2009; Bybee, 2002; Pierrehumbert, 2001). That is, lexical representation of verbs includes the probability of use with syntactic patterns including nominal direct objects or clausal complements. Such knowledge is derived through use and reflects a speaker's experience with language (Bybee, 2010). Are other syntactic probabilities stored lexically?

This current analysis extends Gahl and Garnsey's (2004) finding to examine noun phrase (NP) syntactic patterns. The probabilities considered here are based on grammatical relation probabilities for the noun: the likelihood that each noun will act as a subject of a sentence in overall usage. The role of cumulative use with specific grammatical relations (e.g., subject, object) has not been tested in production. This study uses variationist methodology (Poplack & Meechan, 1998; Poplack & Tagliamonte, 2001) and data from corpora of spoken Puerto Rican Spanish to examine a case of morphosyntactic variation concerning the presentative verb *haber* ('there (be)') that entails a process of regularization by analogy. Results of this current study provide evidence for lexicalized noun grammatical relation-based probabilities. Therefore, the implications of this work are twofold. Specifically, this work contributes a new perspective and innovative methodology to the widely studied problem in Hispanic linguistics of *haber* 'there (be)' regularization in Spanish. Also, importantly, findings in this work contribute to the development of usage-based models of language variation and change by testing new types of probabilistic knowledge and probability measurements.

The present paper is organized as follows. We first detail previous research that informs our discussion of syntactic probability and analogical processes. We next summarize salient aspects of research on *haber* regularization in Spanish that is the source of the data for the current analysis. We then discuss data and methods, followed by results of our quantitative and variable rule analyses. Lastly, we present a discussion of the theoretical and methodological ramifications of the findings.

BACKGROUND

Mechanisms of morphosyntactic change

Studies on morphosyntactic change (Campbell, 1998; Harris & Campbell, 1995; Hopper & Traugott, 2003) have identified two major mechanisms through which morphosyntactic change takes place: analogy and reanalysis.[1] Most examples of analogy are cases of analogical leveling. In analogical leveling an alternation in a paradigm is lost, which yields more uniform paradigms. For this reason, analogical leveling is very often associated with processes of regularization in which unusual (and irregular) forms are replaced by more common (regular) forms. For example, the Spanish verb *cocer* 'to cook' had a strong/irregular preterit form *coxe* 'I cooked', which eventually disappeared, and this verb now follows the regular verb pattern (*cocí* 'I cooked').

Traditionally, analogy is described as an arbitrary process; it unpredictably applies to some forms and not others. Bybee (2001, 2010), however, showed that analogical change is not arbitrary but is generally based on phonological or semantic similarity with existing forms.[2] For example, throughout the history of English, some originally regular verbs have developed irregular forms: *sling-slung, sting-stung, string-strung, fling-flung, hang-hung* (Bybee, 2001:127). These forms arise by analogical extension because of the (semiproductive) pattern of alternation seen in verbs such as *swim-swam, ring-rang*, and *drink-drank*. All these forms share a nasal or velar-final infinitive. Similarly, all the new members of this paradigm (*sling, sting, string, fling, hang*) also end in a velar sound. In this example, analogical extension arises through a "gang effect" (Bybee, 2001:13). This "gang" is established because of the high phonological similarity of the members of the paradigm, and a word's probability of participating in this analogical change (*sling*, for example) is driven by the phonological shape of the word.

An example in which analogical change extends via semantic similarity of new forms with existing forms is seen in the resultative construction SUBJECT + [DRIVE] [X (usually *me*)] *mad* (Bybee, 2010:58). Instead of *mad*, other adjectives or prepositional phrases may occur in the same syntactic slot, but all of them (*crazy*, *up the wall*, *nuts*) are semantically related. As a result, *happy* is very unlikely to occur in this position, because it is a semantic opposite of *mad*.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

Thus, similar to the effect of phonological similarity, the probability of a word occurring in the position of *mad* is determined by similarity (semantic in this case) to existing forms.

Recent studies (Bybee, 2010; Bybee & Torres Cacoullos, 2009; Torres Cacoullos & Walker, 2009; Vergara Wilson, 2009) have also shown that the distribution of lexical items in grammaticalizing constructions is skewed; that is to say, not all the examples of the construction become grammaticalized at the same pace. Rather, grammaticalization processes are promoted by conventionalized combinations of words, also called *prefabs*.[3] As Bybee (2010:35) pointed out, prefabs may be regarded as *chunks* or multiword strings that are created through repetition and acquired lexical strength through their frequency of use. Chunks are stored in the lexicon as any morphologically complex word would be and therefore, may be accessed holistically. As a result, they lose compositionality, which contributes to the semantic bleaching of their constituent parts, which in turn advances the whole grammaticalization process.[4] In addition, prefabs also contribute to the productivity of the grammaticalizing construction because they may attract other verbs in the same semantic class to the construction (e.g., Bybee & Torres Cacoullos, 2009).

Previous usage-based studies, therefore, have shown how processes of language change are shaped by the speaker's experience with language. At the morphological level, innovative forms are very often created on the basis of phonological similarity with preexisting forms, whereas at the morphosyntatic-constructional level, grammaticalization processes are led by prefabs, multiword strings that are accessed holistically by the speaker and that contribute to the general productivity of the grammaticalizing construction by attracting other semantically similar linguistic forms to the same construction. In this study, we will use an example of a grammatical variable taken from Spanish: the pluralization (regularization) of the presentative verb *haber* 'there (be)' in the [*haber* + NP] construction. Results show that the probability of use with a specific grammatical relation (i.e., subject) for each noun shapes the regularization process of *haber*. Thus, an important contribution of this study is to show that syntactic probability, and not just phonological and semantic similarity, can shape analogical processes.

An example of an analogical process: pluralization of existential haber *in Spanish*

In Spanish, *haber* 'there (be)' is a one-argument verb that has a presentative function. The sole argument of this verb or *presentatum* typically occurs in postverbal position and introduces new information in the discourse (Ashby & Bentivoglio, 1997:16), that is, information that is inactive in the hearer's mind (Chafe, 1987). Consider the following excerpt:

(1) Interview 1, 9
I: *A Utuado fui en, . . . hombre hace tiempo que no voy.* ***Y este año no hubo fiestas patronales*** *debido a, el alcalde de Utuado dijo debido y que a los sucesos del 11 de, de septiembre no hubo fiestas patronales en Utuado, yo no*

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

sé qué tiene que ver las twin towers *con Utuado, pero* anyhow.
I: I went to Utuado in . . . well, it's been a while since I've been there. And this year there weren't celebrations [literally: there wasn't celebrations] of their patron saint's holiday, because of, the mayor of Utuado said, because of the events of September 11, they didn't celebrate their patron saint's holiday in Utuado, I don't understand what the twin towers have to do with Utuado, but anyhow.

In (1), the preterit form of *haber* is followed by its sole argument, the NP *fiestas patronales*, which introduces new information in the discourse. Traditionally, the sole argument of *haber* is regarded as the direct object of the construction, because it does not trigger verbal concord [verb (third-person singular), NP (plural)] and it may be replaced by a direct object clitic (the accusative feminine singular pronoun *las*: *no las hubo*). *Haber* is in this way regarded as an impersonal or unipersonal verb (Alarcos Llorach, 1994).

However, in many varieties of Spanish, including Puerto Rican Spanish, this construction coexists with another one in which the verb *haber* agrees in person and number with its sole argument, as (2) illustrates:

(2) Interview 16, 80
M:*Hubieron fi-,* ***hubieron fiestas*** *en todos los pueblos menos en ése*.
M: There were celebrations of their patron saint's holiday in all towns except in that one.

In (2), the verb *haber* occurs in the third-person plural past tense (*hubieron*), in agreement with its sole argument (*fiestas*), which is also in the third-person plural. In cases such as these, it may be argued that the *presentatum* is actually the subject of the construction, since it displays agreement with the verb. In this respect, *haber* behaves like other presentational verbs (e.g., *existir* 'to exist' and *ser* 'to be').

In present-day Spanish, then, both agreement and nonagreement are possible with *haber*, sometimes even in the speech of the same speaker, as (3) illustrates. *Haber* is used three times in (3) with the same plural noun (*accidentes* 'accidents'), twice in singular form (*hubo*, *ha habido*) and once in the plural (*hubieron*).

