

RESEARCH

Probabilistic Grammar: The view from Cognitive Sociolinguistics

Jeroen Claes

KU Leuven, Wambachstraat 8, 2018 Antwerp, BE
jeroen.claes@kuleuven.be

In this paper, I propose that Probabilistic Grammar may benefit from incorporating theoretical insights from Cognitive (Socio)Linguistics. I begin by introducing Cognitive Linguistics. Then, I propose a model of the domain-general cognitive constraints (markedness of coding, statistical preemption, and structural priming) that condition language (variation). Subsequently, three case studies are presented that test the predictions of this model on three distinct alternations in English and Spanish (variable agreement with existential *haber*, variable agreement with existential *there be*, and Spanish subject pronoun expression). For each case study, the model generates empirically correct predictions. I conclude that, with the support of Cognitive Sociolinguistics, Probabilistic Grammar may move beyond description towards explanation.

Keywords: Cognitive Sociolinguistics; Probabilistic Grammar; existential agreement variation; *haber* pluralization; Spanish subject pronoun expression

1 Introduction

In recent years, Probabilistic Grammar has emerged as a booming research tradition at the crossroads of corpus linguistics, variationist linguistics, and theoretical linguistics. The main claim put forward in this area of research is that language is not determined by categorical hard constraints, but rather by the joint action of a multitude of probabilistic soft constraints, which are learned from input and maintained and refined by experience (e.g., Bresnan et al. 2007). To examine these constraints, Probabilistic Grammar performs large-scale corpus studies using the variationist methodology (e.g., Labov 1982; Tagliamonte 2012), examining patterns of correlation between a particular alternation and features such as animacy, constituent length, definiteness, etc. (see e.g., Szmrecsanyi et al. 2016 for a recent example). The sizes, the dimensionality (does the predictor favor or disfavor a particular variant?) and the relative importance of these effects are then interpreted as constituting, on aggregate, speakers' Probabilistic Grammar of the specific alternation (Szmrecsanyi 2013).

The Cognitive Linguistics family of linguistic theories (e.g., Lakoff 1987; Langacker 1987, 1991; Goldberg 1995; Croft & Cruse 2004) shares many of the core assumptions of Probabilistic Grammar, for which there is a great deal of overlap between the two approaches to language. However, from the perspective of Cognitive Linguistics, one shortcoming of Probabilistic Grammar stands out. That is, the contextual features that are used to predict speakers' behavior – and, by extension, the probabilities that are obtained for them – have no independent, prior theoretical motivation relating these features to domain-general cognitive constraints or capacities (see section 2.1 on the generalization and cognitive commitments; Lakoff 1990). Rather, speakers' probabilistic knowledge is claimed to stem from usage, but, in reality, the features that are used to predict speakers' behavior are generally derived from the analysts' intuitions about the particular phenomenon at hand or from a large tradition of corpus-based alternation studies (see

Henry 2002: 277 for a similar observation applied to variationist sociolinguistics and see Geeraerts 2005 for a critique of introspective judgments from the position of Cognitive Linguistics). Because of this, Probabilistic Grammar is unable to formulate predictions about the linguistic features that will condition previously unstudied morphosyntactic alternations and, perhaps more importantly, with what directionality and why. Since these sorts of predictions are the hallmark of any scientific theory, Probabilistic Grammar is more akin to a methodological than to a theoretical framework.

In this paper, I will argue that Cognitive Sociolinguistics may contribute to Probabilistic Grammar the theoretical skeleton it needs to develop into a psychologically plausible perspective on the constraints that shape speakers' probabilistic grammars, resulting in a theoretical framework that is both descriptive and explanatory. However, this rapprochement between the two traditions will require reversing the research questions: rather than inferring hypotheses about the cognitive underpinnings of variation from correlation patterns mined from data, as is customary in Probabilistic Grammar, in this paper, I will test hypotheses about correlation patterns that derive from assumptions about the cognitive underpinnings of language (variation). The result is a theoretical model of the cognitive constraints that govern morphosyntactic variation, which generates predictions about the Probabilistic Grammar of alternations across languages.

The remainder of this paper is organized as proceeds. In section 2, I will briefly introduce Cognitive Linguistics and Cognitive Sociolinguistics, as well as the cognitive constraints on coding/categorization/language production that are assumed in this family of linguistic theories. To test this model, section 3 presents a case study of existential agreement in Caribbean Spanish. Section 4 applies the model to a new, but still related, case of morphosyntactic variation, namely, existential agreement variation in British English. Subsequently, section 5 confronts the theoretical model with Spanish subject personal pronoun expression, a completely unrelated alternation. Section 6 discusses the results and presents some conclusions.

2 Cognitive (Socio)linguistics and cognitive constraints on coding

2.1 Cognitive (Socio)linguistics

Cognitive Linguistics is a theoretical movement that includes frameworks such as Cognitive Construction Grammar (Goldberg 1995, 2006), Cognitive Grammar (Langacker 1987, 1991), usage-based theory (e.g., Bybee 2010), and Word Grammar (e.g., Hudson 2010). What unites these frameworks, which may differ substantially concerning their specific focus and proposals, is the following shared set of guiding assumptions, many of which will sound very familiar to Probabilistic Grammarians (see e.g., Croft & Cruse 2004: Chapter 1 or Geeraerts 2006 for more elaborate overviews of the basic tenets of Cognitive Linguistics).

1. *Linguistic knowledge derives from usage.* Cognitive Linguistics maintains that language use affects language structure, for which quantitative patterns in language such as frequency of co-occurrence are important explanatory constructs (e.g., Croft & Cruse 2004: Chapter 10). Also, because usage inevitably includes variation, Cognitive Linguistics proposes that “it is important not to view the regularities as primary and the gradience and variation as secondary; rather the same factors operate to produce both regular patterns and the deviations” (Bybee 2010: 6; see Geeraerts 2005 as well). Cognitive Sociolinguistics – a subfield of Cognitive Linguistics (see e.g., Geeraerts & Kristiansen 2015 for a concise overview of the field) – even maintains that “a more complete understanding of the usage-based nature of language is only possible if a range of social and cultural factors shaping usage events

- are systematically considered alongside the cognitive ones” (Pütz et al. 2012: 246).
2. *“It’s constructions all the way down!”* (Goldberg 2006: 18). Unlike mainstream generative grammar (e.g., Chomsky 1965, 1995) Cognitive Linguistics does not assume that language consists of lexical items and rules (Goldberg 1995: Chapter 1; Langacker 2008: 5). Rather, it proposes a uniform treatment of abstract morphosyntactic patterns and lexical items (including morphemes) in the form of constructions, form-meaning pairs that provide speakers with the necessary symbolic resources to encode conceptualizations (Croft & Cruse 2004: 257; Langacker 2008: Chapter 1.3.2).
 3. *The generalization and cognitive commitments* (Lakoff 1990). Cognitive Linguistics is committed to describing the general principles that govern all aspects of language (the *generalization commitment*), in accordance with what is known about the functioning of the mind/brain from other disciplines (the *cognitive commitment*). Particularly, Cognitive Linguistics proposes that the human mind is not modular and that in language production – referred to as *coding* or *categorization* – speakers apply nothing but domain-general cognitive capacities that are also used in other tasks, such as, for example, categorization and analogy (Langacker 2008: 8; Bybee 2010: Chapter 1.3).

2.2 Cognitive constraints on coding

Following connectionist models in psycholinguistics (e.g., Dell 1986), Cognitive Linguistics proposes that language production initiates with speakers forming a highly rich conceptualization (Langacker 2008: 31–34). As the conceptualization takes form, domain-general categorization processes compare it to the conceptual import of constructions. In most cases, this rough first pass activates multiple constructions to the degree they match the conceptualization. These start competing for further activation, while also feeding back into the way the conceptualization is structured; this is called *spreading activation* (e.g., Dell 1986; Langacker 2007: 421; 2008: 228–229). Eventually, one construction reaches the highest level of activation and becomes selected to categorize the conceptualization (Langacker 2007: 421; 2008: 228–229).

Of course, given a particular conceptualization, not all constructions will have equal probability of serving as a target for categorization. Since Cognitive Linguistics claims that speakers use domain-general cognitive abilities to retrieve constructions from the network, it seems only fair to assume that domain-general cognitive constraints will also condition the probability of activation of constructions. In this regard, three such factors have been mentioned in the Cognitive Linguistics literature (Langacker 2010: 93): markedness of coding (Langacker 1991: 298), statistical preemption (Goldberg 2006: 94, 2011), and structural priming (Goldberg 2006: 120–125).

Regarding the first of these constraints, the notion of spreading activation entails that the better the conceptualization matches the conceptual import associated with the construction, the more the representation of the construction will become activated. Indeed, in morphosyntax it has been found that a “notion approximating an archetypical conception [tends to be] coded linguistically by a category taking that conception as its prototype” (Langacker 1991: 298). For instance, most speakers of English will prefer *I was hit by a car* (500,000 hits on Google) over *A car hit me* (64,000 hits on Google). This falls out naturally from Langacker’s (1991: 312) schematic definition of the notion of subject as the most conceptually prominent element of the clause, because the first alternative encodes the entity that is most likely to attract the speaker’s attention (i.e., him/herself)

with the grammatical function that signals it as such (i.e., as subject), leading to an optimal correspondence between conceptualization and form. In Cognitive Linguistics, this prototype effect is called *markedness of coding*; *unmarked coding*, referring to a close correspondence between form and meaning, is preferred (Langacker 1991: 298).

A second constraint that influences a representation's level of activation is statistical preemption. This notion indicates that, when the representations of words and constructions are activated frequently together, the compositional expression becomes stored as a single node in the network; this is called *entrenchment* (Bybee 2001: Chapter 5). In turn, because this entrenched expression is more detailed and can be activated faster, it is “preferentially produced over items that are licensed but are represented more abstractly, as long as the items share the same semantic and pragmatic constraints” (Goldberg 2006: 94). This general cognitive constraint has been proposed as a way to explain why speakers do not overgeneralize from input by producing, for example, **stealer* instead of *thief* or **goed* instead of *went* (Goldberg 2006: Chapter 5) and, more generally, why speakers prefer to use grammatical constructions in ways they have predominantly observed them, whereas, in the absence of such experiences, they are perfectly able to accept and produce novel uses of verbs (e.g., Goldberg 2011; Robenalt & Goldberg 2015).

Thirdly, language users tend to pick up and recycle (unintentionally and unconsciously) construction patterns they have (heard) used before, without necessarily repeating the specific words that appear in these structures (e.g., Szmrecsanyi 2008). In the psycholinguistic literature, this tendency is called *structural priming*. Psycholinguistic research into structural priming has revealed that the phenomenon can be accounted for as a residual activation effect: once a particular representation has been visited, it remains more activated than others for a period of time, giving it a head start over its competitors. At the same time, structural priming also appears to be a mechanism of implicit learning, which permanently adapts the ease of activation of constructions to observed patterns of usage (e.g., Goldberg 2006: 120–125; Pickering & Ferreira 2008: 447).

