Cognitive and geographic constraints on morphosyntactic variation

The variable agreement of presentational *haber* in Peninsular Spanish

Jeroen Claes
KU Leuven

In this paper, I examine whether the variation patterns of *haber* pluralization (e.g., *hubo/hubieron fiestas* 'there was/were parties') in Peninsular Spanish corroborate the hypothesis elaborated in earlier work that the phenomenon constitutes a competition between two variants of the presentational construction with *haber* that is constrained by domain-general cognitive constraints on spreading activation. In addition, this paper examines whether *haber* pluralization is incrementing in frequency in particular Peninsular regions and whether or not the phenomenon is spreading geographically. To meet these objectives, I analyze a dataset of more than 7,500 cases of *haber* + plural NP, which were culled from two publicly available data sources: the *Corpus Oral y Sonoro del Español Rural* (which represents only rural speakers born before the 1940s; Fernández-Ordóñez 2005-) and *Twitter* (which represents mainly young and middle-aged speakers). The results of a mixed-effects logistic regression analysis that tests the effects of tense, the absence/presence of negation, typical action-chain position of the noun, the regional origin of the examples, and the data sources support the competition hypothesis. This model also supports that pluralized *haber* is spreading westward from its epicenters (Valencia, Barcelona, and Murcia), while also incrementing in frequency in northern, eastern and southern Spain. However, its frequency appears to be declining in central Spain. A geographically more detailed, but similar picture is obtained with three generalized additive mixed models that test the effects of geography on the total dataset as well as on each of the two subcorpora.

1. Introduction

In Spanish, the most widely used existential/presentational construction is formed with the verb *haber*, which has a similar function to its English analog *there to*

https://doi.org/10.1075/bjl.00002.cla
Belgian Journal of Linguistics Volume 31 (2017), pp. 30–55. issn 0774-5141 | e-issn 1569-9676

be. However, unlike English *there is/there are*, the presentational *haber* construction does not display verb agreement with plural noun phrases (NPs) in normative usage. Since the NP pronominalizes as an accusative pronoun, we may consider that the absence of verb agreement indicates that it functions as a direct object (Gili-Gaya 1980, 78; RAE and ASALE 2009, §41.6.b). Yet, most (if not all) informal varieties of Spanish display some degree of variable agreement with plural noun phrases, as is demonstrated in example (1), where the presence of *-n* marks third-person plural on the verb.

(1) No es que no *hayan* argumentos, es que a los fachas se les bloquea el cerebro cuando se les incita a entrar en razón.
(Twitter, Sant Boi de Llobregat, Barcelona Province, Catalonia)
'It's not that *there aren't* arguments, it's that fascists' brains freeze when they're invited to come to reason.'

For the Spanish Peninsula, the data available in the literature suggest that *haber* pluralization only appears with some frequency in Catalonia, in the Valencian Community, in eastern Aragon, in the east of Castile-La Mancha, Murcia, and in eastern Andalusia (Gili-Gaya 1980, 78; Llorente 1980, 30; RAE and ASALE 2009, §41.6b), for which the isogloss of the phenomenon largely coincides with the (historic) Catalan language area. Therefore, the (former) contact situation with this language, which displays a parallel alternation, may have influenced the emergence of *haber* pluralization in the Spanish varieties of these regions (Claes 2014d, Chapter 7; Blas-Arroyo 1995).

However, variable *haber* agreement does not appear to be limited to eastern Spain. Rather, Lorenzo (1971, 256) already noted in the early 1970s that the use of agreeing *haber* was spreading westward from the bilingual eastern communities. Still, DeMello (1991) does not document the phenomenon in the educated speech of Seville or Madrid. Similarly, in a study of grammatical agreement phenomena in the spoken language of the latter city, Quilis (1983, 94) found only two agreeing examples on a total of more than a thousand cases of presentational *haber* followed by a plural noun. Similar conclusions were reached by Blas-Arroyo (2016) and Paredes-García (2016), who report to have found a small number of agreeing examples in spoken-language corpora from, respectively, Alcalá de Henares, Granada, Malaga and Madrid. But, for rural Spain, Pato (2016) has shown that agreeing cases of *haber* can be documented for each province of Spain that is included in the *Corpus Oral y Sonoro del Español Rural* (COSER). Similarly, Claes (2017), who studies a corpus of Peninsular tweets, finds that *haber* pluralization occurs with a relative frequency of 11% (N = 5,500) in Peninsular Spanish. These data call for a reassessment of the geographical spread of the phenomenon.