(3) Interview 1, 3
I: ***Aquí hubo muchos accidentes,*** *a pesar de que ha habido menos muertes de, de, de accidentes de automóviles,* ***últimamente ha habido unos accidentes que son-, el mismo día hubieron tres accidentes***.
I: Here there were a lot of accidents, even though there have been fewer casualties in car accidents. Lately there have been some accidents that are . . . one and the same day there were three accidents.

According to Fontanella de Weinberg (1992) and Hernández Díaz (2006), the variable *haber* agreement that is attested in present-day Spanish forms part of a

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

larger diachronic change concerning the verb *haber,* which, from a possessive-transitive verb, became a presentational-impersonal verb, and later a presentational-intransitive verb. In Latin, *habēre* 'to have' was a transitive verb that was used to convey possession. This construction was transferred from Latin to Spanish and it survived until the 17th century (Fontanella de Weinberg, 1992:38) and in some fixed expressions until the 19th century. In addition, late Latin developed a new construction in which *habēre* always occurred in the third-person singular and was followed by an NP in the accusative case. This construction gave rise to the presentational-impersonal uses of *haber*. Later on, in the 18th century, the presentational-intransitive construction arose, in which the verb *haber* agrees in person and number with its *presentatum*.[5] As Fontanella de Weinberg (1992:39) pointed out, this construction became more and more common in the 19th and 20th centuries. The presentational-intransitive construction is therefore the more innovative of the two. Both the presentational-impersonal construction and the presentational-intransitive construction co-occur in present-day Spanish.

D'Aquino Ruiz (2008), Díaz-Campos (2003), Hernández Díaz (2006), and Montes de Oca-Sicilia (1994) maintained that the presentative *haber* construction is undergoing a change in progress whereby the construction in which *haber* agrees with its sole argument is likely to eventually replace the impersonal construction. However, other researchers such as Quintanilla-Aguilar (2009) argued that there is not enough evidence (in data from El Salvador) to determine that *haber* regularization represents a change in progress. With our current data, we cannot argue for or against change in progress as opposed to stable variation. Therefore, we will discuss this phenomenon as a case of linguistic variation in present-day Spanish, even though the regularized construction is more innovative than the impersonal construction is.

On this basis, and in line with Waltereit and Detges (2008:26), we argue that the presentative *haber* construction has undergone a process of reanalysis through which the sole argument of *haber* is interpreted as the subject of the construction. This process of reanalysis is possible because of the existence of constructions such as (4), in which both *haber* and the NP argument occur in the singular form. Constructions such as this are open to both interpretations: the NP argument may be regarded as the subject or the direct object of the construction:

(4) Interview 1, 20
E: ¿***Había agua caliente,*** J.?
E: Was there any hot water, J.?

The type of reanalysis that *haber* has undergone in this construction entails a change in the grammatical relation of its sole NP argument, that is, the direct object is reanalyzed as the subject of the construction.[6] The existence of constructions such as (5) in which *haber* occurs in the plural, in agreement with its *presentatum*, indicates that reanalysis has taken place:

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

(5) Interview 15, 157
I: *No, pero* ***habían muchas hormigas*** *y esas hormigas son de las que pican bueno*.
I: No, but there were a lot of ants and they were of the kind that really bites.

The *haber* example in this excerpt shows that, by analogy, the NP argument-verb agreement that we have in (4) is also extended to those contexts in which the NP argument occurs in the plural, as in (5). As is noted by Montes de Oca-Sicilia (1994:16), *haber* undergoes a process of analogical regularization that, by displaying agreement with its subject, levels *haber* with other Spanish verbs. In this study, we will examine the way in which this analogical extension varies, and we will demonstrate that syntactic frequencies of the nouns play a key role.

A process such as the regularization of *haber* in Spanish, whereby plural conjugated forms co-occur with plural NPs, can aptly be described as a case of analogy. A minor construction, the impersonal construction [*haber* + NP], is being replaced by a major construction, the intransitive construction, in which there is person/number agreement between the verb and its sole NP argument, and the plural forms of *haber* are modeled off of existing forms in the inflectional verbal paradigm of Spanish. However, the regularization of *haber* is variable; not all plural NPs co-occur with plural forms of *haber*. There is, in other words, variation. What accounts for this variation?

Usage-based theories have shown that how we use language shapes language structure. We hypothesized that, as part of this regularization process, nouns would play an influential role. Specifically, we hypothesized that nouns typically used with a subject function, and thus typically agreeing with the verb in person and number, would be more apt to promote regularization of *haber* than would nouns that do not typically agree with the verb. Conversely, nouns with a low probability of serving as the subject of a sentence would be less likely to promote *haber* regularization. For instance, in the oral section of Davies (2002–), *estudiantes* 'students' is used with subject function in 55% of the 973 occurrences compared with *directores* 'managers', which has a subject function in only 9% of its 742 occurrences. Our hypothesis would therefore predict that *estudiantes* 'students', which has a high probability of appearing in subject function, is more likely to trigger *haber* regularization than is *directores* 'managers', which presents a very low probability of functioning as subject.

DATA AND METHODS

A primary goal of this study is to demonstrate a statistically significant effect of a nontested notion of grammatical relation probability. To show this effect with a corpus-based approach and statistical model, we also include linguistic factors that have been previously considered for analysis with the variable under study. If a statistically significant effect of this new cumulative, probabilistic measure is found, while controlling other independent factors known to constrain variable

agreement of *haber*, then the results bring new evidence to the study of probabilistic measures.

To determine the linguistic factors that contribute to the regularization of *haber* in Puerto Rican Spanish, we conduct a quantitative analysis in two separate corpora of spoken Puerto Rican Spanish.[7] The first corpus (Cortés-Torres, 2005), henceforth CT, contains approximately 370,000 words of spoken Puerto Rican Spanish representing roughly 27 hours of conversational data from 33 native speakers. These conversations were collected and transcribed by a native speaker in Caguas, Cayey, and San Juan, Puerto Rico in 2000 (Cortés-Torres, 2005). Speakers range in age from 24 to 90 years. Interviews range in duration from one-half hour to three hours and represent sociolinguistic interviews. The second corpus is the *Habla Culta: San Juan* data from the *Corpus del español* (Davies, 2002–), henceforth CE. This second corpus contains approximately 200,000 words of spoken Puerto Rican Spanish. The data in the *Habla Culta: San Juan* portion represent the speech of men and women 25 years and older. All speakers were born and/or raised from a young age in San Juan, lived in this city for at least three quarters of their life, and obtained a degree in higher education. Four types of data collection methodology were employed (DeMello, 1991:446n2): recorded conversations between interviewer and one or more informants, "free conversation" between two informants, secret recordings of spontaneous conversations, and formal language taken from a variety of contexts (e.g., speeches, lectures). The data was compiled throughout the 1970s.

We extract all cases of *haber* with presentational uses with plural nouns for a total of 97 examples from CT and 93 examples from CE. This excludes cases of present indicative form *hay* (1140 in CT and 635 in the CE corpus), which shows no variation in this variety of Spanish. Building on previous analyses of Spanish *haber* regularization and usage-based studies of variation, we code for 10 linguistic factors: (1) proportion of noun in subject function, (2) proportion of instances of noun use in the *[haber* + NP] construction, (3) verbal forms, (4) logarithm of word frequency per million, (5) polarity, (6) word order, (7) presence or absence of a quantifier, (8) definiteness of the NP, (9) human or nonhuman referent of the *presentatum*, (10) corpus.

Proportion of noun in subject function. Previous analyses (Bentivoglio & Sedano, 1989; D'Aquino, 2004; DeMello, 1991; Díaz-Campos, 1999–2000; Domínguez, Guzmán, Moros, Pabón, & Vilaín, 1998) alluded to the important potential role the noun plays in *haber* regularization. Linguistic factors attributed to the noun, and argued to constrain the occurrence of the innovative *haber* form, include animacy and definiteness of the noun argument, as well as the presence of a quantifier within the NP. Notwithstanding potential effects of these identified factors, we hypothesize that the information regarding syntactic usage patterns (grammatical relation probability) will constrain variation. This probabilistic information, which we will argue is lexicalized, can affect use of the [*haber* + NP] construction, the result being a tendency to promote verbal

agreement (*haber* regularization) for nouns with a high probability of use with subject function.