In the following sections, I will review three case studies which demonstrate that these domain-general cognitive constraints predict accurately which linguistic contexts will constrain particular morphosyntactic alternations and with what directionality. To this end, I will start by presenting a case study of existential agreement variation in Caribbean Spanish (Claes 2014a, b, c, 2016). Then, in section 4, I will show that the same theoretical model makes the same accurate predictions for existential agreement variation in British English (Claes & Johnson under review). To show that the model generalizes from agreement variation to other types of morphosyntactic alternations, I will present a third case study of subject personal pronoun expression in Cuban Spanish (Claes under review).

3 Case study I: Existential agreement in Caribbean Spanish

3.1 The phenomenon

In standard Spanish, the existential construction is a subjectless, impersonal structure formed with the verb *haber*. This implies that the NP that appears in this type of sentences is a direct object rather than a subject (as is shown by its accusative pronominalization in example (1)), for which verb agreement does not occur with plural NPs (see example (2)).

(1) (LH15H21/LH1596)

- a. Sí, sí, aquí también los hay.
 yes, yes here as well there.ACC are.SG
 ‘Yes, yes, here there are as well.’
- b. Y yo supongo que los habrá en todos lados.¹

and I suppose that there.ACC will be.SG anywhere
 ‘And I suppose that there will be anywhere.’

- (2) (SJ03H22/SJ327)
 donde era que había fiestas
 where it was that there were.SG parties
 ‘where it was that there were parties’.

However, in all informal varieties of Peninsular (Blas-Arroyo 1995, 2016; Pato 2016; Claes 2017a, b), Canarian (Pérez-Martín 2004), and Latin American Spanish (Vaquero 1996; D’Aquino-Ruiz 2008), speakers variably establish agreement with the NP (as in example (3)).

- (3) (SD04M22/RD437)
 De seguro, no había televisión y, e, no habían computadores.
 surely NEG there was television and er NEG there were.PL computers
 ‘Surely, there was no television and, er, there weren’t computers.’

Earlier investigation of this phenomenon in speech communities worldwide has revealed that agreement with plural NPs occurs more frequently with human-reference NPs, in the imperfect tense, and in the absence of negation. Similarly, the agreement variation has been shown to covary with language-external features such as gender, socioeconomic class and education. Based on these patterns, it has been argued that, at least in Latin American Spanish, the phenomenon constitutes an ongoing change from below geared towards the agreeing forms (D’Aquino-Ruiz 2008; Claes 2014a, b, c, 2015). Let us now consider how a Cognitive Sociolinguistics perspective on this phenomenon may increase our understanding of it.

3.2 Cognitive Constraints at work

Adopting Cognitive Construction Grammar, in earlier work (Claes 2014a, b, c, 2015; 2016) I proposed that the variation between agreeing and non-agreeing *haber* can be conceived of as a competition between two abstract argument-structure constructions. On the one hand, we have the normative construction, which does not have agreement: <ADVP *haber* OBJ>. On the other, we have the variant with a subject, which displays agreement: <ADVP *haber* SUBJ>. Both of these can be considered as nearly synonymous alternatives, except for two conceptual-semantic nuances. Firstly, since Cognitive Linguistics considers that “the grammatical behavior used to identify subject and object do not serve to characterize these notions but are merely symptomatic of their conceptual import” (Langacker 2008: 364), the variant with a subject can be hypothesized to grant more conceptual and formal prominence to the NP, as this is the primary function of subjecthood (Langacker 1991: 294). Secondly, earlier research supports that the social and the stylistic value of the two alternatives is not at all identical. Unfortunately, space limitations impede addressing these aspects of the variation; instead

¹ All Spanish examples were drawn from Claes (2012), a corpus of sociolinguistic interviews that was recorded in Havana, Santo Domingo, and San Juan in March – June 2011. See Claes (2016: Chapter 5) for discussion on the methods that were used in collecting and transcribing the interview data. The codes at the end of the examples represent the following information:

- LH: The token was drawn from the *Havana* section of the corpus (SD: Santo Domingo, SJ: San Juan)
- 15: Participant number 15
- H: Male participant (M: Female)
- 2: 55 + years of age (1: 21–35 years)
- 1: Non-university graduate (2: University graduate)

the reader is referred to Claes (2014a, b, c, 2015, 2016) where these matters are given their due consideration.

Assuming that <ADVP *haber* OBJ> and <ADVP *haber* SUBJ> compete for more or less the same functional space, the cognitive constraints introduced in the previous section can be used to make the following predictions about the Probabilistic Grammar of pluralized *haber*.

- *Markedness of coding*
Cognitively more prominent entities will be encoded more frequently as subject, triggering the use of <ADVP *haber* SUBJ>.
- *Statistical preemption*
If a particular third-person singular tense form of *haber* occurs primarily in <ADVP *haber* OBJ> construction, occurring only sporadically outside of this construction, then this verb tense will disfavor <ADVP *haber* SUBJ>, provided that the conceptual import can be encoded with an entrenched instance of the first construction (i.e., provided that it does not call for aspectual or modal auxiliaries).
- *Structural priming*
If a speaker has just used or processed <ADVP *haber* SUBJ> she will be more likely to use <ADVP *haber* SUBJ> in the following variable context, provided this context occurs within a fairly narrow time window.

Of course, these predictions remain rather abstract, but with additional theoretical support from Cognitive Linguistics, they may be made concrete and specific enough as to be coded into contextual features.

This is especially true for the prediction that refers to markedness of coding, which remains relatively vacuous without a clear definition of what it means for a NP argument to be cognitively prominent. The Cognitive Linguistics literature defines cognitive prominence in relation to the speaker's center of attention: clausal participants on which she has her attention focused are said to be prominent (Langacker 1991: Chapter 7). In turn, Myachykov & Tomlin (2015) show that agents tend to attract more attention than any other type of clausal participants. Therefore, to operationalize markedness of coding, semantic role would be a good candidate.

Nevertheless, the NP of existential expressions cannot be agentive, as the construction presents it as merely being present in a static situation. Still, as argued in earlier work (e.g., Claes 2014a), it is inarguably the case that some entities (say, a lumberjack) are intrinsically more likely than others (say, a tree) to play the agentive role in events. Therefore, with constructions such as existential *haber*, all things being equal, entities like *lumberjack* may be perceived as more potential agents than entities like *tree*, for which the former will be relatively more prominent than the latter.

In Cognitive Linguistics, the semantic roles agent and patient are defined in relation to what Langacker (1991: 283–285) calls the *canonical event model* or the *action-chain model*: the head initiates physical activity, resulting “through physical contact, in the transfer of energy to an external object” (Langacker 1991: 285) and an internal change of state of that entity, the tail of the chain. The semantic roles of agent and patient, in turn, are defined as, respectively, *action-chain head*, and *action-chain tail*. Additionally, events take place in a particular setting, such that the event model minimally includes three elements: action-chain head/agent, action-chain tail/patient, and setting. To classify nouns according to these categories, I relied on the question in (4).

- (4) *Is the referent of the noun highly likely to cause an internal change of state to a second entity without being affected by a third entity first?*
 Yes: typical action-chain head (i.e., more potential agent; e.g., *temblor* ‘earthquake’, *madre* ‘mother’, *carro* ‘car’)
 No: typical action-chain setting or tail (i.e., more potential setting or patient; e.g., *actividad* ‘activity’, *víctima* ‘victim’, *daño* ‘damage’)

Another linguistic feature that correlates closely with speaker’s selective attention is definiteness and specificity (Langacker 1991: Chapter 7). However, because of the discourse function of existential/presentational expressions – which serve to present unknown entities to the hearer (Lakoff 1987: Case Study 3) – the NPs of affirmative existential expressions can only refer to specific indefinite referents (Prince 1992). However, when we negate the existence of a specific entity with an utterance such as *There are no bears in Puerto Rico*, we suspend the reference of the NP *bears* (Keenan 1976: 318) and a generic expression emerges, which can be paraphrased as “the category *bears* does not exist in Puerto Rico”. In other words, under negative polarity, the NP becomes “identifiable only as a type, not as a specific instance or token” (Croft 2003: 132), for which it will be less likely to attract the speaker’s attention (Langacker 1991: 308). Therefore, markedness of coding was operationalized further by coding for polarity.

Let us turn now to statistical preemption. Operationalizing this constraint requires some metric that expresses the relative degree of entrenchment of the different tense forms of *haber* in $\langle \text{ADV}P \text{ haber OBJ} \rangle$. For this case study, I will rely on ΔP (*delta-P*), a measure derived from associative learning theory (e.g., Ellis & Ferreira-Junior 2009). Applied to Spanish existential agreement variation, this metric expresses the probability of observing a third-person singular form of *haber* in the presence of $\langle \text{ADV}P \text{ haber OBJ} \rangle$ minus the probability of observing this form in the absence of that construction. To establish these probabilities, for each form of the verb, I calculated the frequency scores described schematically in Table 1.

With a two-by-two collocations table like Table 1, ΔP may be calculated with the following formula:

$$(5) \quad \Delta P = (\text{Cell A}/(\text{Cell A} + \text{Cell B})) - (\text{Cell C}/(\text{Cell C} + \text{Cell D}))$$

Of course, for this measure to be meaningful, it must be based on frequency counts derived from a large corpus that contains samples of multiple registers of both spoken and written language. Therefore, I turn to the twentieth-century section of the *Corpus del español* (20 million words; Davies 2002-) as an ancillary data source.

Table 1: Collocations table.

Cell A	Cell C
Frequency of word <i>W</i> in construction <i>Cx</i> e.g., frequency of $\langle \text{Adv}P \text{ hubo OBJ} \rangle$	Frequency of words other than <i>W</i> in construction <i>Cx</i> e.g., frequency of $\langle \text{Adv}P \text{ haber OBJ} \rangle$ with forms other than <i>hubo</i>
Cell B	Cell D
Frequency of word <i>W</i> in constructions other than <i>Cx</i> e.g., frequency of non-existential cases of <i>hubo</i>	Frequency of words other than <i>W</i> in constructions other than <i>Cx</i> e.g., frequency of non-existential third-person singular forms of <i>haber</i> other than <i>hubo</i>

The resulting ΔP scores are presented in Figure 1. These data suggest that present-tense *hay*² and preterit-tense *hubo* rarely occur outside of $\langle \text{ADV}P \textit{ haber OBJ} \rangle$, whereas all other forms are either neutral with respect to their preference for occurring in or outside this construction or, in the case of imperfect *había*, display a marked preference for occurring outside of this construction. Also, the ΔP scores support that *hay* and *hubo* are more than twice as deeply entrenched in $\langle \text{ADV}P \textit{ haber OBJ} \rangle$ as any other form of the verb. Therefore, the specific prediction that follows from statistical preemption is that the present and the preterit tense will disfavor $\langle \text{ADV}P \textit{ haber SUBJ} \rangle$, unless encoding the conceptualization requires aspectual or modal auxiliaries, which would bypass the entrenched instances of $\langle \text{ADV}P \textit{ hay OBJ} \rangle$ or $\langle \text{ADV}P \textit{ hubo OBJ} \rangle$. For this reason, statistical preemption was operationalized as: present and preterit tense without aspectual/modal auxiliaries vs. all others.