Regarding the linguistic factors that impact on *haber* pluralization, previous variationist explorations in Canarian, Latin American, and Peninsular Spanish reveal a recurring pattern. For example, many studies find that agreeing *haber* is favored by human reference, the absence of negation, the imperfect indicative, compound tenses and aspectual/modal auxiliaries (Bentivoglio and Sedano 2011; D'Aquino Ruiz 2004; Diaz-Campos 2003). For the Canary Islands, Samper-Padilla and Hernández-Cabrera (2012, 749–750) equally document more pluralized cases in the imperfect tense (24.1% vs. 20.5%). For Castellón de la Plana (Valencian Community), Blas-Arroyo (2016) notes that the pluralized variant is favored by human reference and is more visible in the imperfect indicative. Similarly, of the twelve pluralized cases documented in Madrid by Paredes-García (2016, 224), ten correspond to the imperfect indicative. As I have argued elsewhere (Claes 2014a, 2014b , 2014c, 2014d, 2016), these quantitative similarities across different varieties of Spanish are too striking to be coincidental. Rather, relying on Cognitive Construction Grammar (Goldberg 1995, 2006) I have argued in previous work that *haber* pluralization involves a competition between two variants of the presentational construction, which is conditioned by three general cognitive constraints proposed by Cognitive (Socio)linguistics: markedness of coding, statistical preemption, and structural priming. In Section 3, we will consider this hypothesis with some more detail.

In sum, while the variation patterns of *haber* pluralization in American, Canarian, and eastern Peninsular Spanish are quite well understood, its geographic distribution in Peninsular Spanish and its potential geographic spread over time remain subject to debate. It also remains to be explored whether the analysis that was proposed for Caribbean Spanish applies to the Spanish Peninsula as well. Therefore, this paper follows up on Claes (2017), where I investigate the geographic distribution of *haber* pluralization in Peninsular Spanish using a corpus of tweets. Particularly, in this paper, I will analyze an extended dataset of geographically annotated tokens of *haber* + plural NP, which was culled from two publicly available sources: Twitter (*N* = 5,500; analyzed in Claes 2017) and Fernández-Ordoñez's (2005-) *Corpus Oral y Sonoro del Español Rural* (*N* = 2,031), a rural dialect corpus of Spanish that only includes older speakers. With these data, I intend to address the following two research questions:

- Do the results support portraying *haber* pluralization as a competition between two variants of the presentational construction with *haber* that is constrained by domain-general cognitive constraints on spreading activation, as was proposed for Caribbean Spanish?

- What is the geographic distribution of *haber* pluralization in Peninsular Spanish? Does *haber* pluralization appear to be incrementing in frequency while also spreading geographically in apparent time?

In exploring these research questions in the light of interviews recorded with older, nonmobile rural speakers – the preferred data source of traditional dialectology – and samples drawn from Twitter, this paper will also investigate how data generated by new social media may open a new and detailed window on the geographic distribution of morphosyntactic variation and its evolution over time.

The remainder of this paper is structured as follows. In Section 2, I discuss the methods of this investigation. In Section 3, I introduce the theoretical framework. Section 4 presents the hypotheses, and the way these hypotheses were operationalized. Section 5 is concerned with the results, and Section 6 offers some concluding remarks.

2. Methods

2.1 Data

As I announced in the introduction, the data were culled from two publicly available sources that record the geographic origin of examples: the *Corpus Oral y Sonoro del Español Rural* (COSER, henceforth; Fernández-Ordóñez 2005-) and Twitter. The first data source constitutes a large-scale, ongoing project that aims to develop a publicly available corpus of the language of older, nonmobile rural speakers of Peninsular Spanish. To date, the corpus includes interviews with 2,248 speakers, distributed across 1,124 rural communities. Of these, 147 interviews representing 141 localities and 183 hours of speech can be searched through the corpus website (http://www.corpusrural.es/). The mean age of the informants at the moment of the interview (the first of which were recorded in 1988) was 71.55 years, which implies that all speakers included in the corpus were born before the 1940s.

The COSER data were collected by performing searches with regular expressions using the search engine on the corpus website. The searches focused on *haber* in non-present tenses as well as all forms of the following modal verb constructions, where agreement is visible on the auxiliary rather than *haber* itself: *acabar de haber* 'there has just been', *deber haber* 'there has to be', *deber de haber* 'there has to be', *ir a haber* 'there is going to be', *poder haber* 'there can be', *seguir habiendo* 'there continues to be'. The exclusion of the present tense was motivated by the fact that this tense has been shown to display no variation in Peninsular Spanish (e.g., Blas-Arroyo 2016; Paredes-García 2016; Gómez-Molina 2013).

For 106 out of the 141 localities, at least one case of presentational *haber* + plural NP could be collected. Since the transcripts reflect the dialectal pronunciation as closely as possible, there is a substantial amount of variation. This impeded the automatic extraction of the tokens of presentational *haber* + plural NP from the web interface. Therefore, I selected all applicable tokens manually and I handcoded them for the following variables: absence/presence of negation, typical action-chain position, verb tense, noun that appears with *haber* (see Section 4), and the locality. Subsequently, I used the *geocode* function of the R (R Core Team 2016) package *ggmaps* (Kahle and Wickman 2016) to annotate the examples with the longitude, latitude, province, and autonomous community of the locality from which the token was drawn.