To test this, we devise a measure for the grammatical relation probability for each noun. Calculations of probabilistic measures, particularly measures of cumulative patterns of use such as the one we are testing, can be more reliably determined by examining multiple instances of use. For this reason, we base our calculations on the oral section of Davies (2002–), which includes Spanish varieties other than Puerto Rican and has approximately 5 million words. We use the lemma frequency calculation for each noun. For each noun, we determine the number of occurrences it has with subject function as determined by its use in context. For nouns with a textual frequency high enough to provide representation values in our CT corpus (arbitrarily set at greater than 500 tokens), we use the CT corpus as opposed to the CE corpus (Davies, 2002–). The number of noun examples used with subject function is divided by the total occurrences of the same noun overall in the corpus. The result is a proportion of noun instances that occur with subject function in the corpus, expressed as a percentage (similar to methods used in determining verb biases from large corpora). For example, the noun lemma *poeta* 'poet' occurs 322 times in the oral section of Davies (2002–). Of these, 75 instances of the noun are used with subject function (e.g., *Los poetas, sin embargo, siempre han sido más respetados* 'poets, however, have always been more respected' [entrevista ABC]). Thus, the noun *poeta* is given a value for its proportion of noun subject use of 23% (75 of 322).

This measure can be taken as the syntactic probability for each noun (independently of other semantic characteristics such as +/– human), and, importantly, this measure is distinct from the probability of occurring with or without a pluralized verb. This differentiates our measurement from other common measures of probability (Bock, 1986; Chang, Dell, & Bock, 2006; Jaeger, 2010; Szmrecsányi, 2005; Tily et al., 2009) in which the dependent variable (the predicted linguistic form) and the probabilistic measure predicting the linguistic outcome are derived from the same context, and thus such contextually dependent effects can be seen as occurring online.

In discretizing the continuous data, the tokens were sorted numerically lowest to highest with the goal of grouping the same number of tokens (approximately 63; one-third of the total 190) in each category (low, mid, high). Identical numerical values or lexical items were not sorted into distinct groups. Rather, in such a case, the division between groups was set at the closest previous or subsequent token in the sorted list with a different numerical value. Tokens with a proportion of noun in subject function falling in the bottom third are considered low frequency (the noun is used relatively infrequently with a subject function in discourse), those in the middle third are coded as medium frequency, and those in the highest third are considered high frequency. This same procedure for grouping continuous values was used for the proportion of instances of noun use in the *[haber* + NP] construction and log of word frequency factors.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

*Proportion of instances of noun use (generally in the language) in the [*haber + *NP] construction.* In cases of language variation, it has been shown that lexical representations of patterns are strengthened through high frequency. Such strong lexical representations make linguistic forms more resistant to regularizing patterns (Bybee, 1985). Thus, we code individual noun lexical frequencies in the [*haber* + NP] construction in order to test whether certain frequent [*haber* + NP] units could be stored holistically as chunks through repeated use. Usage-based approaches would lead us to hypothesize that the more often a noun occurs with *haber* (and its irregular nonagreement pattern), the less likely it would be to regularize in accord with the conserving effect of frequency of use (Bybee & Thompson, 1997).

To determine the strength of this linguistic factor in constraining variable agreement of *haber*, we use methods similar to those outlined for the proportion of noun as subject calculation. For each noun lemma, we calculate the number of occurrences with any form of presentational *haber* in the oral section of Davies (2002–). For nouns with a textual frequency high enough to provide representation values in our CT corpus (again, arbitrarily set at greater than 500 tokens), we use the CT corpus as opposed to the CE corpus. In line with previous methodological approaches to morphosyntactic constructions (see Bybee & Torres Cacoullos, 2009, for the Spanish progressive *estar* 'to be' + gerund), each instance of the noun in the [*haber* + NP] construction is counted together irrespective of the tense or aspect of the verb *haber* (e.g., *había*, *hubo*, *hay*) and without differentiating cases with and without lexical material intervening between the noun and *haber*. The number of noun examples used with a form of *haber* is divided by the total occurrences in the corpus. The result is a proportion of noun instances that occur with *haber* in the corpus, expressed as a percentage. For example, *poeta* 'poet' occurs 322 times in the oral Davies (2002–) corpus, and is used with *haber* 9 times (e.g., *pero hay poetas, fíjate, como la Gabriela Mistral que . . .* 'but there are poets, you know, like Gabriela Mistral who . . .' [Habla Culta: Santiago]). Thus, *poeta* has a proportion of noun use with *haber* of 3%. This measure can be taken as a gauge of the construction strength with the specific noun, or the likelihood that it is stored lexically as a chunk. Tokens with a percentage falling within the range in the bottom third of our data are considered low frequency (noun occurs relatively infrequently in the [*haber* + NP] construction), those in the middle third are coded as medium frequency, and those in the highest third are considered high frequency.

Verbal forms. Hernández Díaz (2006) maintained that compound forms of *haber* (e.g., *tienen que haber* 'there have to be', *han habido* 'there have been') are more likely to regularize than are simple forms of *haber* because in compound forms agreement is placed on the auxiliary verb and not directly on the *haber* form. However, results from previous quantitative studies do not support Hernández Díaz's predictions. For example, D'Aquino Ruiz (2004), in a quantitative analysis of *haber* regularization in Venezuelan Spanish, showed that compound forms disfavor the innovative form, whereas all simple forms of

haber favor pluralization except the preterit. This may be because, as is noted by Freites Barros (2004), the plural preterit form *hubieron* is stigmatized in Venezuelan Spanish. In contrast, Vaquero (1978) suggested the same is not true for Puerto Rican Spanish. Therefore, we code verbal forms into two groups: simple forms (imperfect, preterit, present subjunctive) and compound forms (modal uses, periphrastic future, and perfect tenses).

Log of word frequency per million. Previous usage-based studies (e.g., Bybee, 2001) have shown that lexical frequency affects rates of change. If the regularization pattern we find for *haber* in Spanish were propelled by a frequent noun (regardless of how often it appears as a subject or in a construction with *haber*), an analysis of noun word frequency could reveal this. Thus, for each noun, we find the lemma frequency per million in the oral Davies (2002–) corpus and use the log(10) of the frequencies in the analysis. Nouns in the bottom third of tokens are coded as low frequency, those in the middle third are coded as having medium frequency, and nouns in the highest third are coded as high frequency lexical items.

Polarity. The rationale for considering this factor is that negative clauses contain nonreferential NPs. As Du Bois (1980:208) pointed out, "a noun phrase is *referential* when it is used to speak about an object as an object, with continuous identity over time." Nonreferential NPs typically refer to "the quality defined by the noun, rather than the potential of the noun for concrete meaning" (Du Bois 1980:209). Among the major categories of nonreferential NPs, Du Bois mentions negative clauses. Because NPs within the negative scope in a clause are nonreferential, the difference between singular and plural does not apply to them. Therefore, we would predict that negative clauses would be less apt to regularize than affirmative clauses would. This prediction is supported by previous studies, such as D'Aquino Ruiz (2004), who found that affirmative polarity favors regularization. Thus, we code each example for affirmative or negative polarity.

Word order. Spanish is considered a flexible subject-verb-object language (López Meirama, 1997). Yet in conversational Puerto Rican Spanish, the vast majority of subjects (96%) occur in preverbal position (Subject-Verb-Object) (Brown & Rivas, 2011). Therefore, in order to test whether speakers interpret the preverbal NP as subject, due to the overwhelming frequency of subjects in this position in this particular variety, we code the position of the NP in relation to the conjugated form of *haber*. That is, if speakers regard the order NP + verb (in this case *haber*) as indicative of the syntactic pattern subject + verb, we would predict increased regularization in preverbal (as opposed to postverbal or null) NPs. This would be evidence of syntactic priming (e.g., Bock, 1986; Szmrecsányi, 2005). We distinguish between preverbal uses, postverbal uses, and nonapplicable in cases in which the NP is not overt (null).

Presence or absence of a quantifier. Bentivoglio and Sedano (1989) and D'Aquino Ruiz (2004), using data from Venezuela, argued that the presence of a numeral in the *presentatum* contributes to the pluralization of *haber* as do other

surface markers of plurality such as those found in indefinite quantifiers (*muchos* 'many', *algunos* 'some'). In contrast, in her study of Mexican Spanish, Castillo-Trelles (2007) reported that the presence of a quantifier disfavors pluralization of *haber,* whereas the absence of a quantifier favors agreement. To test whether the presence of a quantifier has any effect on pluralization in our Puerto Rican data, we code each instance of *haber* for the presence of a numeral, the presence of an indefinite quantifier, or the absence of any such markers.