Finally, as for structural priming, the data were coded for the type of last token that was uttered by the interviewer (comprehension-to-production priming) and the participant (production-to-production priming) and the number of conjugated verbs that occur between these tokens and the case at hand. Since the initial results displayed long-lasting priming effects independent of lexical repetition, structural priming was operationalized as follows: first occurrence/distance 20+ clauses, primed with $\langle \text{ADV}P \textit{ haber SUBJ} \rangle$, and primed with $\langle \text{ADV}P \textit{ haber OBJ} \rangle$.

3.3 Data and methods

To test how these operationalized cognitive constraints impact the competition between $\langle \text{ADV}P \textit{ haber OBJ} \rangle$ and $\langle \text{ADV}P \textit{ haber SUBJ} \rangle$ I analyze a collection of 3×24 recording sessions with native-speaker residents of Havana (Cuba), Santo Domingo (the Dominican Republic), and San Juan (Puerto Rico), comprising some 78 hours of speech (Claes 2012). The data are stratified by age (21–35 years vs. 55+ years), education (no university degree vs. university degree), gender (female vs. male), and data elicitation method (interview, sentence completion task in story, and sentence completion task in questionnaire).

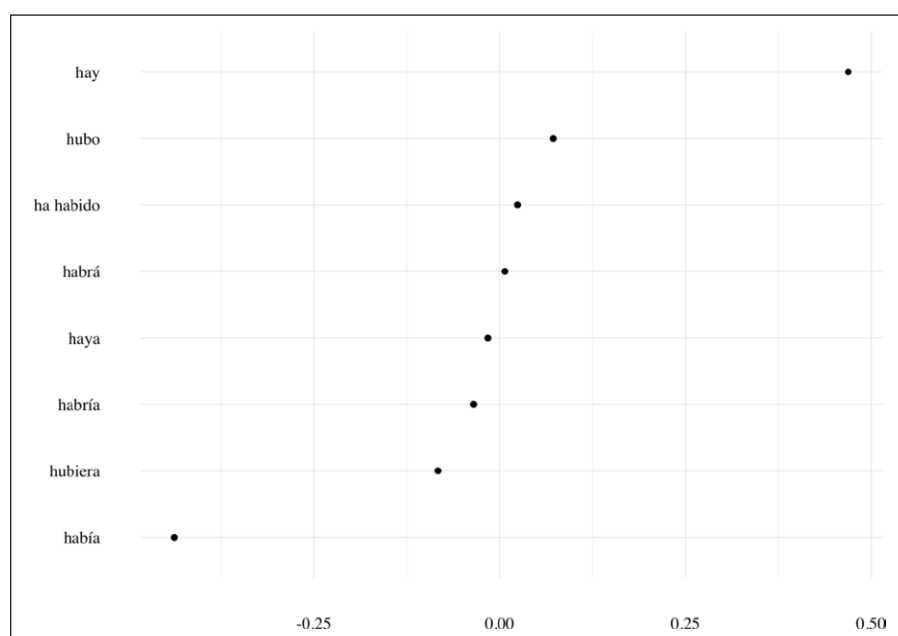


Figure 1: ΔP scores for the different tensed forms of *haber*.

² While rare, the vernacular plural form *hayn* does occur in Latin American Spanish (e.g., Vaquero 1996).

After transcription, all cases of presentational *haber* followed by a plural NP were manually selected from the transcripts and coded for the predictors described in the previous section. This yielded a database of 5,589 eligible tokens, which was explored with by-city parallel mixed-effects logistic regression analyses using the *lme4* package (Bates et al. 2016) for R (R Core Team 2016). In these analyses, the tokens of presentational *haber* were grouped together according to the speakers and the nouns that appear in the token.

Model selection was informed by second-order Akaike Information Criterion (AIC_c in the *MuMIn* R package; Bartón 2015), which is a sample-size-adjusted measure that expresses how useful the information provided by the candidate model is for predicting plural agreement (Burnham & Anderson 2002: 66). To select a parsimonious model, I started out with full models including the random intercepts, the demographic information recorded by the corpus, as well as all the predictors described in the previous section. Then, I ran and evaluated candidate models for all possible subsets of these fixed effects, using the *pdredge* function of the *MuMIn* package. The output of this model selection procedure is a list of candidate models ordered by their AIC_c score. The model with the lowest AIC_c value was selected as the starting point in the posterior model fitting process. To evaluate whether interactions and random slopes improved the model fit, those were added one at a time. If the addition of an interaction or slope lowered the AIC_c value of the model, it was included in the final model, provided the candidate model converged and did not overfit the data (evaluated with confidence intervals based on 1,000 bootstrap repetitions with the *confint* function of the *lme4* package). For the Santo Domingo and the San Juan data, adding interactions did not result in better fits. For Havana, an interaction between tense and data collection method was detected. By-speaker random slopes for typical action-chain position improved the fits of the Cuban and the Dominican models, as did a by-speaker random slope for tense for the Puerto Rican model. By-noun random slopes did not converge, which is probably due to the Zipfian distribution of nouns in naturalistic discourse.

3.4 Results

When it comes to markedness of coding, the regression results in Tables 2–4 show that speakers of all three Caribbean varieties of Spanish are more likely to use plural presentational *haber* when the NP refers to an entity that can easily be imagined as the starting point of a series of events, such as *alumnos* ‘students’ in example (6). The magnitude of this effect also appears to be more or less the same for the three varieties.

- (6) (LH10M22/LH1261)
 habían por lo menos veinte y pico, treinta alumnos
 there were.PL at least twenty-some thirty students
 ‘there were at least twenty-some, thirty students’.

In contrast, polarity did not contribute useful information for modeling the variation in Dominican Spanish, whereas speakers of Cuban and Puerto Rican Spanish disfavor plural agreement under negative polarity with a similar effect size. When an alternative regression model is fitted to the Dominican data, the same directionality of effect is obtained for polarity. Yet, as is shown by Figure 2, which plots the effect of polarity on the Log Odds of plural agreement, the effect size of this predictor is minimal for this variety (0.091 Log Odds). These results support that markedness of coding conditions existential agreement variation.

Turning now to statistical preemption, Tables 2–4 show that speakers of Cuban, Dominican and Puerto Rican Spanish are much less likely to use plural presentational *haber* for the present and the preterit tense, whereas they favor agreement for all other tenses. This suggests that the entrenched instances <ADVP *hay* OBJ> or <ADVP *hubo* OBJ> block

Table 2: Logistic generalized linear mixed-effects model of presentational *haber* pluralization in Havana (sum contrasts): numbers, percentages, and coefficients for pluralized presentational *haber*.³

Fixed effects	Havana		
	N	%	Coefficient
(Intercept)			-1.023
Verb tense			
All others	819/1298	63.1	1.663
Synthetic expressions in present or preterit tense	115/795	14.5	-1.663
Production-to-production priming			
Pluralized presentational <i>haber</i> construction	556/817	68.1	0.653
First occurrence/distance 20+ clauses	83/297	27.9	-0.268
Singular presentational <i>haber</i> construction	295/979	30.1	-0.385
Comprehension-to-production priming			
Pluralized presentational <i>haber</i> construction	113/239	47.3	0.503
Singular presentational <i>haber</i> construction	73/204	35.8	-0.151
First occurrence/distance 20+ clauses	748/1650	45.3	-0.353
Typical action-chain position of the noun's referent			
Heads	467/925	50.5	0.248
Tails and settings	467/1168	40.0	-0.248
Polarity			
Positive	708/1523	46.48	0.188
Negative	226/570	39.65	-0.188
Model summary			
C-index of concordance			0.89
Pseudo-R ²			0.60
AIC _c			1974.9

the use of the more abstract pattern <ADVP *haber* SUBJ>. Further evidence for this claim can be found when we compare the distribution of present- and preterit-tense existentials that involve aspectual or modal auxiliaries (see example (7)) with those that do not.

- (7) (SD20H12/RD2706)
 Pueden haber expresiones que, que tengan una acepción diferente
 there-can-be.PL expressions that that may-have a different meaning
 ‘There can be expressions that, that may have a different meaning.’

³ Besides the predictors that appear in this table, the regression model also includes education, data collection method and a data collection method × tense interaction. Space restrictions inhibit us from discussing these results. The reader is kindly referred to Claes (2016: Chapter 8) for a discussion of the social covariates of *haber* pluralization in Cuban Spanish. In computing the models the *bobyqa* optimizer for *glmer* was used.

Table 3: Logistic generalized linear mixed-effects models of presentational *haber* pluralization in Santo Domingo (sum contrasts): numbers, percentages, and coefficients for pluralized presentational *haber*.⁴

Fixed effects	Santo Domingo		
	N	%	Coefficient
(Intercept)			-0.224
Verb tense			
All others	720/1103	65.3	1.446
Synthetic expressions in present or preterit tense	140/739	18.9Î	-1.446
Production-to-production priming			
Pluralized presentational <i>haber</i> construction	484/711	68.1	0.780
First occurrence/distance 20+ clauses	123/337	36.5	-0.125
Singular presentational <i>haber</i> construction	253/794	31.9	-0.654
Comprehension-to-production priming			
Pluralized presentational <i>haber</i> construction	151/264	57.2	0.507
Singular presentational <i>haber</i> construction	63/185	34.1	-0.189
First occurrence/distance 20+ clauses	646/1393	46.4	-0.317
Typical action-chain position of the noun's referent			
Heads	439/815	53.9	0.463
Tails and settings	421/1027	41.0	-0.463
Polarity			
Positive			
Negative			
Model summary			
C-index of concordance			0.87
Pseudo-R ²			0.54
AIC _c			1844.0

This is represented in Figure 3, which shows that across the board, plural agreement is much more frequent when such additional constructions are present. As a matter of fact, the rate of plural agreement that is observed with aspectual and modal auxiliaries is virtually identical to the one that is documented for any other tense. This seems to confirm that the entrenched instances of <ADVP *hay* OBJ> and <ADVP *hubo* OBJ> only preempt the use of the more abstract construction <ADVP *haber* SUBJ> when both could encode the conceptualization equally well, as is predicted by statistical preemption.

Turning now to structural priming, Tables 2–4 support that once speakers have used a particular variant of the presentational construction with *haber*, they are much more likely

⁴ Besides the predictors that appear in this table, the regression model also includes gender. Space restrictions inhibit us from discussing these results. The reader is kindly referred to Claes (2016: Chapter 8) for a discussion of the social covariates of *haber* pluralization in Dominican Spanish. In computing the models the *bobyqa* optimizer for *glmer* was used.