These data were supplemented with the collection of Peninsular tweets analyzed in Claes (2017). The combination of these two data types was motivated by both practical concerns and the research questions. Regarding the former, the type of analysis that will performed in this paper requires a denser grid of localities than the 141 sites for which the COSER provides data. Regarding the latter, confronting the speech patterns of older speakers with data from Twitter – which represent mostly young to middle-aged speakers – we may gain more insight into whether or not *haber* pluralization is incrementing and spreading geographically over (apparent) time (see Labov 1994: Chapter 1 for discussion on 'real' vs. 'apparent' time).

To collect the Twitter data, I used the *TwitteR* (Gentry 2016) package for R to perform batches of automated searches with the *Search API,* which targeted the same forms as those that were considered for the COSER. Each batch combined a global search in a radius of 580 kilometers around Madrid with more restricted searches that targeted a radius of 200 kilometers around each of the capitals of the autonomous communities (excluding non-contiguous areas, i.e., the Canary Islands, the Balearic Islands, Ceuta, and Melilla). The searches yielded an initial dataset of some 50,000 cases of *haber,* which was pruned in the following way. First, all retweets were filtered out. Then, all duplicated tweet ID's and duplicated tweet texts were removed. The remaining tweets were part-of-speech tagged with the *Stanford POS Tagger* (Toutanova et al. 2003), to allow for their automatic filtering and annotation. After tagging, I applied a regular expression to select tokens preceded by a plural pronoun (*e.g., Las había* 'them there were'), the relative *que* and a plural noun (*Las cosas que había* 'the things that there were') and those followed by a plural noun (e.g., *había cosas* 'there were things'), an adjective and a plural noun (*e.g., había grandes cosas* 'there were big things'), a determiner and a plural noun (e.g., *había varias cosas* 'there were various things'), or the adverb *más* and a plural noun (e.g., *había más cosas* 'there were more things'). All remaining tokens that did not match any of these types were discarded.

Because the GPS data that are offered by Twitter are unreliable when the goal is to establish speaker's geographic origin,[1] tokens were located geographically according to the locality users specify as their home location in their user profile. To this end, the location names were transformed to lowercase and variant location names (e.g., *xixón, gijón, gijon*) were recoded to one variant each (*gijon*, in this case). Afterwards, I used the *geocode* function included in the R package *ggmaps* to standardize the location field further and to code the data for the longitude, latitude, province, and autonomous community corresponding with the locality. Localities for which the geocoder could not return results were coded manually; Tweets by users who did not specify a valid location were excluded. Subsequently, the remaining tokens were coded automatically for the same variables as those that were considered for the COSER data. The resulting corpus represents 4,809 unique Twitter users spread over 450 localities. After filtering and coding, the Twitter data were merged with the COSER data and a variable was added to record the origin of the examples.

2.2 Statistical analysis: Mixed-effects logistic regression and generalized additive mixed modeling

With these data, a mixed-effects logistic regression was performed with the *lme4* package (Bates et al. 2016) in R, in which the localities and the nouns that appear with *haber* were included as random effects. To establish the regression model, I started out with a full model including the random effects, the fixed effects (which will be described in full in the following section), as well as all theoretically relevant interactions between them. The full model also included by-locality random slopes for the linguistic predictors; by-noun random slopes failed to converge.

To establish a parsimonious model, I first removed the random slopes one by one, then I removed interactions one by one, after which I started removing the fixed effects one at a time. On each iteration, I compared the Second-Order Akaike Information Criterion of the model (*AICc* in the *MuMin* package for R; Bartón 2016) with that of the full model or a previous more parsimonious model. Following Burnham and Anderson (2002, 70), single-term deletions which implied a reduction of the AICc statistic by two units or more were taken to suggest that there was substantially more evidence in favor of the simpler model, for which the random slope, interaction, or fixed effect was dropped. The model fitting procedure eliminated the by-locality random slope for typical action chain position. Only the interactions between tense and typical action-chain position

[1] For instance, if a user from Madrid happens to be in Sevilla when he or she publishes a tweet, this tweet will become geocoded as representing Sevilla, rather than Madrid.

and that between region and corpus were withheld. To guard against overfitting, bootstrap confidence intervals were computed for the final model (using the *confint* function of the *lme4* package). I also evaluated overdispersion and multicollinearity, which were not issues (summed squares of Pearson residuals < residual degrees of freedom and Variance Inflation Factors < 5; Speelman 2014).