Definiteness of the NP. As has been shown in typological studies (Comrie, 1989; Dixon, 2010; Rivas, 2004), prototypical subjects have human and definite referents. Because agreement is one of the defining characteristics of subjects in Spanish, and the innovative forms of *haber* entail agreement with their NP argument, we hypothesize that definite referents will favor the phenomenon of pluralization. Following Du Bois (1980), we consider definite NPs to be those with a definite article, demonstrative or possessive marker, and indefinite to be all other cases. Definiteness is in this way regarded as a grammatical category and not a pragmatic category.

Human or nonhuman referent of the presentatum. DeMello (1991) showed that in Puerto Rican Spanish, NPs with a human referent favor the use of the plural form, at least in the imperfect. Studies on Venezuelan Spanish (Bentivoglio & Sedano, 1989; Díaz-Campos, 1999–2000; Domínguez et al., 1998) also pointed out that a human NP favors pluralization of the verb. Thus, we code each token for human or nonhuman, according to the referent of the NP.

Corpus. All tokens were coded as to whether they were extracted from the CT corpus (Cortés-Torres, 2005) or the CE corpus (Davies, 2002–).

We submit our coded data to variable rule analysis using Varbrul (Rand & Sankoff, 2001). The following section presents the results of our quantitative and variable rule analyses.

RESULTS

Table 1 provides the total number of examples ($N = 190$) of presentational *haber* in which variation is possible (i.e., *haber* used with a plural NP argument) in the two corpora we are considering.[8] As can be seen in Table 1, in those contexts in which *haber* introduces a plural NP,[9] 44% of the examples present plural forms of *haber*.

To determine which of the coded factors make a significant contribution to the regularization of *haber*, we submit our data to a variable rule analysis using Varbrul (Rand & Sankoff, 2001). This enables us to determine the independent contribution of each factor group while controlling for all the other independent variables (Guy, 1993). Through this analysis, we are able to determine the statistical significance of each factor group—determined by both a p value and by the log likelihood (Sankoff, 1988). Further, Varbrul enables us to determine the relative strength of each factor group. The greater the range of the factor group, the greater the

TABLE 1. *Forms of* haber *with plural NP in Puerto Rican data*

Form of *haber*	*n*	%
Singular (e.g., *había*)	107	56
Plural (e.g., *habían*)	83	44
Total	190	100

magnitude of effect. Therefore, the factor group with the greatest range is the group that contributes most significantly to constraining the plural form of *haber*. Lastly, we can determine a constraint hierarchy through the Varbrul analyses. Within each factor group, the individual factors are ranked according to their factor weight. These weights reflect the degree to which they favor (>.50) or disfavor (<.50) the application of the dependent variable.

We include the following linguistic variables in the variable rule analyses: proportion of noun in subject function (low, mid, high), proportion of noun use with *haber* (low, mid, high), verbal form (simple versus compound), log of the word frequency per million of the noun (low, mid, high), polarity (positive versus negative), word order (preverbal, postverbal, null), quantifier (numeric, other quantifier, none), definiteness (definite NP versus indefinite NP), human referent (yes or no), and corpus (CT, CE).

We summarize the findings of the variable rule analysis in Table 2. Of the 10 factor groups considered for analysis, Varbrul selected as significant just two: the proportion of noun use as subject and the corpus (CT, CE). No other linguistic factor group was selected as significant for *haber* regularization in Puerto Rican Spanish. The only significant linguistic factor is the previously unidentified and unanalyzed factor group pertaining to the syntactic probability of the noun. The likelihood of *haber* regularization directly reflects the degree to which the noun being presented is used with subject function in spoken Spanish. That is, nouns with a low rate of usage with subject function trigger *haber* regularization less than do nouns with a high rate of usage with subject function.

If the noun being presented has a low proportion of use as a subject, regularization of *haber* is disfavored with a factor weight of .28 (e.g., *casas* 'houses', *Entonces, también, no había en Río Piedras las casas de hospedaje tan buenas que hay hoy día* 'So then, also, there weren't the great boarding houses like the ones we have nowadays in Río Piedras'[Davies, 2002–, *Habla Culta: San Juan*]). The verb *haber* is expressed in the plural with these low proportion nouns with a rate of 29%. The factor weight for plural nouns used with subject function with middle range frequency hovers close to .50 (.52), and regularization for these tokens occurs in 44% of the instances of use. Additionally, if the noun has a high frequency of use in spoken Spanish with a subject function, regularization of *haber* is strongly favored with a factor weight of .69 (e.g., *estudiantes* 'students', *Como habían estudiantes de bachillerato. . .*, 'Since there were high school students . . .' [Davies, 2002–, *Habla Culta: San Juan*]). With this class of nouns, *haber* regularizes in 56% of the instances.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

TABLE 2. *Factors favoring* haber *regularization in Puerto Rican oral Spanish Input: .42, N = 190 χ^2 per cell = .9678*

	Factor Weight	% Plural *haber* (*habían*)	*n*	% Data
Proportion of noun use as subject				
High	.69	56	64	33
Mid	.52	44	65	34
Low	.28	29	61	32
Range	41			
Corpus				
Cortés Torres	.70	58	97	51
Davies (2002–)	.29	27	93	48
Range	41			
Proportion of noun use in *haber* construction				
High	[.57][a]	46	60	31
Mid	[.53]	51	64	33
Low	[.41]	33	66	34
Verbal form				
Simple	[.52]	46	161	84
Compound	[.38]	27	29	15
Word frequency[b]				
High	[.55]	37	59	31
Mid	[.52]	50	66	34
Low	[.44]	43	65	34
Polarity				
Positive	[.51]	46	156	82
Negative	[.44]	29	34	17
Word order				
Preverbal	[.40]	44	18	9
Postverbal	[.49]	41	153	80
Null	[.69]	57	19	10
Quantifier				
Numeral	[.55]	58	34	17
Indefinite	[.42]	36	38	20
Absent	[.51]	41	118	62
Definiteness of NP				
Definite	[.64]	42	21	11
Indefinite	[.48]	43	169	88
Human				
Yes	[.57]	50	54	28
No	[.48]	41	136	71

Notes: [a] Factor weights between brackets are not significant. [b] We noted the crossover between factor weight and percentage pluralization for high-token frequency nouns. Following Paolillo (2002:89–91), we identified a suspected interaction between word frequency and corpus. The bulk of the high-token frequency words (76%, $n = 45$ of 59) come from the Davies corpus, which has a significantly lower percentage of pluralization (27% versus 58%). The low- and mid-frequency words only have 34% ($n = 22$ of 65) and 39% ($n = 26$ of 66) of their tokens from the Davies corpus. We did runs on the corpora separately (see note 11) and found no evidence of crossovers or interactions.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

This innovative probabilistic measure (subject grammatical relation probability) is found to be significant while controlling for multiple factors implicated in this regularization process. In this process of reanalyzing the object NP as a subject NP in the *haber* construction, individual nouns and their probabilistic features significantly constrain the variation. We argue that in addition to semantic and phonological information being stored lexically (exemplar model; Bybee, 2001), individual nouns, through repeated use with specific grammatical functions, acquire a grammatical relation probability. It is the probabilistic features of individual nouns that trigger *haber* regularization and not the abstract grammatical relation of subject itself. In other words, these results do not argue that *haber* projects two separate constructions: [*haber* + object NP] (not pluralized) and [*haber* + subject NP] (pluralized), and that a token with high proportion of use in subject function is more likely to occur in the latter construction. If it were the case that nouns more likely to be used as subjects in the grammar were also more likely to be used as subjects in the pluralized construction, then we would anticipate that the majority of nouns with a high proportion of use in subject function would occur with a pluralized *haber* form. In fact, the rate of use of both mid and high proportion nouns is roughly equal in both pluralized and nonpluralized forms.[10]

This study contributes a new probabilistic measure to the study of variation. However, the measurement we introduce is also innovative in how it views the predictive power of probabilities. Typically, the probabilistic measurement and the predicted linguistic outcome (e.g., durational shortening, syntactic reduction) are both contextually determined. That is, the probabilistic measurement is calculated based on the probability of use in the very specific context where the linguistic outcome is being predicted. Conversely, in the present analysis, the probability measurement (proportion of noun use as subject) is context-independent. That is, the linguistic outcome (pluralized *haber* forms) is not determined by the likelihood that a noun type of low, mid, or high proportion of use in subject function occurs in a pluralized (unambiguously subject) or a nonpluralized construction. Rather, the significant result we report in Table 2 reflects speakers' awareness of a cumulative probability based on their overall experience with the noun usage patterns generally in the language, in line with studies reporting significant effects of cumulative measures of frequency and probabilities on phonological reduction (Brown & Raymond, 2012; Bybee, 2002; Raymond & Brown, 2012).