Table 4: Logistic generalized linear mixed-effects models of presentational *haber* pluralization in San Juan (sum contrasts): numbers, percentages, and coefficients for pluralized presentational *haber*.⁵

Fixed effects	San Juan		
	N	%	Coefficient
(Intercept)			-0.974
Verb tense			
All others	622/1014	61.3	1.766
Synthetic expressions in present or preterit tense	62/641	9.7	-1.766
Production-to-production priming			
Pluralized presentational <i>haber</i> construction	352/558	63.1	0.597
First occurrence/distance 20+ clauses	88/246	35.8	-0.155
Singular presentational <i>haber</i> construction	244/851	28.7	-0.442
Comprehension-to-production priming			
Pluralized presentational <i>haber</i> construction	92/175	52.6	0.452
Singular presentational <i>haber</i> construction	30/125	24.0	-0.266
First occurrence/distance 20+ clauses	562/1355	41.5	-0.186
Typical action-chain position of the noun's referent			
Heads	350/773	45.3	0.418
Tails and settings	348/882	37.9	-0.418
Polarity			
Positive	559/1225	45.6	0.341
Negative	125/430	29.1	-0.341
Model summary			
C-index of concordance			0.89
Pseudo-R ²			0.62
AIC _c			1517.9

to use the same constructional alternative in the following variable context, provided this context occurs within a fairly narrow time window. A similar, but smaller effect is observed for comprehension-to-production priming.

Finally, the model summaries at the bottom of Tables 1–3 support that the models that were presented in this section perform highly accurately at modeling speakers' behavior for the three speech communities. For each, a C-index in the high eighties is observed, suggesting excellent discriminative ability (Hosmer & Lemeshow 2000: 162). Also, the models capture more than 50% of the variability that is observed in the three datasets, as is shown by the Nakagawa & Schielzeth (2013) conditional pseudo-R² values.

⁵ Besides the predictors that appear in this table, the regression model also includes gender and education. Space restrictions inhibit us from discussing these results. The reader is kindly referred to Claes (2016: Chapter 8) for a discussion of the social covariates of *haber* pluralization in Puerto Rican Spanish. In computing the models the *bobyqa* optimizer for *glmer* was used.

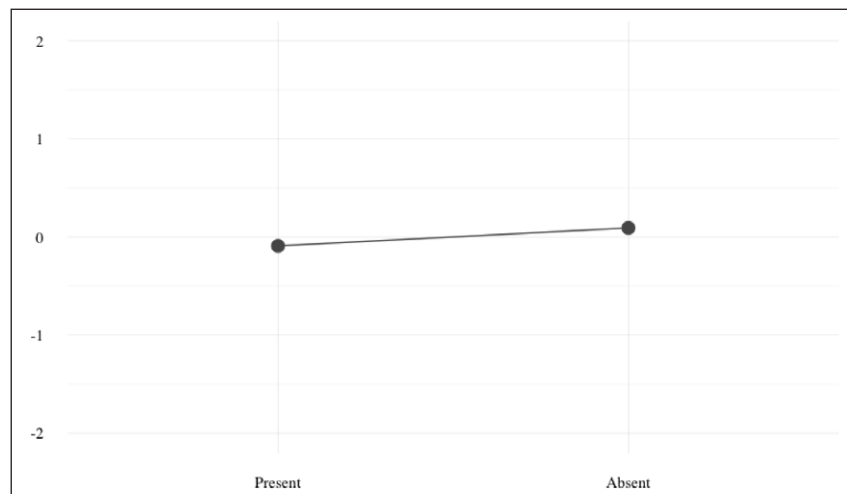


Figure 2: Plot of the effect of polarity on the Log Odds of plural presentational *haber* in Dominican Spanish.

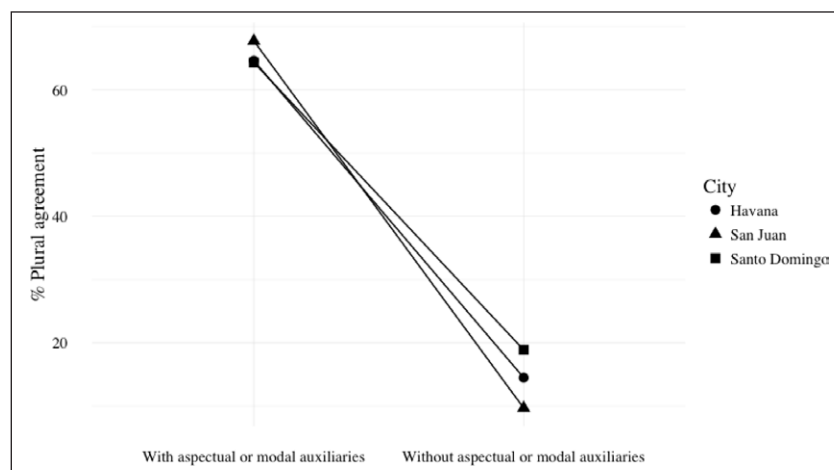


Figure 3: Percentages of plural agreement in the present and the preterit tense: with aspectual/modal auxiliaries vs. without aspectual/modal auxiliaries.

3.5 Discussion

In this first case study, robust priming effects were uncovered from production to production and from comprehension to production. Since priming effects without lexical repetition are generally accepted as evidence that expressions depend on shared cognitive representations (e.g., Goldberg 2006: 120–125), this supports that *haber* pluralization involves a competition between two distinct presentational *haber* constructions. Also, because highly similar tendencies were uncovered for the independent variables, which pattern as predicted by the hypotheses, the results of the first case study are highly favorable to the view that probabilistic patterns in language (variation) reflect domain-general cognitive constraints on language production: markedness of coding, statistical preemption, and structural priming.

However, the promising results of this section were obtained for closely related varieties, for which they do not provide the necessary evidence to support such a far-reaching claim. If existential agreement variation and morphosyntactic variation generally is indeed governed by these domain-general cognitive constraints, similar if not identical patterns should be found for a related phenomenon in another language. This will be the topic of the second case study.

4 Case study II: Existential agreement variation in British English

To test whether the model of (existential agreement) variation described in section 2 generalizes from the original case study for which it was developed to other types of alternations, Claes & Johnson (under review) contrast the patterns of existential agreement variation in British English with the data presented in the previous section.

4.1 The phenomenon

Whereas the Spanish presentational construction is a non-agreeing structure that displays variable agreement, English existential *there to be* is an agreeing construction that shows variable non-agreement in the present and the past tense, as is shown in examples (8) and (9).

(8) *There are foxes alive and well in Bristol (BNC, JNB, PS4C6, 582).*⁶

(9) *There is and there was cairns where they used to rest the, the coffin (BNC, FXP, PS238, 31)*

Research into this phenomenon typically finds some effects of polarity, tense, contractedness, determiner type, and the distance between the form of *to be* and the noun (e.g., Tagliamonte & Baayen 2012). For the first of these constraints, polarity, location-specific patterns have been found. For instance, Tagliamonte (1998) indicates that the presence of sentence negation disfavors the absence of agreement. This is also the effect that is generally documented in the U.S. (Tagliamonte & Baayen 2012), but in New Zealand English polarity does not seem to constrain the variation (Hay & Schreier 2004). In turn, for tense, contractedness, and distance, similar results have been documented in English worldwide: the present tense, the full forms of *to be*, and shorter distances between *to be* and its noun favor agreement over its absence (e.g., Hay & Schreier 2004; Crawford 2005; Tagliamonte & Baayen 2012). As for the language-external covariates of this alternation, research supports that non-agreement is more common in informal registers and appears to correspond to an ongoing change towards non-agreement (Tagliamonte & Baayen 2012). Let us now consider how the cognitive constraints described above allow us to construct a predictive, probabilistic model of existential agreement variation in English.

4.2 The cognitive constraints at work

As was the case for existential *haber*, both agreeing and non-agreeing *there be* can be considered to fulfill the same function in discourse. For agreeing *there be*, we can hypothesize that the construction treats the postverbal nominal as a subject and *there* as a grammaticalized adverbial, similar to *y* in French *il y a*, *hi* in Catalan *hi ha*, or *ci* in Italian *ci'è*. Schematically, this would yield the following structure: < *There be* **SUBJ** >. In turn, the variant without agreement can be conceptualized as treating *There* as an impersonal subject (similar to *it* in *it's a dog* or *it's raining*), and the nominal as a complement to that impersonal subject (< *There.IMP-SUBJ be* **COMP** >).

Assuming this competition, we can propose that markedness of coding and structural priming will have exactly the same effects in British English as in Caribbean Spanish.

⁶ The English examples were drawn from the British National Corpus (British National Corpus Consortium 2007). The codes should be interpreted as follows:

- *BNC*: British National Corpus
- *JNB*: Informant JNB
- *PS4C6*: Text PS4C6
- *582*: Sentence unit 582

That is, markedness of coding predicts that more prominent NPs (in terms of polarity and typical action-chain position) will tend to be encoded as subject, favoring the <There *be* **SUBJ**> construction. Also, as was the case for presentational *haber*, structural priming leads to the expectation that speakers will tend to use this variant more often in contexts following an agreeing *there to be* sentence. As for statistical preemption, applying the ΔP metric to frequency data culled from the full *British National Corpus* (British National Corpus Consortium 2007) yields the estimates provided in Figure 4. These data show that, whereas *are* is strongly attracted to the agreeing presentational construction, *were* is not. Therefore, the prediction that follows from statistical preemption is that the past tense will disfavor agreement, while the present will favor it.

4.3 Data and methods

The English *there is/are/was/were* data were culled from the 10.4-million-word spoken component of the *British National Corpus* (British National Corpus Consortium 2007). As the initial search returned 34,197 hits of *there* plus any form of *be*, Claes & Johnson (in evaluation) filtered the data in R to make manual inspection and coding possible. This involved the reduction of sets of adjacent tokens (e.g., *I think from that report there was a, there was a requirement or request...*) to one token each and the exclusion of compound forms (e.g., *will be, should be, etc.*), as well as tokens that did not have a plural noun. Contracted forms (e.g., *there's*) were also excluded from the dataset, as earlier research (e.g., Crawford 2005; Walker 2007) gives reason to believe that these forms constitute chunks (in the sense of Bybee 2010), which act as formulaic sequences that have grown to be largely independent from other cases of *there be*. Applying these filters resulted in a total of 1,932 tokens, to which the same analytical procedures were applied as those described in section 3.3.

4.4 Results

For markedness of coding Table 5 shows that, contrary to the data that were obtained for Caribbean Spanish, typical action-chain position did not turn out to be a relevant predictor for British English. Indeed, an alternative model that includes this predictor estimates the size of the effect to be only 0.073 Log Odds. Still, as Figure 5 shows, the directionality of the effect remains identical with respect to Caribbean Spanish, as typical action-chain heads favor agreement.