To explore the geographical distribution of *haber* pluralization in Peninsular Spanish, this model was complemented with a generalized additive mixed model (the *bam* function of the *mgcv* package for R; Wood 2016). Generalized additive mixed models are an extension of generalized linear models (e.g., the type implemented in *lme4*). Like generalized linear models, generalized additive models are based on the assumption that each change in a predictor value corresponds with a linear, constant increase or decrease of the log-odds of obtaining a particular value for the dependent variable. However, unlike generalized linear models, generalized additive models allow relaxing this assumption for certain predictors, by estimating their effects not with the linear function, but rather with an unspecified smoothing function. For studies focusing on geographically conditioned variation (e.g., Wieling et al. 2014), this has the advantage that geography can be modeled as a nonlinear continuum, rather than having to divide this continuum into more or less arbitrary discrete entities, such as e.g., provinces or states. As such, with GAMs, the latitude and longitude of research sites can be incorporated directly into the model. This enables us to incorporate geography into our models of linguistic variation, while also controlling for other fixed or random effects. The result of such a model are parametric estimates of the effects of the variables that are evaluated using a linear model, as well as an estimate of the effect of the smoothed predictor (in our case, longitude/latitude pairs) on the outcome. The smoothed terms may be plotted on a map, which offers a fine-grained projection of the relative likelihood of obtaining a particular variant in a particular region in space. For the present paper, I specified generalized additive models that included the same predictors as the mixed-effects model, but the random effect for locality was substituted by a non-parametric smoothing term for the latitude/longitude combination that corresponds to the localities. Let us now consider the theoretical framework of this study and the cognitive constraints on language production that can be derived from it.

3. Cognitive Construction Grammar and cognitive constraints on *haber* pluralization

Following connectionist models in psycholinguistics (e.g., Dell 1986), Cognitive Construction Grammar – as Cognitive Linguistics generally – 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.

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 activation probability 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). 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).

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'. However, preliminary analyses revealed that very few cases occur in the vicinity of a previous token of presentational *haber*+plural NP, for which this cognitive constraint could not be taken into account. Instead, the discussion will focus on the other two.

4. Cognitive constraints on presentational *haber* pluralization

Adopting Cognitive Construction Grammar (Goldberg 1995, 2006) and the above model of cognitive constraints on language variation, I have argued elsewhere (Claes 2014a, 2014b, 2014c, 2015, 2016) that we can conceptualize *haber* pluralization as a competition between two constructions: <**AdvP** *haber* **Subj**> (agreement with a plural NP subject) and <**AdvP** *haber* **Obj**> (non-agreement with a plural NP object). The two alternatives can be claimed to be nearly synonymous. However, 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), it is expected that the variant that has a subject will attribute more semantic prominence to the NP argument than the latter.[2]

Assuming that the variants of the presentational *haber* construction 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 distribution of pluralized and singular *haber* across linguistic contexts.

Markedness of coding

Particularly, since subjecthood is "symptomatic of some special cognitive salience that makes it particularly accessible" (Langacker 1991, 306), markedness of coding leads to the following prediction.

> Cognitively more prominent entities will be encoded more frequently as subject, triggering agreement. (Hypothesis 1, Markedness of coding)

Of course, this prediction remains relatively vacuous unless we operationalize the notion of cognitive prominence in some empirically testable fashion. In this regard, research in Cognitive Linguistics suggests that cognitively prominent participants are those on which the speaker has his or her attention focused and

[2] A second important semantic consideration is the social and stylistic meaning of the variants. This aspect of the variation does not fall within the scope of the present article, but see Claes (2014a, b, c, d, 2015, 2016) for the social groups associated with *haber* pluralization in Caribbean Spanish.

that agents attract more attention than any other semantic role (Myachykov and Tomlin 2015). Therefore, semantic role would be a good candidate to operationalize markedness of coding, even more so because agenthood correlates rather closely with subjecthood (cf. Du Bois 2003, 20–21; Langacker 1991, Chapter 7; Myachykov and Tomlin 2015).

However, the NP of existential expressions cannot be agentive, as the construction presents it as merely being present in a static situation. Still, as I argued in earlier work (e.g., Claes 2014a), it is inarguably the case that some nominal 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*, entities like *lumberjack* may nevertheless be perceived as more potential agents – and as thus as more conceptually prominent – than entities like *tree*.

In Cognitive Linguistics, the semantic roles 'agent' and 'patient' are defined in relation to a conception called the 'action-chain model': archetypically, the head of the chain "volitionally initiates physical activity, resulting, through physical contact, in the transfer of energy to an external object" (Langacker 1991, 285), causing an internal change of state of that entity, the tail of the chain. Agents, then, are defined as 'action-chain heads'. Therefore, in order to test the first hypothesis, each NP was coded for the typical action-chain position of the referent of its head noun, relying on the question in (2).