As also is evident in Table 2, *haber* pluralization is favored in the CT corpus (factor weight .70) compared with the CE (Davies, 2002–) (factor weight .29). Indeed, in the CT corpus, most tokens of *haber* regularize (58%), whereas in the CE corpus, regularization is less common (27%).[11] This result likely reflects the type of data represented in each of the corpora: one being spontaneous recorded conversations among friends and family (CT), and the other representing more formal registers. Further, if *haber* regularization is, indeed, a change in progress as some suggest (D'Aquino Ruiz, 2008; Díaz-Campos, 2003; Hernández Díaz, 2006; Montes de Oca-Sicilia, 1994), it is noteworthy that

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

the CE data represent speech from the 1970s, whereas the CT data was recorded in 2000, thus providing indirect evidence in support of viewing this phenomenon as change. The potential role of extralinguistic factors in the variable realization of plural *haber* with presentative function lies outside the focus of this study.

DISCUSSION AND CONCLUSIONS

This study presents results of a previously untested notion regarding the role of syntactic probabilities in analogical processes. We have shown that the frequency of occurrence of a noun as subject in language usage conditions the regularization of *haber*. This suggests for each noun, therefore, that the speaker not only stores semantic information (such as gender, number, and count/mass), but also associates the noun with certain grammatical relations (propensity to be subject, for instance). This knowledge of usage patterns affects production and may guide analogical processes.

This syntactic probability measure was applied to the widely studied phenomenon of *haber* regularization in Spanish specifically. The quantitative analysis of each occurrence of variable presentational *haber* in oral Puerto Rican Spanish did not support previously held assumptions regarding the linguistic factors that constrain the variation of *haber* (verbal form, presence/absence of quantifiers, polarity, and human NP) determined from analyses of other varieties of Spanish. Rather, the innovative methodology employed in this analysis revealed that *haber* regularization is affected by speakers' probabilistic knowledge of grammatical relations. Nouns with a higher probability of use in discourse with a subject function promote *haber* regularization, whereas nouns with low probability of subject use disfavor *haber* pluralization.

This effect of proportion of noun use as subject may have been detected via correlation in previous research (Bentivoglio & Sedano, 1989; DeMello, 1991; Díaz-Campos, 1999–2000; Domínguez et al., 1998) that found that human referents significantly favor the regularized form over nonhuman referents. *Haber* is used in our corpus to present both human (e.g., *había pasajeros* 'there were passengers') and nonhuman NP referents (e.g., *había un restaurante* 'there was a restaurant'), albeit not with the same frequency. *Haber* is used more frequently in our corpus to present a nonhuman NP referent (96 types, 136 tokens) than a human NP referent (32 types, 54 tokens). However, the rate of *haber* regularization is higher (50%) for plural human NP referents than for plural nonhuman NP referents (41%). This results ties in with a tendency for human entities to function as subjects. As is noted by Coco and Keller (2009:275):

> Animacy is known to play a role in language production; in particular, it can influence the assignment of grammatical functions and word order. . . . Animate entities are conceptually more accessible than inanimate ones . . . and therefore privileged during syntactic encoding. This is reflected by the fact that animate entities are

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

TABLE 3. *Noun referents coded as nonhuman*

Proportion of Noun Use as Subject	*N*	% Pluralization
Low (0–7%, average 4%)	54	29
Mid (8–16%, average 18%)	55	40
High (20–66%, average 33%)	27	67

Note: Difference between low and high: $p = .0016$, $\chi^2 = 10.13767$; difference between mid and high: $p = .0232$, $\chi^2 = 5.154286$; difference between low and mid: not significant.

more likely to be encoded with the grammatical function subject, while inanimate entities occur mostly with the function object.

It is true that, for the NP referents in our data, the human NPs have a higher probability on average of functioning as subjects than do nonhuman NPs (26% and 13%, respectively). However, to attribute the human/nonhuman distinction a causal function in the regularization of *haber* fails to capture the bigger picture. That is to say, the human NPs act as subject more than nonhuman NPs do, but our results suggest this effect is gradient and that the binary +/– human distinction used in previous analyses captures less precisely the regularization process.[12]

Indeed, if the explanation for *haber* regularization were simply a matter of the categorization of the NP referent as human or nonhuman, we would not expect to see different rates of regularization within one of those categories. That is to say, all NPs with a human referent would display similar rates of pluralization, as would NPs with nonhuman referents. This is not the case in our data. For instance, if we examine the plural, nonhuman NP referents, there is a significant difference ($p = .002$) in rates of *haber* regularization for the group of NPs infrequently used as subject (e.g., *ventanas* 'windows') compared with the group of NPs most often used with subject function (e.g., *chismes* 'gossip'). This is summarized in Table 3.[13] Within the group of nonhuman NP plural referents, there is *haber* regularization in 29% of the cases with nouns used infrequently as subjects. This rate of regularization is significantly lower than the 67% regularization for the group of nouns used frequently in discourse with a subject function. Thus, the lexical effect of the noun's cumulative syntactic function probability is also evident for nonhuman NP referents.

Another usage-based factor group we examined in this study (the proportion of use of noun in the *haber* construction) was not selected as significant. In line with Goldberg (1995, 2006) and Croft (2001), in this study [*haber* + NP] is described as a construction. This construction has a direct form (V-*haber* + NP) ~ function (presentational) pairing in which the verb slot is fixed, as it is always occupied by a form of *haber*, and the NP slot is open. We predicted, based on Bybee and Thompson (1997) and Bybee (2003), that the innovative *haber* construction could be promoted with nouns that are used infrequently in the [*haber* + NP]

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

construction. As Bybee (2003) argued, analogical leveling affects low frequency items first. Thus, NP arguments frequently occurring in the [*haber* + NP] construction would be predicted to be more resistant to regularization due to the conserving effect of frequency (Bybee & Thompson, 1997).[14] If the loss of the unusual presentative construction were the result of analogical leveling with the regularized pattern of NP subject + verb agreement, we would expect that the innovative construction would arise in contexts in which the NP argument is infrequent in the *haber* construction. Our results show, however, that the regularization of *haber* we find in our Puerto Rican data does not reflect this tendency.

Much of the theory and method of this study builds largely on usage-based theory generally (e.g., Bybee, 2001, 2002, 2010) and the growing body of literature on the effects of linguistic probabilities. We argue that the lexical representation of words contains information regarding syntactic probabilities. Gahl and Garnsey (2004:751) identified three different criteria that data must meet to be considered part of the grammar:

> First, one would certainly ask that the relevant probabilities be based on phrase types, rather than specific lexical items; second, they should make reference to syntactic relationships, rather than simple adjacency or cooccurrence of syntactic phrases; and third, they should be distinguishable from probabilities based on real-world plausibility, that is, the likelihood that a given sentence may be true.