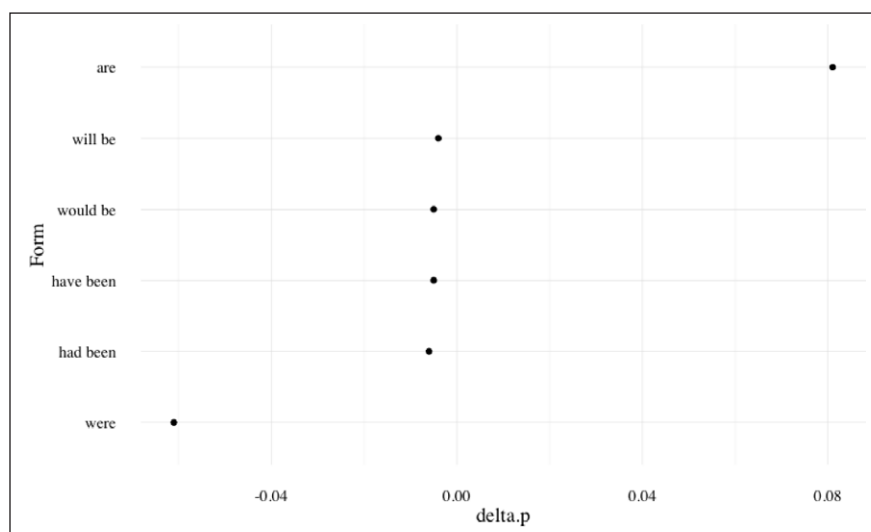


Figure 4: ΔP scores for different forms of *to be* in the *British National Corpus* (British National Corpus Consortium 2007).

Table 5: Logistic generalized linear mixed-effects model of agreement with *there be* in plural existentials in British spoken English (sum contrasts): numbers, percentages, and coefficients for agreeing *there be*.⁷

Fixed effects	N	%	Coefficient
(Intercept)			2.595
<i>Verb tense</i>			
Past	250/696	64.1	-1.403
Present	1203/1236	97.3	1.403
<i>Production-to-production priming</i>			
Agreeing <i>There be</i>	309/331	93.4	0.476
First occurrence/distance 10+ sentence units	925/1066	86.8	0.116
Non-agreeing <i>There be</i>	415 /535	77.6	-0.591
<i>Polarity</i>			
Positive	1482/1712	86.6	0.219
Negative	167/220	75.9	-0.219
Model summary			
Pseudo-R ²			0.62
C-index			0.95
AIC _c			1100.7

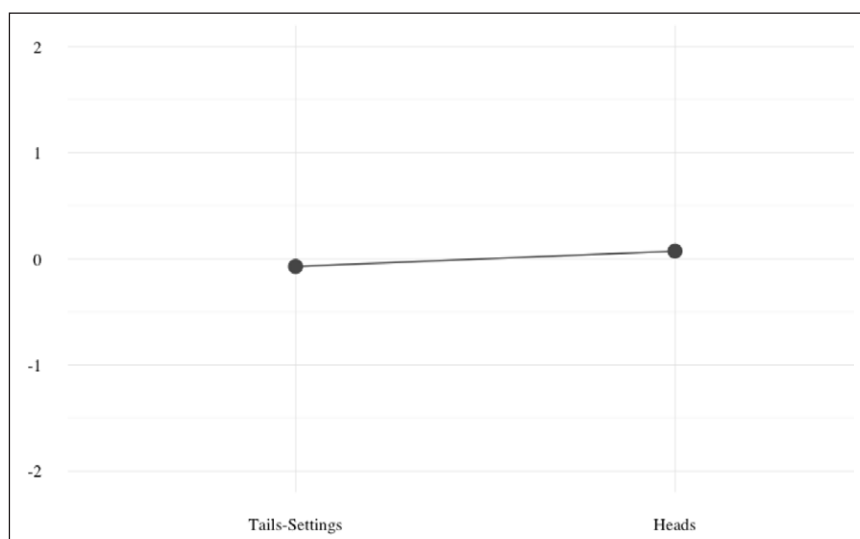


Figure 5: Plot of the effect of typical action-chain position on the Log Odds of plural existential agreement in British English.

Turning now to the effects of polarity, Table 5 shows that, as was the case in Caribbean Spanish, speakers of British English establish agreement more often with *there to be* when the clause has positive polarity, whereas they are more inclined to use a non-agreeing existential when negation is present.

⁷ Also included in the final model: domain/register. In the spoken British National Corpus, as is observed throughout the Anglophone world, agreeing *there be* appears to be more common in formal language, whereas non-agreeing *there be* is more common in informal language (Tagliamonte & Baayen 2012).

When it comes to statistical preemption, Table 5 indicates that speakers are far more likely to use the *<There be SUBJ>* construction in the present tense (at least when contracted cases are not taken into account), whereas they are more likely to use *<There. IMP-SUBJ be COMP>* for the past tense. Since Figure 4 indicates that the present tense is far more entrenched in the agreeing existential construction, this supports that speakers use the entrenched instance *<There are SUBJ>* whenever they see the chance, as statistical preemption would predict.

Regarding structural priming, Table 5 shows that speakers are more likely to use *<There be SUBJ>* when they have just used an instance of this construction. Unfortunately, the British National Corpus data do not allow exploring comprehension-to-production priming, as only 40 examples occurred within a twenty-clause window after another speaker included in the corpus had used an existential expression.

Finally, as was the case for Caribbean Spanish, the model summary at the bottom of Table 5 indicates that the model has outstanding discriminative ability ($C > 0.90$; Hosmer & Lemeshow 2000: 162). Also, the conditional pseudo- R^2 value indicates that the predictors capture more than 50% of the variability. Both of these metrics suggest that the model taps into the constraints to which speakers are sensitive.

4.5 Discussion

The second case study has revealed that existential agreement variation in both Caribbean Spanish and British English is sensitive to the same contextual features, which constrain the variation with the same directionalities of effects. Despite these qualitative similarities, a considerable quantitative difference was documented between the effect sizes that are found in the two languages for typical action-chain position. This difference may relate to the fact that in Spanish the contrast between the competing constructions is one of object versus subject, which are maximally distinct grammatical functions (Langacker 1991: Chapter 7). In turn, in English, the alternation involves a contrast between subject and complement, which are conceptually and formally distinct grammatical functions, but not exactly each other's opposites. Therefore, it is partially predictable that markedness of coding may have stronger effects in Spanish than in English.

In any case, the qualitative correspondences that were uncovered in these first two case studies between the Probabilistic Grammars of existential agreement variation in three varieties of Spanish and in British English are too striking to be coincidental. Because these correspondences extend across two distinct languages and multiple speech communities, it is problematic to claim that the similarities in the behavior of speakers simply emerge from their similar experiences with language. Rather, they contribute strong support for the claim that (existential agreement) variation is constrained by the domain-general cognitive constraints on spreading activation that were sketched in section 2. Let us now consider whether this model can withstand the test of being applied to a totally different type of alternation. This will be the topic of the third and final case study.

5 Case study III: Subject personal pronoun expression in Spanish

The previous two case studies have shown that the model of morphosyntactic variation that was outlined in section 2 can account for the patterns of existential agreement variation in English and Spanish. In this case study, we will move beyond this type of alternations, in an attempt to evaluate whether the model generalizes to morphosyntactic variation at large. To this end, I will perform a case study of Spanish subject pronoun expression (SPE) in Cuban Spanish (Claes under review).

5.1 The phenomenon

It is a well-known property of Spanish to be a so-called *pro-drop* language. This means that the rich verb morphology allows speakers not only to signal tense, aspect, and mood, but also person and number. As a result, the expression of pronominal subjects is optional, as is shown in examples (10) and (11).

(10) (LH01M22/16–17)
 también conozco, de cierta manera, conozco a las personas
 as well know.1-SG in a way know.1-SG the people
 ‘I also know, in a way, I know the people’.

(11) (LH01M22/258)
 No era, no, no fue porque yo la escogí.
 NEG was NEG NEG was because I it.ACC picked
 ‘It wasn’t, it, it wasn’t because I picked it.’

This is a phenomenon of which the Probabilistic Grammar is generally assumed to be very well understood; Bayley et al. (2013: 22) even label it as a *showcase variable* of variationist sociolinguistics. Indeed, Carvalho et al. (2015: 13) present the case of Spanish SPE as an example that “lends support to the notion that structured linguistic variation is an intrinsic part of speakers’ grammatical knowledge”. In both cases, the optimism stems from the fact that studies typically find a highly consistent, recurring pattern of contextual conditioning, with SPE being favored by first- and second-person singular subjects, non-generic subjects, potentially ambiguous verb tenses, priming, stative (e.g., *estar* ‘to be located’) and cognitive (e.g., *creer* ‘to think’) verbs, independent main clauses, and changes in the reference of the subject with respect to the subject of the previous clause.

While it is true that these predictors produce highly similar results for most monolingual and bilingual communities considered to date (see e.g., Flores-Ferrán 2007 and Carvalho et al. 2015 for overviews), recent work suggests that they model only a small portion of the variability. For example, Otheguy & Zentella (2012) report that their models capture some 18% of the variance ($R^2 = 0.18$). In part, this may be due to Zipfian biases and the skewing caused by high-frequency pronoun-verb collocations, which have only recently begun to attract the attention of researchers working on SPE (e.g., Travis & Torres-Cacoulllos 2012; Orozco 2015; Posio 2015). Also, earlier work has not reached a consensus on how to explain the effects of these predictors, which is mostly due to the fact that the common denominator of the results does not support any analysis completely (Travis 2007). Let us now investigate whether the cognitive constraints model could bring a change to this.

5.2 The cognitive constraints at work

A first issue that needs to be addressed is how to portray SPE. In this regard, Travis & Torres-Cacoulllos (2012) propose that the variation between the presence and the absence of subject personal pronouns (SPPs) involves the variable realization of the SPP in an invariant <(SPP) Verb> construction. However, this view presupposes that the default alternative is the presence of a SPP, and that the deviation that only occurs under certain discourse conditions (cf. Goldberg 2005) is its absence. This is not at all compatible with the *pro-drop* nature of Spanish, where the default alternative is the absence of pronominal subjects.

Therefore, this case study proposes a competition between two distinct constructions: <Verb> and <SPP Verb>. Both of these constructions can be assumed to do the same referential work, but, as was the case for agreement variation, they can be assumed to be minimally distinct in relation to their social and stylistic meaning and when it comes to

the relative prominence they attribute to their subjects. That is, since the <SPP Verb> construction encodes the subject explicitly through a pronoun, it can be assumed that this construction grants somewhat more prominence to this constituent.

When we apply markedness of coding, statistical preemption, and structural priming to this working hypothesis, we obtain a series of detailed predictions about the absence/presence of SPPs in discourse:

- *Markedness of coding*
<SPP Verb> will be favored with conceptually more prominent subjects.
- *Statistical preemption*
The strongest mental representations of verb forms can be ranked on a continuum ranging from <SPP Verb>-based collocation to <Verb>-based collocation. Towards the extremes of the continuum, markedness of coding and structural priming will only make a minor contribution to explaining the variation, as speakers will generally favor the collocation.
- *Structural priming*
Producing/processing <SPP Verb> or <Verb> will incite speakers to use the same construction in the following variable context, regardless of variations in grammatical person and number, tense-aspect-mood, and verb, provided this second context falls within a fairly narrow time window after the first.