(2) 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/potential agent
No: Other

Polarity is another measure that can serve to establish the relative prominence of the referent of the NP of existential expressions. Following Prince (1992, 299), the NP of an affirmative existential clause, regardless of its formal definiteness, must be interpreted as referring to a specific instance or token that is unknown to the hearer. Under negative polarity, in contrast, the reference of the NP becomes suspended (Keenan 1976, 318), such that it becomes "identifiable only as a type, not as a specific instance or token" (Croft 2003, 132), and therefore less likely to attract the speaker's attention (Langacker 1991, 308). Therefore, markedness of coding predicts lower rates of pluralized *haber* in clauses that contain negation.

Statistical preemption

In Spanish, not all tensed forms of *haber* occur with equal frequency in the presentational construction compared to their uses in other constructions. In addition, the results of diachronic investigations give reason to believe that the agreeing construction constitutes a posterior evolution of the non-agreeing construction (e.g., Claes 2014d, Chapter 7). As a result, different forms of *haber* would have been entrenched to varying degrees in the non-agreeing Spanish <**AdvP** *haber* **Obj**> construction before the pluralized construction came into usage. Therefore, just as markedness of coding can be translated into features that can be annotated for as typical action-chain position and polarity, statistical preemption can be operationalized as tense-based groupings.

The prediction that follows from statistical preemption is that the stronger a particular tensed form of *haber* is entrenched in the non-agreeing 'older' variant of the existential construction, the less that verb tense will favor the 'newer' agreeing construction. Of course, since statistical preemption includes the provision that forms need to be near-synonyms, this effect will only hold when both the entrenched instance and a novel expression based on the 'new' construction could encode the conceptualization equally well (Hypothesis 2, Statistical preemption).

This raises the question as to how we can measure how strongly each particular verb form of *haber* is entrenched in the 'old' existential constructions. The Cognitive Linguistics literature offers various suggestions for measures of the association between words and construction slots (see e.g., Schmid and Küchenhoff 2013 for a critical overview). Most of these methods depend on a two-by-two contingency table such as Table 1.

In recent work, ΔP has proven to be a viable way to establish how frequently a particular lexical item occurs in a specific construction as opposed to its occurrences in other constructions (e.g., Ellis and Ferreira-Junior 2009; Schmid and Küchenhoff 2013). ΔP is a unidirectional measure that expresses the probability of observing a construction (Cx) in the presence of a word (W), minus the probability of observing the construction in the absence of the word. With a two-by-two table like Table 1, ΔP can be obtained with the formula in (3).

Table 1. Collocations table

Cell A	*Cell C*
Frequency of word *W* in construction *Cx* e.g., Frequency of <**AdvP** *hubo* **Obj**>	Frequency of words other than *W* in construction *Cx* e.g, Frequency of <**Adv** *haber* **Obj**> with forms other than *hubo*
Cell B	*Cell D*
Frequency of word *W* in constructions other than *Cx* e.g., Frequency of non-presentational cases of *hubo*	Frequency of words other than *W* in constructions other than *Cx* e.g., Frequency of non-presentational third-person singular forms of *haber* other than *hubo*

(3) ΔP = (Cell A/(Cell A + Cell B)) – (Cell C/(Cell C + Cell D))

The higher the resulting ΔP, the deeper the word is entrenched in that particular construction.

To ensure that the ΔP measures represent speakers' earlier experience with language, I calculated delta-p scores for occurrence rates of *haber* in the twentieth-century section of *Corpus del español* (Davies 2002-), which is a large, balanced corpus of Spanish that includes multiple genres of both spoken and written Spanish. Table 2 provides an overview of the frequency readings and ΔP measures that were obtained for the different tense forms of *haber*. The table shows that two large groups can be distinguished: the present tense *hay* (ΔP 0.469) and the preterit tense *hubo* (ΔP 0.072) form one group, which occurs relatively less frequently outside of existential expressions and is more than twice as deeply entrenched in <**AdvP** *haber* **Obj**> than other forms of the verb. The other group reunites all other forms, which are either not at all entrenched in <**AdvP** *haber* **Obj**> or occur too infrequently to assume that they are stored as entrenched instances of any construction (in the case of future *habrá*, conditional *habría*, imperfective subjunctive *hubiera*, and present perfect *ha habido*; see the columns *Cell A* and *Cell B* in Table 2). This result supports operationalizing the abstract hypothesis 2 as follows for Spanish: the present and preterit tense will disfavor <**AdvP** *haber* **Subj**>, provided that the conceptualization can be expressed with the entrenched instances <**AdvP** *hay* **Obj**> or <**AdvP** *hubo* **Obj**> (i.e., provided that coding the conceptual import does not call for aspectual or modal auxiliaries). However, since *hay* does not display any variation in Peninsular Spanish, the data were coded as: *hubo/hubieron* vs. all others.