Our data meets these three criteria. First, the probabilistic measure we report, rather than phrase type, refers to grammatical relation and not to particular lexical items or identified lexical effects such as word frequency, which was not selected as significant in our variable rule analyses. Second, our measure is not contingent on co-occurrence frequencies. Recall, the frequency with which the noun occurs in the [*haber* + NP] construction is not significant in this regularization process in Puerto Rican Spanish. Lastly, we argue the syntactic probability effect we report is lexicalized and not based solely on real-world plausibility. It could be argued that the high syntactic probability for humans to be subjects is determined by the extralinguistic (real-world) tendency for humans to act as agents. And, as Comrie (1989) pointed out, prototypical subjects are agents, and therefore *haber* regularization could reflect online cognitive processes and not lexicalized probabilities.[15] Evidence against this argument lies in usage patterns detected in the corpus (Davies, 2002–) regarding classes of nouns (i.e., human versus nonhuman). For instance, two seemingly comparable nouns describing groups of professionals (*abogados* 'lawyers', *maestros* 'teachers') could feasibly, in plausible terms, act as subjects with equal likelihood. However, it is the actual grammatical probability that predicts the regularization pattern and not the mere possibility of acting as a subject. These nouns vary significantly ($p = .0000$, $\chi^2 = 164.6702$) in their grammatical relations; in this illustration, *abogados* is used in subject function in only 13% of cases ($N = 579$), whereas *maestros* is used with subject function in 45% of cases ($N = 896$), and these probabilities are tied to specific nouns.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press

This significant result of high proportion of noun use in subject function favoring *haber* pluralization we have interpreted as a lexical effect, that is, having to do with information about syntactic contexts of use stored with lexical items. Other analyses of syntactic probabilities employ an information-theoretic approach to language production, as illustrated, for example, in Jaeger (2010) and Tily et al. (2009). These authors propose the principle of Uniform Information Density (UID), where information is "measured in such a way that the more probable an item in a sequence, the less informative it is, and conversely the less probable, the more informative" (Bresnan & Ford, 2009:57). The UID principle assumes if "the rate at which information is conveyed in the speech stream is roughly constant, then more predictable words, which carry little information, should take less time to pronounce during production than less predictable words" (Bresnan & Ford, 2009:57). As such, the UID principle is straightforwardly applied to the study of reductive processes. Although we analyze a case of analogy that does not involve a reductive process (but perhaps the opposite), our results might lend partial support to the UID proposal (Jaeger, 2010). If we consider that a less predictable noun for the *haber* construction can be interpreted as one typically used as subject, and not object, then our findings could be accounted for within this view. Tokens with low proportion of noun use as subject do not favor regularization with the accompanying additional phonological material entailed by the morphological expression of plural.

Other explanations for our results might appeal to lexical access during speech production. For instance, using a connectionist model of sentence production, Chang et al. (2006:246) argued that thematic role distinctions contribute to the selection of structures in language production. On this basis, these authors consider the possibility that lexical items may be closely linked with syntactic knowledge. We support this view in our analyses, and our data demonstrate a significant effect of grammatical relation distinctions among nouns. As Chang et al. (2006:243) pointed out, speech production involves "incremental competition between words that are activated by the message. The sequencing system attempts to make a grammatical sequence out of the winners of this competition." Such competition between nouns with different proportion of uses as subject and their relative activation may thus bias production and the morphological form of *haber*.

In this analysis, we provide statistical control for both previously identified linguistic factors attributed to the *haber* regularization in Spanish, as well as novel factors, and find that the proportion of nouns in subject function significantly constrains variable agreement of *haber*. We would not advocate a view that any one source alone could account for the pluralization pattern we observe for *haber* and acknowledge a need for future tests of this factor that we identify. For instance, we lack control for lexical and/or syntactic priming in the current analysis. Given the multifaceted ways in which priming is shown to have effects on all levels of language production (e.g., Chang et al., 2006; Gries, 2005; Szmrecsányi, 2005; Torres Cacoullos & Travis 2010), future analysis should test for potential effects of syntactic and lexical priming.

The results reported here, therefore, describe a new type of linguistic information stored in the lexicon. The spread of this analogical process (*haber* regularization in Spanish) is not determined by phonological and/or semantic similarity as has been shown for other changes. Rather, syntactic probabilities of nouns in subject function are shown to significantly constrain this variation. This type of probabilistic information attributed to lexical items should be tested on other linguistic variables to determine how pervasive it is in processes of language variation and change.

NOTES

1. Harris and Campbell (1995) also included borrowing among the mechanisms of syntactic change.

2. Bybee (2010:66) noted that analogical leveling is also modeled by frequency, because (1) it applies to low-frequency verbs first (*weep-weeped/wept*, *leap-leaped/leapt*), high-frequency verbs being more resistant to leveling (*keep-kept/?keeped*, *sleep-slept/?sleeped*), and (2) the new forms are created on the basis of the most frequent pattern in the paradigm (*-ed* preterit forms in this particular example).

3. Erman and Warren (2000:31) described prefabs as "a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization." Examples of prefabs are *I'm afraid* (used as a softener of bad news) (cf. with **I'm scared* or **I'm frightened*, which are not acceptable as softeners of bad news), *I can't see a thing* (**I can't see an object*), or *intensive care* (**intensive attention*).

4. As is noted by Brinton and Traugott (2005), Hopper and Traugott (2003), and Lehmann (2002), *inter alia*, semantic bleaching is one of the typical characteristics of grammaticalization.

5. Hernández Díaz (2006) pointed out that the earliest examples of presentational-intransitive *haber* are from the 16th century. However, she also indicates that there are very few attested examples before the 20th century.

6. From a minimalist perspective, Rodríguez Mondoñedo (2006) maintained that the regularized construction and the impersonal construction are actually independent of each another; that is to say, this author points out that in diachronic terms the regularized construction does not derive from the impersonal construction or vice versa (Rodríguez Mondeño, 2006:382n34). This perspective is challenged by Fontanella de Weinberg (1992) and Hernández Díaz (2006). Furthermore, Rodríguez Mondoñedo (2006) maintained that the *presentatum* is not the subject of the construction, even when it agrees in number with the verb. Following Chomsky (1965), this author considers subject to be defined structurally as the constituent that occupies the specifier of tense (T) phrase (P). As a result, Rodríguez Mondoñedo (2006) accounted for the syntactic structure of existential *haber* in Spanish by postulating the existence of a *v*P in which *v* has only the feature of number. In constructions in which *haber* agrees with the NP, T also has a person feature that accounts for agreement. Our notion of subject, however, follows functionalist analyses such as Keenan (1976) and Givón (2000), in which subject is defined as a gradual category that has functional as well as formal (coding and behavior-and-control) properties. From this perspective, *haber* agreement makes the NP more subject-like than lack of agreement does, because agreement is one of the coding properties of subject in Spanish. In addition to this, even though Rodríguez Mondoñedo (2006) argued that the accusative clitic may occur even if the verb agrees with the NP (e.g., *Hubieron dos hombres en la fiesta* 'there were two men at the party': *los hubieron* 'ACCUSATIVE CLITIC were' [Rodríguez Mondoñedo, 2006:327]), we have not found any example of this construction in either of the corpora that we have used for this study.

7. We use two separate corpora owing to the low textual frequency of this construction.

8. This excludes third-person singular present form *hay* because there is no plural counterpart in our data. Although Montes Giraldo (1982) reported the existence of plural forms such as *haen* and *hayn* in some dialects of Colombia, in most varieties of Spanish in which variable agreement of *haber* occurs, including Puerto Rican Spanish, the present tense does not present variation. One of the reasons for this may be that, in morphological terms, *hay* is in itself an irregular form, because it derives from the amalgamation of the third-person singular form *ha* plus the locative deictic *y* (García, 1986). For this reason, there is no plural analog in the system, unlike what happens with other tenses such as imperfect (*había-habían*) or preterit (*hubo-hubieron*).

9. In the CT corpus, we find variation with a few singular NPs. We find two instances of pluralization of *haber* with the singular NP referent *gente* 'people' (i.e., *Pero él, él sufría porque habían gente que tenían dinero* 'But he, he suffered because there were people who had money' [Interview 7, p. 57]). These cases may be regarded as examples of *ad sensum* agreement, because they are concerned with a collective noun,

which triggers singular agreement in Spanish but has a plural referent. Interestingly, *gente* has a proportion of use as subject that is extremely high (53%). Using just the CT corpus, we conduct a separate analysis including all instances of *haber* with either a singular ($n = 248$) or a plural ($n = 97$) referent. The Varbrul analysis we conduct on these data includes all the factor groups considered in the analysis of plural NPs, with an additional group coding the number of the NP (singular or plural). Varbrul selected the following groups as significant: number of the NP referent (plural [factor weight .98], singular [factor weight .17]), and proportion of noun as subject (low [factor weight .28], mid [factor weight .45], high [factor weight .75]) ($p = .003$, log likelihood = –56.136, chi-square per cell = 1.1702).

10. Tokens of high proportion of noun use in subject function are not significantly more likely ($p = .1573$) to be used in the pluralized construction ($n = 36$) than in the nonpluralized construction ($n = 28$). The same lack of significant difference ($p = .2195$) holds for the tokens with mid proportion use as subject ($n = 29$ for pluralized constructions and $n = 36$ for nonpluralized constructions). In contrast, there is a significant difference in use in the low category ($p = .0000$, $n = 61$, $\chi^2 = 20.4918$), where nouns with low-proportion use with subject function are used significantly more often (70%) in the nonpluralized construction ($n = 43$) than in the pluralized construction ($n = 18$). This result partially supports a view where simply a noun used frequently as subject anywhere in the grammar would also be likely to be used elsewhere in the grammar where subjects are permissible.