As for the earlier two case studies, these somewhat abstract predictions need to be operationalized before they can be tested on corpus data. To this end, we can again turn to Cognitive Linguistics.

When it comes to markedness of coding, it was already mentioned above that Cognitive Linguistics assumes that cognitive prominence coincides rather closely with speakers' center of attention. In this regard, Langacker (1991: Chapter 7) argues that speakers are mostly concerned with themselves and their interlocutors, whereas they are less likely to focus attention on others. Therefore, I coded the data for the empathy hierarchy (*speaker* > *hearer* > *other*; Langacker 1991: 305). For this variable, markedness of coding predicts that referents that refer to the hearer and the speaker are more likely to be encoded as SPPs than other referents. Additionally, since speakers are more likely to focus attention on highly agentive participants (Myachykov & Tomlin 2015), I also coded for Lakoff's (1977) agentivity features volitionality and referentiality; other features (e.g., responsibility, control) were not withheld as these proved to be either nearly impossible to code consistently or to collide with others. For volitionality and referentiality, markedness of coding predicts that subjects that are explicitly portrayed as volitional in the event and subjects that refer to concrete referents will more likely be encoded with <SPP Verb>. Additionally, the literature suggests that highly transitive clauses (e.g., *John hit Pat*) foreground the event (Hopper & Thompson 1980: 253; Lakoff 1977: 244), for which markedness of coding predicts SPE to be less frequent with this type (cf. Posio 2011). To test this prediction, I coded for the transitivity features proposed by Hopper & Thomson (1980) that are most closely related to the event (aspect, kinesis, and individuation of the object, coded here as animacy; punctuality was discarded, as it collides with kinesis).⁸ Finally, on a discourse

⁸ Hopper & Thompson (1980) approach aspect as a binary distinction between imperfective/atelic and perfective/telic. However, since the goal is to model the amount of attention that is turned to the predicate, a three-way distinction between continuous, imperfective, and perfective aspect may be more appropriate. That is, continuous aspect is a type of imperfective aspect that presents the event in its course, for which it favors the focusing of attention on the event rather than the subject.

Table 6: Collocations table.

Cell A	Cell B
<i>Corpus del español</i> frequency of the verb form with its corresponding SPP (e.g., <i>yo creo</i>)	<i>Corpus del español</i> frequency of all other tokens of <SPP Verb> (e.g., <i>él cree, yo bailo, ella duerme</i>)
Cell C	Cell D
<i>Corpus del español</i> frequency of the verb form outside of <SPP Verb> (e.g., <i>creo</i>)	<i>Corpus del español</i> frequency of all other verb forms outside of <SPP Verb> (e.g., <i>trabajo, dice, corre</i>)

level, newly introduced referents can be hypothesized to attract relatively more attention (and, hence, to be expressed more frequently with <SPP Verb>) than well-established, continuous topics. Therefore, as other investigations of SPE in Spanish (e.g., Bayley et al. 2013; Otheguy & Zentella 2012; Shin 2014), I also coded for the referential continuity of the subject, as well as the distance (in number of conjugated verbs) between coreferential subjects.

Hypothesis 2 claims that with certain verb forms speakers will preferentially use either a <Verb>-based or a <SPP Verb>-based collocation. Therefore, to test the effects of statistical preemption, I performed *distinctive collexeme analyses* (e.g., Stefanowitsch & Gries 2005) on frequency data culled from the 20-million 20th century section of *Corpus del español* (Davies 2002-). Applied to SPE, this type of analysis consists in calculating the positive/negative base-ten logarithm of a *p*-value obtained with a Fisher-Yates Exact test for Table 6, depending on whether or not the observed frequency of Cell A exceeds its expected frequency. The further the resulting *collostruction strength* deviates from zero, the stronger the association between the verb form and either <SPP Verb> (positive strengths) or <Verb> (negative strengths; Levshina 2015: 232, 242–243).

Finally, to test for priming effects, the tokens were coded for the type of last variant that was used by the speaker (production-to-production priming) and the hearer (comprehension-to-production priming), the distance between the primes and the targets, and whether or not the speaker repeated the exact same verb form. Since clauses with human reference subjects tend to occur in clusters, the maximum distance for priming effects was set to five clauses. Also, although lexical repetition increased the magnitudes of the priming effects – a recurrent finding in structural priming research (Pickering & Ferreira 2008) – the levels were collapsed into broader categories, as we are currently only interested in establishing whether or not SPE can be primed.

5.3 Data and methods

The data for this case study were culled from the interview sections of the Havana subsection of Claes (2012). The interviews were part-of-speech tagged with the *Stanford POS Tagger* (Toutanova et al. 2003) to allow for the semi-automatic extraction and annotation of conjugated verbs. Subsequently, all pronominal and non-overt human-reference subjects were filtered out and annotated semi-manually for semantic and formal predictors.

To be eligible, verbs had to occur in contexts where both subject pronouns and verbal markings alone could occur. Following Otheguy & Zentalla (2012) and Otheguy et al. (2007), this implied that verbs with impersonal and inanimate subjects were not considered as instances of the variable (e.g., meteorological verbs such as *llueve* ‘it rains’; existentials with *haber* e.g., *hay cosas* ‘there are things’ and *hacer* e.g., *hace años* ‘years ago’, as well as *se*-passives e.g., *se pide ayuda* ‘help is requested’). Verbs that occurred with a lexical subject (e.g., *Marta pide ayuda* ‘Martha requests help’) or in a subject-headed relative (*El que con cojos anda* ‘He who walks with cripples’) were not included as instances of the variable either. In turn, contrastive contexts, which have been excluded from some earlier

studies of subject pronoun expression, were included in the corpus provided the pronoun was not the focus of contrast (Matos-Amaral & Schwenter 2005). Also, following Otheguy & Zentella (2012), verb forms accompanied by a topicalized SPP (e.g., *Yo lo que Ø quiero es* ‘I what [I] want is’) were coded as bare verbs, as it would have been possible to insert a pronoun directly before the verb (e.g., *Yo lo que yo quiero es* ‘I what I want’). However, the approach of this paper diverges from Otheguy & Zentella’s (2012: 234–235) in not excluding verb forms that occur in highly fixed set phrases (e.g., *no sé* ‘dunno’, *qué sé yo* ‘what do I know’, *tú sabes* ‘you know’, etc.), because the existence of such highly fixed formulaic sequences and their idiosyncratic behavior is of interest to hypothesis 2.

To examine the effects of the cognitive constraints, a mixed-effects logistic regression analysis was performed with the *lme4* package (Bates et al. 2016) in R. Since hypothesis 2 claims that the overall entrenchment of tokens in <Verb> or <SPP Verb> will modulate the effects of all other predictors, tokens were grouped together by the conjugated verb forms they instantiate. The tokens were also grouped together by speaker.

5.4 Results

Regarding markedness of coding, Table 7 shows that speakers use <SPP Verb> more often when talking about themselves or their interlocutor and much less frequently when referring to human referents that are not present in the interview context, such as the speaker and his childhood friends in example (12). These results reflect the empathy hierarchy (Speaker > Hearer > Other).

- (12) (LH05M21/12)
 A la playa íbamos también solos.
 to the beach went.1-PL also alone
 ‘To the beach we also went alone.’

Additionally, the data reveal that speakers are more likely to use <SPP Verb> for subjects that are perceived as volitional in the context of the event encoded by the clause and for subjects that refer to specific entities. Both of these properties are illustrated in example (13).

- (13) (LH05M21/131)
 No, yo no me meto allí a casa de nadie.
 NEG I NEG myself put over there in the house of noone
 ‘No, I don’t go over to other people’s houses over there.’

In turn, the results for aspect and kinesis reveal that speakers disfavor <SPP Verb> in clauses with continuous or perfective aspect, as in example (14), or in clauses that refer to energetic events, as in example (15), as Posio (2011) had already observed for first- and second-person singular SPPs.

- (14) (LH01H22/253)
 y eso fue lo que estudié
 and that was what studied.1-SG
 ‘and that was what I studied’

- (15) (LH22M11/861)
 preferiría que llegues
 would prefer.1-SG that arrive.2-SG
 ‘I would prefer that you arrived’

Table 7: Logistic generalized linear mixed-effects model of SPE in Cuban Spanish: numbers, percentages, and coefficients for pronominal subjects (sum contrasts).⁹

	N	%	Coefficient
<i>(Intercept)</i>	2194/7849	27.95	-1.68
<i>Collostruction strength</i>			
Collostruction strength	Numeric	Predictor	0.662
<i>Empathy</i>			
Speaker – Hearer	1618/4542	35.62	0.412
Other	576/3307	17.42	-0.412
<i>Production-to-production priming</i>			
<SPP Verb>	832/2051	40.57	0.339
First/5+ clauses	181/566	31.98	-0.037
<Verb>	1181/5232	22.57	-0.303
<i>Referential continuity and referential distance</i>			
Switch: non-adjacent	1133/3169	35.75	0.333
Switch: new	169/796	21.23	0.266
Switch: adjacent	99/387	25.58	-0.082
Continuity: adjacent	793/3497	22.68	-0.516
<i>Aspect</i>			
Imperfective	1719/5928	29	0.314
Perfective	445/1756	25.34	-0.106
Continuous	30/165	18.18	-0.208
<i>Kinesis</i>			
Non-action	1627/5151	31.59	0.281
Action	555/2565	21.64	-0.281
<i>Volition</i>			
Volitional	1300/4929	26.37	0.166
Non-volitional	894/2920	30.62	-0.166
<i>Animacy of the object</i>			
Inanimate	681/2571	26.49	0.131
Absent	1391/4643	29.96	0.09
Animate	122/635	19.21	-0.221
<i>Referentiality</i>			
Referential	1749/5504	31.78	0.061
Non-referential	445/2345	18.98	-0.061
<i>Comprehension-to-production priming</i>			
<SPP Verb>	75/1136	33.01	0.095
<Verb>	400/1562	25.61	-0.034
First/5+ clauses	1419/5151	27.55	-0.061

(contd.)

<i>Number of words between verb and SPP site</i>			
Number of words between verb and SPP site	Numeric	Predictor	-0.098
Random effects		Variance	Std. Deviation
Verb form		0.844	0.919
Speaker		0.261	0.511
Model summary			
Pseudo-R ²			0.41
AIC _c			7840.2
C index			0.85

In addition, Table 7 reveals that speakers are less likely to use <SPP Verb> when an animate object is present in the clause (see example (16)). In turn, when no object is present in the clause or when the object has inanimate reference, speakers are more inclined to use <SPP Verb>.