Table 2. Frequency counts and ΔP for different third-person singular forms of *haber* in the twentieth-century section of *Corpus del español*

	Cell A	*Cell B*	*Cell C*	*Cell D*	*ΔP*
Había	4559	21810	23154	14384	−0.438
Hubiera	329	3418	27384	32776	−0.083
Habría	514	1952	27199	34242	−0.035
Haya	1072	1965	26641	34229	−0.016
Habrá	971	1025	26742	35169	0.007
ha habido	685	32	27028	36162	0.024
Hubo	2334	450	25379	35744	0.072
Hay	17249	5542	10464	30652	0.469

5. Results

5.1 General distribution

With the methods that were described in Section 2, I collected a corpus of 7,531 instances of presentational *haber* followed by a plural NP, which is the largest dataset that has been analyzed to date for this alternation. Of these tokens, 5,500 were culled from Twitter; the remaining 2,031 tokens were drawn from the COSER. This corpus represents a dense grid of 550 localities, distributed across the Peninsula as is shown in Figure 1.

When it comes to the overall distribution of singular and pluralized *haber,* the top row of Table 3 shows that, even though *haber* pluralization occurs across the entire Spanish peninsula, it is far less common in Peninsular than in Latin American Spanish, where pluralized *haber* typically represents some 40–80% of the cases, depending on the variety, the number of tokens, and the sociolect that are considered (e.g., Claes, 2016: Chapter 2; D'Aquino-Ruiz 2008; Bentivoglio and Sedano 2011; Lastra and Martín-Butragueño 2016). As a matter of fact, the fre-

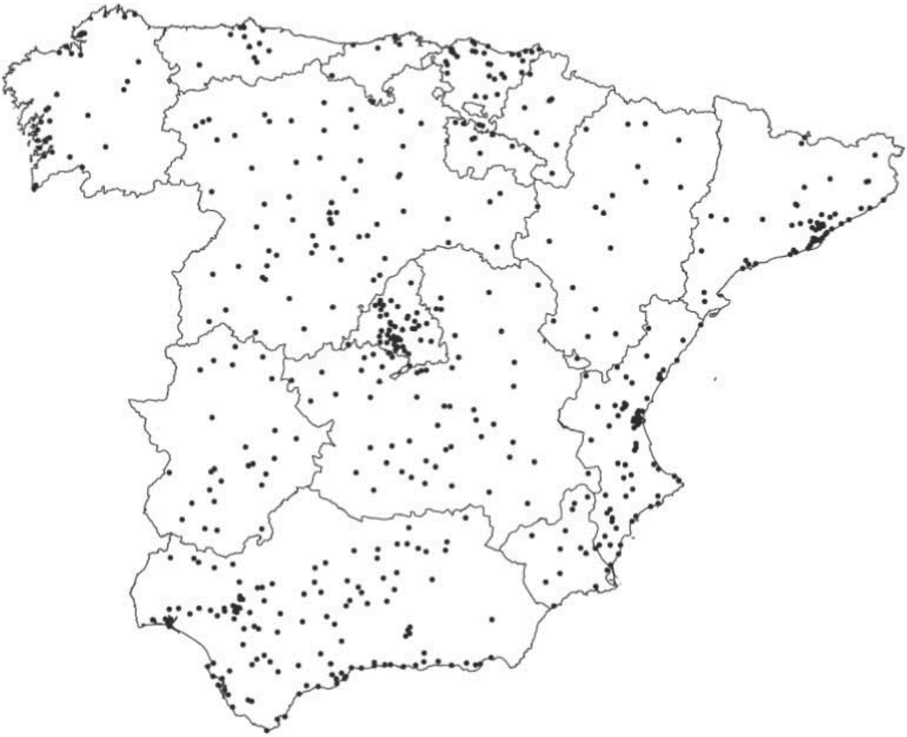

Figure 1. Localities included in the present study

quency of the pluralized variant does not rise above the 10% mark in Peninsular Spanish.

5.2 Cognitive constraints

Turning now to the variables that model markedness of coding (i.e., the typical action-chain position of the referent of the noun and the absence/presence of negation), the regression estimates in Table 3 support that speakers of Peninsular Spanish are somewhat more likely to use pluralized *haber* when the noun refers to an entity that can easily be imagined as a starting point of a series of events (i.e., with typical action-chain heads; 0.059 log-odds). For other types of nouns, speakers are less likely to do so (−0.059 log-odds). In addition, the interaction between this variable and the verb tense shows that the small effect size (0.118 log-odds) that is obtained for this variable may be explained by the fact that speakers appear to be more sensitive to differences in typical action-chain position for the preterit tense (0.133 log-odds); for the other verb tenses, the effect dissipates (−0.133 log-odds). As to the absence/presence of negation, Table 3 shows that speakers of Peninsular Spanish are more likely to use pluralized *haber* when negation is absent (0.188 log-odds). When negation is present, they prefer the singular variant (−0.188 log-odds).