11. Two Varbrul analyses conducted on the plural NP data from each corpus separately [CE: $N = 93$, CT: $N = 97$)] selected the same factor group as significant (proportion noun as subject) with the same ordering of factors and direction of effect. The factor group "proportion of noun use as subject" significantly ($p = .007$) constrains variation in the CE (low: .13, mid: .57, high: .68) (input = .23, log likelihood = –41.375, $\chi^2 = 57.4650$), as well as having a significant ($p = .009$) effect in the CT corpus (low: .34, mid: .51, high: .78) (input = .60, log likelihood = –57.554, $\chi^2 = 82.9245$). In both analyses, this was the only factor group selected as significant. This factor group (proportion of noun use as subject) was also found to be significant in a paper analyzing other varieties of Latin American Spanish (Grammon, 2012).

12. We did separate runs on the data. One including the human/nonhuman characteristic of the NP as a factor group while excluding the proportion of noun use as subject, and another with noun use as subject to the exclusion of +/– human. Separate runs indicated that the model incorporating the proportion of noun use as subject provides a significantly better fit for the data ($p > .005$, $df = 1$, log likelihood +/– human = –112.304, log likelihood proportion of noun use as subject = –106.196).

13. A similar direction of effect is noted for the plural human NPs (low-proportion subject [$N = 7$]: 40% pluralization of *haber*, mid-proportion subject [$N = 10$]: 70% pluralization of *haber*, high-proportion subject [$N = 37$]: 51% pluralization of *haber*), although based on a lower number of tokens.

14. In line with the conserving effect of frequency, Waltereit and Detges (2008) argued that the reanalysis of the NP argument of *haber* as subject starts in low-frequency tenses owing to speakers' linguistic insecurity regarding the syntactic structure of the construction [*haber* + NP]. Waltereit and Detges maintained that frequency explains why pluralization is especially disfavored in the present (*hayn/haen/hain*), which is the tense that displays the highest frequency of use for presentative constructions. Similarly, in nonpresent tenses, *haber* regularization is more frequent in the imperfect (*habían*) than in the preterit (*hubieron*), the imperfect being the less frequent of the two Spanish past tenses. Even if we assume that the preterit tense is more frequent than the imperfect tense—in the oral section of Davies (2002–), the token frequency of the imperfect ($N = 57{,}134$) is actually higher than the token frequency of the preterit ($N = 54{,}079$)—the imperfect displays a higher token frequency than the preterit does if we only take into account the occurrence of both tenses in the presentational *haber* construction. For example, in the CT corpus the imperfect (*había/habían*) corresponds with 15% ($N = 229$) of the total examples of presentative *haber*, whereas the preterit only occurs in 4% ($N = 65$) of cases.

15. However, in Spanish, the subject is used to encode not only agents, but also other semantic roles whose referent is not necessarily human, such as themes (Moure, 1995). In this same line, Langacker (1990) also showed that the subject is not necessarily a human being. He understood relational predications as involving a figure and ground relationship. The figure, which Langacker (1990) identified with the subject, is a perceptual prominent element (a thing-like element) because it has a shape, coherence, and structure. In addition to this, in conversational data, agent-patient constructions, that is to say, cardinal transitive constructions (Hopper & Thompson, 1980), are highly infrequent (see, e.g., Thompson & Hopper [2001] for English, Vázquez Rozas & García Miguel [2006] for Spanish).

REFERENCES

Alarcos Llorach, Emilio. (1994). *Gramática de la lengua española*. Madrid: Espasa Calpe.

Ashby, William J., & Bentivoglio, Paola. (1997). Strategies for introducing new referents into discourse: A comparative analysis of French and Spanish presentational structures. In R. M. Hammond &

M. MacDonald (eds.), *Linguistic studies in honor of Bodhan Saciuk*. West Lafayette, IN: Learning Systems Inc. 9–26.
Beckner, Clay, Blythe, Richard, Bybee, Joan, Christiansen, Morten H., Croft, William, Ellis, Nick C., Holland, John, Ke, Jinyun, Larsen-Freeman, Diane, & Schoenemann, Tom. (2009). Language is a complex adaptive system: position paper. *Language Learning* 59(Suppl. 1):1–26.
Bell, Alan, Brenier, Jason M., Gregory, Michelle, Girand, Cynthia, & Jurafsky, Dan. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1):92–111.
Bentivoglio, Paola, & Sedano, Mercedes. (1989). Haber: ¿un verbo impersonal? Un estudio sobre el español de Caracas. *Estudios sobre español de América y lingüística afroamericana. Ponencias presentadas en el 45 Congreso Internacional de Americanistas (Bogotá, julio de 1985)*. Bogotá: Instituto Caro y Cuervo. 59–81.
Bock, J. Kathryn. (1986). Syntactic persistence in language production. *Cognitive Psychology* 18 (3):355–387.
Bresnan, Joan, & Ford, Marilyn. (2009). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1):168–213.
Brinton, Laurel J., & Traugott, Elizabeth C. (2005). *Lexicalization and language change*. Cambridge: Cambridge University Press.
Brown, Esther L., & Raymond, William. (2012). How discourse context shapes the lexicon: Explaining the distribution of Spanish *f-/h-* words. *Diachronica* 29(2):139–161.
Brown, Esther L., & Rivas, Javier. (2011). Subject ~ verb word-order in Spanish interrogatives: A quantitative analysis of Puerto Rican Spanish. *Spanish in Context* 8(1):23–49.
Bush, Nathan. (2001). Frequency effects and word-boundary palatalization in English. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 255–280.
Bybee, Joan. (1985). *Morphology. A study of the relation between meaning and form*. Amsterdam: John Benjamins.
_____. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
_____. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14:261–290.
_____. (2003). Mechanisms of change in grammaticization: The role of frequency. In B. Joseph & R. Janda (eds.), *The handbook of historical linguistics*. Oxford: Blackwell. 602–623.
_____. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.
Bybee, Joan, & Thompson, Sandra A. (1997). Three frequency effects in syntax. *Berkeley Linguistics Society* 23:65–85.
Bybee, Joan, & Torres Cacoullos, Rena. (2009). The role of prefabs in grammaticization: How the particular and the general interact in language change. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (eds.), *Formulaic language*. Vol. 1. *Distribution and historical change*. Amsterdam: John Benjamins. 187–217.
Campbell, Lyle. (1998). *Historical linguistics. An introduction*. Cambridge: The MIT Press.
Castillo-Trelles, Carolina. (2007). La pluralización del verbo *haber* impersonal en el español yucateco. In J. Holmquist, A. Lorenzino, & L. Sayahi (eds.), *Selected proceedings of the third workshop on Spanish sociolinguistics*. Somerville, MA: Cascadilla Proceedings Project. 74–84.
Chafe, Wallace. (1987). Cognitive constraints on information flow. In R. S Tomlin (ed.), *Coherence and grounding in discourse*. Amsterdam: John Benjamins. 21–51.
Chang, Franklin, Dell, Gary S., & J. Kathryn Bock. (2006). Becoming syntactic. *Psychological Review* 113(2):234–272.
Chomsky, Noam. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.
Coco, Moreno, & Keller, Frank. (2009). The impact of visual information on reference assignment in sentence production. In N. Taatgen & H. van Rijn (eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society*. Amsterdam: CogSci 2009. 274–279.
Comrie, Bernard. (1989). *Language universals and linguistic typology: Syntax and morphology*. 2nd ed. Oxford: Blackwell.
Cortés-Torres, Mayra E. (2005). *La perífrasis estar + -ndo en el español puertorriqueño: variación dialectal o contacto lingüístico?* Ph.D. dissertation, University of New Mexico.
Croft, William. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
Davies, Mark. (2002–). *Corpus del Español (100 million words, 1200s–1900s)*. Available at: http://www.corpusdelespanol.org.