- (16) (LH23H21/637)
 que cogieron a mi amigo
 when caught.3-PL my friend
 ‘when they caught my friend’

These data support that features of lower transitivity (the absence of an object or non-individuated/inanimate objects, imperfective aspect, verbs that refer to non-energetic events; Hopper & Thompson 1980) favor <SPP Verb>, whereas features typical of higher transitivity favor <Verb> (Posio 2011). This pattern is highly favorable to the view that speakers prefer <Verb> when relatively more attention is focused on the event, whereas they prefer <SPP Verb> when relatively more attention is focused on the subject, as hypothesis 1 predicts.

Regarding the discourse-oriented prominence of the subject, preliminary analyses revealed that a three-way factorized version of referential distance (referent occurs in adjacent clause, referent occurs in non-adjacent clause, new) provided a much better fit than the continuous variable, which has a highly skewed distribution. However, this factorized version collides with referential continuity, as subjects that form a reference chain with the subject of the previous clause also appear in that clause. Therefore, these two predictors were collapsed into one regressor. As in earlier work on SPE, the results reveal that subjects that imply a switch in reference with regard to the subject of the previous clause are more likely to be expressed with <SPP Verb> than subjects that continue to refer to the same entity. However, this tendency is severely mitigated by referential distance. That is, when there is a switch in reference, speakers slightly favor <Verb> when the switch occurs across adjacent clauses. In contrast, when they switch to a referent that has not appeared in prior discourse or to a referent that was mentioned further away in discourse, they favor <SPP Verb>. These results support that discursively prominent referents are more likely to be encoded with <SPP Verb>, as hypothesis 1 predicts.

⁹ Besides the predictors described here, the model also includes speakers' gender and age. The reader is kindly referred to Claes (under review) where these predictors are given their due consideration. In computing the model the *bobyqa* optimizer for *glmer* was used.

With regard to hypothesis 2, the results for collocation strength reveal that speakers become gradually more likely to use <SPP Verb> as the collocation strength of the verb token rises. In other words, the data support that, when confronted with the choice between <Verb> and <SPP Verb>, speakers draw on their past experience with language and use the construction alternative they have witnessed most consistently with a particular verb form. This is exactly the pattern one would expect in the light of statistical preemption. In addition, Table 7 also shows that when one or more words are inserted between the verb and the SPP site, the use of <SPP Verb> becomes less likely with each element that is inserted. Since the presence of such elements inhibits the use of a prefabricated <SPP Verb> expression, these results add further support to hypothesis 2 and the more general claim that statistical preemption constrains SPP variation.

When it comes to the effects of structural priming, as in earlier research of SPE (e.g., Otheguy & Zentella 2012; Shin 2014), <SPP Verb> is preferred whenever speakers have just used or processed a case of this construction and vice versa. Whenever they have not been exposed to any of the alternatives, they are less likely to use the <SPP Verb> construction, in line with the overall tendency in Spanish to omit subject pronouns.

Finally, the model summary at the bottom of Table 7 shows that the predictors have excellent discriminative ability, suggesting the model represents speakers' choice making very adequately. However, the conditional R-squared value suggests that the model represents only some 40% of the variance. While this is substantially less than the amount of variance that was accounted for in the previous two case studies (0.54–0.62), it doubles the R-squared that is reported in other studies of SPE (e.g., Otheguy & Zentella 2012; Shin 2014).

5.5 Discussion

The data reported in this case study support that speakers are more likely to use <SPP Verb> with subjects that are more likely to attract their attention at the level of the clause or in discourse. The results also suggest that speakers' experience with the token-level preference of verb forms for one construction or the other is a stringent constraint on this alternation. Priming, both from comprehension to production as from production to production, also turned out to be an important determinant of SPE. These results are highly favorable to the view that SPE constitutes a competition between two constructions that may be constrained by markedness of coding, statistical preemption, and structural priming. Let us now turn to some general conclusions that can be derived from the case studies.

6 Conclusions

In this paper, I have proposed that morphosyntactic variation is constrained by three domain-general cognitive constraints that are assumed in Cognitive (Socio)linguistics: markedness of coding, statistical preemption, and structural priming. Then I have explored, analyzing three alternations in two languages, whether this claim is supported by empirical data. The results of these case studies suggest that, when contextual features are defined in such a fashion that they operationalize the cognitive constraints, highly predictive models of the individual alternations can be obtained. These models do not lose their predictive and explanatory accuracy when they are applied to the same phenomenon in different varieties of the same language, to a similar phenomenon in an unrelated language, or to an unrelated phenomenon in the same language.

How does this portray the notion of a Probabilistic Grammar? The case studies suggest that speakers' grammars of linguistic alternations are not merely the result of exposure to/the internationalization of (largely arbitrary) probabilistic patterns in usage. Rather, the consistency in these patterns across speakers, varieties, and languages reflect the joint

action of domain-general cognitive constraints on spreading activation. In other words, if speakers of Spanish use more plural agreement with typical agents, this does not indicate that they have learned from input that this form has the higher probability in that context (e.g., Bresnan 2007), but rather that markedness of coding increases this construction's level of activation for conceptualizations that include this type of nominal referent. Of course, speakers will still have to learn the constructions and their specific semantic details from input. Crucially, the model does not deny the influence of usage and usage patterns either; these have a profound effect on linguistic representation and encoding, particularly through statistical preemption.

In summary, this paper has illustrated how combining the methods that define Probabilistic Grammar with the theoretical tools of Cognitive Sociolinguistics leads to a psychologically plausible answer to a question that has puzzled corpus linguists, cognitive linguists, and variationists for decades: *why does linguistic variation pattern the way it does?* The result is a theoretical model of morphosyntactic variation that generates empirically falsifiable predictions for multiple alternations and even for multiple languages. Beyond morphosyntax, some version of these constraints may also be operative. Geeraerts (2016), for example, makes such a claim for lexical choice. In any case, I hope to have demonstrated how Cognitive Sociolinguistics may assist Probabilistic Grammar in moving from description towards explanation.

Abbreviations

ACC = accusative pronoun, ADVP = adverbial phrase, COMP = complement, IMP-SUBJ = impersonal subject, NEG = sentence negation, NP = noun phrase, OBJ = object, PL = plural, SG = singular, SPE = subject pronoun expression, SPP = subject personal pronoun, SUBJ = subject

Acknowledgements

I dedicate this paper to my niece Roos Claes, who was born when I was revising the final manuscript. I would also like to thank Daniel Ezra Johnson for a fruitful collaboration in developing further some of the ideas that form the core of this paper.

Competing Interests

The author has no competing interests to declare.

References

- Bartón, Kamil. 2015. *MuMIn: Model selection and model averaging based on information criteria (AICc and alike)*. <https://cran.r-project.org/web/packages/MuMIn/index.html>.
- Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2016. *lme4: Linear mixed-effects models using 'Eigen' and S4*. <https://cran.r-project.org/web/packages/lme4/index.html>.
- Bayley, Robert, Kristen Greer & Cory Holland. 2013. Lexical frequency and syntactic variation: A test of a linguistic hypothesis. *University of Pennsylvania Working Papers in Linguistics* 19(2). Art. 4.
- Blas-Arroyo, José-Luis. 1995. A propósito de un caso de convergencia gramatical por causación múltiple en el área de influencia lingüística catalana. *Análisis sociolingüístico. Cuadernos de investigación filológica* 21–22. 175–200. DOI: <https://doi.org/10.18172/cif.2356>
- Blas-Arroyo, José-Luis. 2016. Entre la estabilidad y la hipercorrección en un antiguo 'cambio desde abajo': *Haber* existencial en las comunidades de habla castellanenses. *Lingüística Española Actual* 38(1). 69–108.

- Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 77–96. Berlin/Boston, MA: De Gruyter.
- Bresnan, Joan, Ana Cueni, Tatiana Nikitina & Rolf Harald Baayen. 2007. Predicting the dative alternation. In Gerolf Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
- British National Corpus Consortium. 2007. *The British National Corpus, version 3 (BNC XML Edition)*. <http://www.natcorp.ox.ac.uk/XMLedition/URG/>.
- Burnham, Kenneth & David Anderson. 2002. *Model selection and multimodel inference*. New York, NY: Springer. DOI: <https://doi.org/10.1007/b97636>
- Bybee, Joan. 2001. *Phonology and language use*. Cambridge, MA: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511612886>
- Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge, MA: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511750526>
- Carvalho, Ana-María, Rafael Orozco & Naomi Shin. 2015. Introduction. In Ana-María Carvalho, Rafael Orozco & Naomi Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 13–23. Georgetown, DC: Georgetown University Press. DOI: <https://doi.org/10.13109/wdor.2015.45.1.3>
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Boston, MA: The MIT Press.
- Chomsky, Noam. 1995. *The minimalist program*. Boston, MA: The MIT Press. DOI: <https://doi.org/10.7551/mitpress/9780262527347.001.0001>
- Claes, Jeroen. 2012. *Caribe: A sociolinguistic corpus of Caribbean Spanish (Havana, Santo Domingo, and San Juan)*. 72 speakers, 78 hours of speech, 100 million words. Antwerp: University of Antwerp.
- Claes, Jeroen. 2014a. A Cognitive Construction Grammar approach to the pluralization of presentational *haber* in Puerto Rican Spanish. *Language Variation and Change* 26(2). 219–246. DOI: <https://doi.org/10.1017/S0954394514000052>
- Claes, Jeroen. 2014b. La pluralización de *haber* presentacional y su distribución social en el español de La Habana, Cuba: Un acercamiento desde la gramática de construcciones. *Revista Internacional de Lingüística Iberoamericana* 23. 165–187.
- Claes, Jeroen. 2014c. Sociolingüística comparada y gramática de construcciones: Un acercamiento a la pluralización de *haber* presentacional en las capitales antillanas. *Revista Española de Lingüística Aplicada* 27(2). 338–364. DOI: <https://doi.org/10.1075/resla.27.2.05cla>
- Claes, Jeroen. 2015. Competing constructions: The pluralization of presentational *haber* in Dominican Spanish. *Cognitive Linguistics* 26(1). 1–30. DOI: <https://doi.org/10.1515/cog-2014-0006>
- Claes, Jeroen. 2016. *Cognitive, social, and individual constraints on linguistic variation: A case study of presentational haber pluralization in Caribbean Spanish*. Berlin/Boston, MA: De Gruyter Mouton. DOI: <https://doi.org/10.1515/9783110524154>
- Claes, Jeroen. 2017a. Cognitive and geographic constraints on morphosyntactic variation: The variable agreement of presentational *haber* in Peninsular Spanish. *Belgian Journal of Linguistics*. In press.
- Claes, Jeroen. 2017b. La pluralización de *haber* presentacional en el español peninsular: Datos de Twitter. *Sociolinguistic Studies* 11(1). In press.
- Claes, Jeroen & Daniel Ezra Johnson. Under review. Cognitive Linguistics and the predictability of effects: Agreement in English and Spanish presentational constructions. *Linguistics*.