When it comes to statistical preemption, the results for the verb tense show that speakers are far more likely to use pluralized *haber* for tenses other than the synthetic preterit (0.975 log-odds). Indeed, the frequency of the pluralized variant is consistently higher for non-preterit forms of *haber*, as is shown in Table 4.

The effects of typical action-chain position, negation, and the verb tense appear to be stable across the two corpora and the regional varieties of peninsular Spanish, because interaction terms between the linguistic variables and the other two predictors did not lower the AICc statistic. This finding, coupled with the fact that the results pattern as predicted by hypotheses 1 and 2 provides support for the claim that *haber* pluralization constitutes a competition between two variants of the presentational construction with *haber* that is constrained by markedness of coding and statistical preemption. The fact that the same predictors have been shown to constrain *haber* pluralization in Caribbean Spanish (e.g., Claes 2014a, b, c, d, 2015, 2016) with the same direction of effects adds even more support to this interpretation. Let us turn now to the geographic distribution of *haber* pluralization and its potential incrimination and diffusion over (apparent) time.

Table 3. Generalized linear logistic mixed model of *haber* pluralization in Peninsular Spanish (sum contrasts)

Variable	*N*	*%*	*Estimate*
(Intercept)	747/7531	9.91	−4.418
Negation			
Absent	613/5834	10.51	0.188
Present	134/1697	7.90	−0.188
Typical Action-Chain Position			
Heads	245/2505	9.78	0.059
Other	502/5026	9.99	−0.059
Verb tense			
Others	719/6787	10.59	0.975
Preterit	28/744	3.76	−0.975
*Typical Action-Chain Position * Verb tense (interaction)*			
Heads: Others	237/2354	10.07	−0.133
Heads: Preterit	8/151	5.30	0.133
Tails and settings: Others	482/4433	10.87	0.133
Tails and settings: Preterit	20/593	3.37	−0.133
Region			
Center	138/2552	5.41	−0.621
East	434/1627	26.67	1.836
North	52/1645	3.16	−0.989
South	123/1707	7.21	−0.386
Corpus			
COSER	136/2031	6.70	−0.320
Twitter	611/5500	11.11	0.320
*Corpus*Region (Interaction)*			
Center*Twitter	103/2135	4.82	−0.858
East*Twitter	370/1130	32.74	0.619
North*Twitter	22/740	2.97	−0.014
South*Twitter	116/1495	7.76	0.252
Center*COSER	35/417	8.39	0.858
East*COSER	64/497	12.88	−0.619
North*COSER	30/905	3.31	0.014

Table 3. *(continued)*

Variable	*N*	*%*	*Estimate*
South*COSER	7/212	3.30	−0.252
Random effects		*Variance*	*Standard Deviation*
Locality		3.810	1.952
Locality:Negation		0.005	0.069
Locality: Tense		0.376	0.613
Noun		0.648	0.805
Model Summary			
AICc			3916.84
Pseudo R^2			0.62
C-Index			0.91

Note : The *bobyqa* optimizer for *glmer* was used in calculating the models. I report Nakagawa and Schielzeth's (2013) conditional pseudo-R^2, obtained with the *r.squared.glmm* function of the *MuMIn* package.

Table 4. Pluralized *haber* across tenses in Peninsular Spanish

Tense	*N*	*%*
Preterit	28/744	3.76
Present perfect	5/108	4.63
Imperfect	251/2922	8.59
Subjunctive present	118/1372	8.6
Conditional	31/301	10.3
Aspectual/modal auxiliaries	128/942	13.59
Future	186/1142	16.29

5.3 Geographic constraints

For the apparent-time and the geographic distribution of *haber* pluralization, the mixed-effects regression model in Table 3 estimates a higher likelihood for pluralized *haber* in the Twitter data (0.320 log-odds). Since Twitter users can be assumed to be much younger than the speakers that are included in the COSER corpus, these results may suggest that the frequency of *haber* pluralization is incrementing. This interpretation is supported by Paredes-García's (2016, 231) results, who compares a recent sample of university graduates from Madrid with a similar sample recorded in the late sixties and early seventies. In the recent sample, pluralized *haber* occurs more frequently, while also occuring in the speech of a larger proportion of the individuals.