D'Aquino Ruiz, Giovanna. (2004). Haber impersonal en el habla de Caracas. Análisis sociolingüístico. *Boletín de Lingüística* 21:3–26.
_____. (2008). El cambio lingüístico de *haber* impersonal. *Núcleo* 25:103–123.
DeMello, George. (1991). Pluralización del verbo "haber" impersonal en el español hablado culto de once ciudades. *Thesaurus* 46(3):445–471.
Díaz-Campos, Manuel. (1999–2000). La pluralización del verbo *haber* en dos áreas dialectales de Hispanoamérica. *Anuario de Lingüística Hispánica* 15–16:235–245.
_____. (2003). The pluralization of *haber* in Venezuelan Spanish: A sociolinguistic change in real time. *IU Working Papers in Linguistics* 03–05. https://www.indiana.edu/~iulcwp/pdfs/03-Diaz-Campos05.pdf Access date: 09/14/2012.
Dixon, Robert M. W. (2010). *Basic linguistic theory.* Vol. 2. *Grammatical topics*. Oxford: Oxford University Press.
Domínguez, Carmen, Guzmán, Blanca, Moros, Luis, Pabón, Maryelis, & Vilaín, Roger. (1998). Personalización de *haber* en el español de Mérida. *Lengua y Habla* 3(1):23–36.
Du Bois, John W. (1980). Beyond definiteness: The trace of identity in discourse. In W. Chafe (ed.), *The Pear Stories: Cognitive, cultural and linguistic aspects of narrative production*. Norwood, NJ: Ablex Publishing Corporation. 203–274.
Erman, Britt, & Warren, Beatrice. (2000). The idiom principle and the open choice principle. *Text* 20:29–62.
Fontanella de Weinberg, María B. (1992). Variación sincrónica y diacrónica de las construcciones con haber en el español americano. *Boletín de Filología de la Universidad de Chile* 33:35–46.
Freites Barros, Francisco. (2004). Pluralización de haber impersonal en el Táchira: Actitudes lingüísticas. *Boletín de Lingüística* 22:32–51.
Frisch, Stephan A., Large, Nathan R., & Pisoni, David B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42 (4):481–496.
Gahl, Susanne, & Garnsey, Susan M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80:748–875.
García, Erica. (1986). Cambios cuantitativos en la distribución de formas: ¿Causa y síntoma de cambio semántico? In A. D. Kossoff, R. H. Kossoff, G. Ribbans, & J. Amor y Vázquez (coords.), *Actas del VIII congreso de la Asociación Internacional de Hispanistas: 22–27 agosto 1983*. Madrid: Istmo. 557–566.
Givón, Talmy. (2000). *Syntax: An introduction*. Amsterdam: John Benjamins.
Goldberg, Adele. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
_____. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
Grammon, Devin. (2012). Nuevas aproximaciones al fenómeno de la pluralización de *haber* impersonal: Un análisis basado en el uso. Paper presented at the 2nd CU Graduate Conference Framing Narratives, Boulder, University of Colorado, April 13–14, 2012.
Gries, Stefan Th. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34:365–399.
Guy, Gregory. (1993). The quantitative analysis of linguistic variation. In D. R. Preston (ed.), *American dialect research.* Amsterdam: John Benjamins. 223–249.
Harris, Alice, & Campbell, Lyle. (1995). *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.
Hernández Díaz, Axel. (2006). Posesión y existencia: La competencia de *haber*, *tener* en la posesión y *haber* existencial. In C. Company Company (ed.), *Sintaxis histórica de la lengua española. Primera parte: la frase verbal.* Vol. 2. México: Universidad Nacional Autónoma de México: Fondo de Cultura Económica. 1053–1160.
Hopper, Paul, & Thompson, Sandra A. (1980). Transitivity in grammar and discourse. *Language* 56 (2):251–299.
Hopper, Paul & Traugott, Elizabeth C. (2003). *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.
Jaeger, T. Florian. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1):23–62.
Jurafsky, Daniel, Bell, Alan, Gregory, Michelle, & William D. Raymond. (2001). Probablilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 229–254.
Keenan, Edward. (1976). Towards a universal definition of 'subject'. In C. Li (ed.), *Subject and topic*. New York: Academic Press. 303–333.

Langacker, Ronald. (1987). *Foundations of cognitive grammar: Theoretical prerequisites.* Vol. 1. Stanford: Stanford University Press.
_____. (1990). Settings, participants, and grammatical relations. In S. Tsohatzidis (ed.), *Meanings and prototypes: Studies on linguistic categorization*. London: Routledge & Kegan Paul. 213–238.
Lehmann, Christian. (2002). *Thoughts on grammaticalization*. Erfurt, Germany: Universität Erfurt.
López Meirama, Belén. (1997). Aportaciones de la tipología lingüística a una gramática particular: el concepto de orden básico y su aplicación al castellano. *Verba* 24: 45–82.
Montes de Oca-Sicilia, M. del Pilar. (1994). La concordancia con *haber* impersonal. *Anuario de Letras* 32:7–35.
Montes Giraldo, J. Joaquín. (1982). Sobre el sintagma "Haber + sustantivo." *Thesaurus* 37(2):383–385.
Moure, Teresa. (1995). Sobre el controvertido perfil del complemento directo. *Moenia* 1:47–110.
Paolillo, John. (2002). *Analyzing linguistic variation: Statistical models and methods*. Stanford: CSLA Publications.
Pierrehumbert, Janet B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 137–157.
Poplack, Shana, & Meechan, Marjory. (1998). Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2:127–138.
Poplack, Shana, & Tagliamonte, Sali. (2001). *African American English in the diaspora*. Maiden, MA: Blackwell Publishers.
Quintanilla-Aguilar, José R. A. (2009). *La (des)pluralización del verbo haber existencial en el español salvadoreño: ¿Un cambio en progreso?* Ph.D., University of Florida.
Rand, David, & Sankoff, David. (2001). *GoldVarb: A variable rule application for Macintosh*. Toronto: University of Toronto, Department of Statistics.
Raymond, William D., Dautricourt, Robin, & Hume, Elizabeth. (2006). Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18:55–97.
Raymond, William D., & Brown, Esther L. (2012). Are effects of word frequency effects in of context of use? An analysis of initial fricative reduction in Spanish. In S. Th. Gries & D. S. Divjak (eds.), *Frequency effects in language. Vol 2: Learning and processing*. The Hague: Mouton de Gruyter. 35–52.
Rivas, Javier. (2004). *Clause structure typology: Grammatical relations in cross-linguistic perspective*. Lugo: Tris Tram.
Rodríguez Mondoñedo, Miguel. (2006). Spanish existentials and other accusative constructions. In C. Boeckx (ed.), *Minimalist essays*. Amsterdam: John Benjamins. 326–394.
Sankoff, David. (1988). Variable rules. In U. Ammon, N. Dittmar, & K. J. Mattheier (eds.), *Sociolinguistics: An international handbook of the science of language and society*. New York: Walter de Gruyter. 984–997.
Szmrecsányi, Benedikt. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1):113–149.
Thompson, Sandra A., & Hopper, Paul. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In J. Bybee & P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 27–60.
Tily, Harry, Gahl, Susanne, Arnon, Inbal, Snider, Neal, Kothari, Anubha, & Bresnan, Joan. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2):147–165.
Torres Cacoullos, Rena, & Travis, Catherine E. (2010). Variable *yo* expression in New Mexico: English influence? In S. Rivera-Mills & D. J. Villa (eds.), *Spanish in the U.S. Southwest: A language in transition*. Madrid: Iberoamericana. 185–206.
Torres Cacoullos, Rena, & Walker, James A. (2009). The present of the English future: Grammatical variation and collocations in discourse. *Language* 85(2):321–354.
Vaquero, María T. (1978). Enseñar español, pero ¿qué español? *Boletín de la Academia Puertorriqueña de la Lengua Española* 6:127–146.
Vázquez Rozas, Victoria, & García-Miguel, José M. (2006). Transitividad, subjetividad y frecuencia de uso en español. *VII congrés de lingüística general. Actes, del 18 al 21 d'abril de 2006*. Barcelona: Universitat de Barcelona. CD-Rom.
Vergara Wilson, Damián. (2009). From "remaining" to "becoming" in Spanish: The role of prefabs in the development of the construction *quedar(se)* + ADJECTIVE. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (eds.), *Formulaic language*. Vol. 1. *Distribution and historical change*. Amsterdam: John Benjamins. 273–295.

Waltereit, Richard, & Detges, Ulrich. (2008). Syntactic change from within and from without syntax: A usage-based analysis. In R. Waltereit & U. Detges (eds.), *The paradox of grammatical change: Perspectives from Romance*. Amsterdam: John Benjamins. 13–30.

https://doi.org/10.1017/S0954394512000154 Published online by Cambridge University Press