- Crawford, William. 2005. Verb agreement and disagreement: A corpus investigation of concord variation in existential *there + be* constructions. *Journal of English Linguistics* 33(1). 35–61. DOI: <https://doi.org/10.1177/0075424204274001>
- Croft, William. 2003. *Typology and universals*. Cambridge, MA: Cambridge University Press.
- Croft, William & Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge, MA: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511803864>
- D'aquino-Ruiz, Giovanna. 2008. El cambio lingüístico de *haber* impersonal. *Núcleo* 20(25). 103–124.
- Davies, Mark. 2002-. *Corpus del español. 100 million words (1200s–1900s)*. <http://www.corpusdelespanol.org/>.
- Dell, Gary. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 92(3). 283–321. DOI: <https://doi.org/10.1037/0033-295X.93.3.283>
- Ellis, Nick & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220. DOI: <https://doi.org/10.1075/arcl.7.08ell>
- Flores-Ferrán, Nydia. 2007. A bend in the road: Subject personal pronoun expression in Spanish after 30 years of sociolinguistic research. *Language and Linguistics Compass* 1(6). 624–652. DOI: <https://doi.org/10.1111/j.1749-818X.2007.00031.x>
- Geeraerts, Dirk. 2005. Lectal variation and empirical data in Cognitive Linguistics. In Francisco Ruiz de Mendoza Ibáñez & Sandra Peña Cervel (eds.), *Cognitive Linguistics: Internal dynamics and interdisciplinary interaction*, 163–189. Berlin/Boston, MA: De Gruyter.
- Geeraerts, Dirk. 2006. A rough guide to Cognitive Linguistics. In Dirk Geeraerts (ed.), *Cognitive Linguistics: Basic readings*, 1–28. Berlin/Boston, MA: De Gruyter. DOI: <https://doi.org/10.1515/9783110199901.1>
- Geeraerts, Dirk. 2016. Entrenchment as onomasiological salience. In Hans-Jörg Schmid (ed.), *Entrenchment, memory and automaticity: The psychology of linguistic knowledge and language learning*, 153–174. Berlin & Boston, MA: De Gruyter. DOI: <https://doi.org/10.1515/9783110199901.1>
- Geeraerts, Dirk & Gitte Kristiansen. 2015. Variationist linguistics. In Dagmar Divjak & Ewa Dabrowska (eds.), *Handbook of Cognitive Linguistics*, 366–389. Berlin & Boston, MA: De Gruyter. DOI: <https://doi.org/10.1515/9783110292022-018>
- Goldberg, Adele. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago, IL: Chicago University Press.
- Goldberg, Adele. 2005. Constructions, lexical semantics, and the correspondence principle: Accounting for generalizations and subregularities in the realization of arguments. In Nomi Erteschik-Shir & Tova Rapoport (eds.), *The syntax of aspect: Deriving thematic and aspectual interpretation*, 215–236. Oxford: Oxford University Press.
- Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- Goldberg, Adele. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics* 22(1). 131–153. DOI: <https://doi.org/10.1515/cogl.2011.006>
- Hay, Jennifer & Daniel Schreier. 2004. Reversing the trajectory of language change: Subject-verb agreement with *be* in New Zealand English. *Language Variation and Change* 16(2). 209–235. DOI: <https://doi.org/10.1017/s0954394504163047>
- Henry, Allison. 2002. Variation and syntactic theory. In J. K. Chambers, Peter Trudgill & Natalie Chilling-Estes (eds.), *The handbook of language variation and change*, 267–282. Oxford: Blackwell. DOI: <https://doi.org/10.1002/9780470756591.ch10>

- Hopper, Paul & Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299. DOI: <https://doi.org/10.1353/lan.1980.0017>
- Hosmer, David & Stanley Lemeshow. 2000. *Applied logistic regression*. Oxford: Wiley. DOI: <https://doi.org/10.1002/0471722146>
- Hudson, Richard. 2010. *An introduction to Word Grammar*. Cambridge, MA: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511781964>
- Keenan, Edward. 1976. Towards a universal definition of subject. In Charles Li (ed.), *Subject and topic*, 305–333. New York, NY: Academic Press.
- Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam & Philadelphia, PA: John Benjamins. DOI: <https://doi.org/10.1075/cilt.24.06lab>
- Lakoff, George. 1977. Linguistic Gestalts. In Woodford Beach, Samuel Fox & Shulamith Philosph (eds.), *Papers from the Thirteenth Regional Meeting of the Chicago Linguistics Society, April 14–16, 1977*, 236–287. Chicago, IL: Chicago Linguistics Society.
- Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: Chicago University Press. DOI: <https://doi.org/10.7208/chicago/9780226471013.001.0001>
- Lakoff, George. 1990. The invariance hypothesis: Is abstract reason based on image schemas? *Cognitive Linguistics* 1(1). 39–74. DOI: <https://doi.org/10.1515/cogl.1990.1.1.39>
- Langacker, Ronald. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Langacker, Ronald. 1991. *Foundations of Cognitive Grammar, Vol. 2: Descriptive application*. Stanford, CA: Stanford University Press.
- Langacker, Ronald. 2007. Cognitive Grammar. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford handbook of Cognitive Linguistics*, 421–462. Oxford: Oxford University Press.
- Langacker, Ronald. 2008. *Cognitive Grammar: A basic introduction*. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780195331967.001.0001>
- Langacker, Ronald. 2010. Cognitive Grammar. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 87–110. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780199738632.013.0017>
- Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam & Philadelphia, PA: John Benjamins. DOI: <https://doi.org/10.1075/z.195>
- Matos-Amaral, Patricia & Scott Schwenter. 2005. Contrast and the (non-)occurrence of subject pronouns. In David Eddington (ed.), *Selected proceedings of the 7th Hispanic linguistics symposium*, 116–127. Somerville, MA: Cascadilla.
- Myachykov, Andriy & Russel Tomlin. 2015. Attention and salience. In Ewa Dabrowska & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics*, 31–52. Berlin & Boston, MA: De Gruyter Mouton. DOI: <https://doi.org/10.1515/9783110292022-003>
- Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4. 133–142. DOI: <https://doi.org/10.1111/j.2041-210x.2012.00261.x>
- Orozco, Rafael. 2015. Pronominal variation in Colombian Costeño Spanish. In Ana-Maria Carvalho, Rafael Orozco & Naomi Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 17–38. Georgetown, DC: Georgetown University Press.
- Otheguy, Ricardo & Ana-Celia Zentella. 2012. *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199737406.001.0001>
- Otheguy, Ricardo, Ana-Celia Zentella & David Livert. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* 83(4). 771–802. DOI: <https://doi.org/10.1353/lan.2008.0019>


- Pato, Enrique. 2016. La pluralización de *haber* en español peninsular. In Carlota de Benito Moreno & Álvaro Octavio de Toledo (eds.), *En torno a haber: Construcciones, usos y variación desde el latín hasta la actualidad*, 357–392. Berlin: Peter Lang.
- Pérez Martín, Ana-María. 2004. Pluralización del verbo *haber* en el habla de la isla de El Hierro: datos parciales. *Interlingüística* 15(2). 1125–1130.
- Pickering, Martin & Victor Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin* 134(3). 427–459. DOI: <https://doi.org/10.1037/0033-2909.134.3.427>
- Posio, Pekka. 2011. Spanish subject pronoun usage and verb semantics revisited: First and second person singular subject pronouns and focusing of attention in spoken Peninsular Spanish. *Journal of Pragmatics* 43. 777–798. DOI: <https://doi.org/10.1016/j.pragma.2010.10.012>
- Posio, Pekka. 2015. Subject pronoun usage in formulaic sequences: Evidence from Peninsular Spanish. In Ana-María Carvalho, Rafael Orozco & Naomi Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 59–79. Georgetown, DC: Georgetown University Press.
- Prince, Ellen. 1992. The ZPG letter: Subjects, definiteness, and information-status. In William Mann & Sandra Thompson (eds.), *Discourse description: Diverse linguistic analyses of a fund-raising text*, 295–326. Amsterdam & Philadelphia, PA: John Benjamins. DOI: <https://doi.org/10.1075/pbns.16.12pri>
- Pütz, Martin, Justyna Robinson & Monika Reif. 2012. The emergence of Cognitive Sociolinguistics. *Review of Cognitive Linguistics* 10(2). 241–26. DOI: <https://doi.org/10.1075/rcl.10.2.01int>
- R Core Team. 2016. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robenalt, Clarice & Adele Goldberg. 2015. Judgment evidence for statistical preemption: It is relatively better to *vanish* than to *disappear* a rabbit, but a lifeguard can equally well *backstroke* or *swim* children to shore. *Cognitive Linguistics* 26(3). 467–503. DOI: <https://doi.org/10.1515/cog-2015-0004>
- Shin, Naomi. 2014. Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change* 26(2). 303–330. DOI: <https://doi.org/10.1017/S095439451400012X>
- Stefanowitsch, Anatol & Stephan Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43. DOI: <https://doi.org/10.1515/cllt.2005.1.1.1>
- Szmrecsanyi, Benedikt. 2008. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin & Boston, MA: De Gruyter. DOI: <https://doi.org/10.1515/9783110197808>
- Szmrecsanyi, Benedikt. 2013. Diachronic Probabilistic Grammar. *English Language and Linguistics* 19(3). 41–68. DOI: <https://doi.org/10.17960/ell.2013.19.3.002>
- Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert & Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28(1). 1–29. DOI: <https://doi.org/10.1017/S0954394515000198>
- Tagliamonte, Sali. 1998. *Was/were* variation across the generations: View from the city of York. *Language Variation and Change* 10(2). 153–191. DOI: <https://doi.org/10.1017/S0954394500001277>
- Tagliamonte, Sali. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Oxford: Wiley-Blackwell.
- Tagliamonte, Sali & Rolf Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178. DOI: <https://doi.org/10.1017/S0954394512000129>

- Toutanova, Kristin, Christopher Manning, Kan Klein & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, 252–259. North American Chapter of the Association for Computational Linguistics. DOI: <https://doi.org/10.3115/1073445.1073478>
- Travis, Catherine. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19. 101–135. DOI: <https://doi.org/10.1017/S0954394507070081>
- Travis, Catherine & Rena Torres-Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical, and constructional factors in variation. *Cognitive Linguistics* 23(4). 711–748. DOI: <https://doi.org/10.1515/cog-2012-0022>
- Vaquero, María. 1996. Antillas. In Manuel Alvar-López (ed.), *Manual de dialectología hispánica: El español de América*, 51–67. Barcelona: Ariel.
- Walker, James. 2007. “There’s bears back there” Plural existentials and vernacular universals in (Quebec) English. *English World Wide* 28(2). 147–166. DOI: <https://doi.org/10.1075/eww.28.2.03wal>

How to cite this article: Claes, Jeroen. 2017. Probabilistic Grammar: The view from Cognitive Sociolinguistics. *Glossa: a journal of general linguistics* 2(1): 62.1–30, DOI: <https://doi.org/10.5334/gjgl.298>

Published: 29 June 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 