Regarding the geographic distribution of *haber* pluralization in Peninsular Spanish, the mixed-effects regression model also indicates that pluralized *haber* is somewhat more likely in the Spanish South (Andalusia, Extremadura, and Murcia; −0.386 log-odds) than in central (Castile-La Mancha, Castile and Leon, and the Community of Madrid; −0.621 log-odds) and northern Spain (Asturias, Basque Country, Cantabria, Galicia, La Rioja, and Navarre; −0.989 log-odds). However, it is for eastern Spain (Aragón, Catalonia, and Valencian Community) that the model estimates the highest likelihood for pluralized *haber* (1.836 log-odds). This confirms the results of earlier variationist and dialectological work (e.g., Blas-Arroyo 1995, 2016; Gómez-Molina 2013; Llorente 1980).

In addition, the interaction between the corpora and these broad geographical regions shows that the favorable effect of the eastern region of Spain is larger for the Twitter data (0.619 log-odds) than it is for the COSER data (−0.619 log-odds); the same is true for the southern region (with an effect margin of 0.504 log-odds). In contrast, the regression supports that, in the COSER data, *haber* pluralization is more common in the center (0.858 log-odds) than it is in the Twitter data (−0.858 log-odds). For the north, pluralized *haber* is also slightly less common in the Twitter corpus (−0.014 log-odds) than it is in the COSER data (0.014 log-odds). If we approach the two corpora as two samples of different points in apparent time, this may suggest that *haber* pluralization is progressing in eastern and southern Spain, whereas it is retroceding in northern and central Spain.

To shed more light on the geographic spread of *haber* pluralization in apparent time, I specified a generalized additive model for the entire corpus, as well as a model for each of the two subcorpora. The results are displayed in Figures 2–4. In these figures, the darker shades of gray represent areas for which the model estimates lower likelihoods for pluralized *haber*. The brighter shades of gray indicate areas where pluralization is more likely. The regions for which there is not enough data are left without coloring. The light gray lines that appear on the maps can be considered as quantitative isoglosses, as they mark the contours of areas for which the model estimates different log-odds of pluralized *haber*.

Turning now to the estimates of the first model – which is based on the full dataset – Figure 2 shows that, as has been observed in traditional dialectology and earlier variationist work on Peninsular Spanish, the pluralized variant can more readily be found in eastern Spain. Within this region, the Valencian Community stands out as the geographic area that favors *haber* pluralization above all others (0–1 log-odds). A second, less pronounced hotspot can be found in the city of Barcelona and the surrounding areas (−0.5 to 0.5 log-odds), and a third is located in Murcia and the eastern half of its autonomous community (−0.5 to 0 log-odds).

As we move from east to west, pluralization becomes gradually less likely. Indeed, the province of Lerida (Catalonia), the occidental half of Aragón, the provinces of Cuenca and Albacete (Castile-La Mancha), and the western half of Murcia form a transition zone between the area that favors pluralization (−1.5 to −0.5 log-odds) and central Spain, which generally disfavors pluralization (−2.5 to −2 log-odds). However, the area around Toledo appears to constitute a special enclave, which has a higher likelihood of pluralized *haber* (−2.5 log-odds vs. −2 log-odds). In the South, the entire coastal area of Andalusia also displays a higher likelihood of pluralized *haber* than the rest of the country (−2 to −1 log-odds).

When we compare Figure 2 with Figures 3 and 4, we find that, due to the differing sizes of the datasets, the separate corpora do not allow to predict the distribution of pluralized and singular *haber* with the same level of detail. Still, contrasting the predictions of the generalized additive models for the Twitter (Figure 3) and the COSER data (Figure 4) confirms the results that were obtained with the generalized linear mixed model. Particularly, for both the Twitter data and the COSER data, the Valencian Community and the eastern half of Murcia appear as the driving force behind *haber* pluralization in Peninsular Spanish. For Catalonia, the COSER does not provide enough data points to make predictions.

As the Twitter data and the overall model, the COSER model supports that the Eastern part of Aragon and the Castilian provinces of Cuenca and Albacete form a transition zone between the Valencian community and the rest of the country (−2 to 0 log-odds). Although the data is sparse, the fact that this area extends into the western half of Murcia gives reason to believe that already for the older speakers included in the COSER, *haber* pluralization is more common in this area. When the two plots are compared, it also becomes evident that the transition zone extends further towards the west in the Twitter data. For this dataset, the −1.5 log-odds isogloss passes through Zaragoza, whereas this city is in the area between the −4 and −3 log-odds isoglosses in the COSER data. In addition, in the Twitter data most of the Andalusian province of Almeria is included in the transition zone (−1.5 to −1 log-odds), whereas the COSER model estimates a likelihood of −3 log-odds for this province.

Turning now to central and northern Spain, the mixed-effects model suggested that *haber* pluralization is retroceding in these two areas. For central Spain, this is supported by the comparison of the generalized additive models for the two datasets. Particularly, for the Twitter data, the −2 log-odds isogloss is situated in central Andalusia, leaving entire Castile-La Mancha and most of Andalusia in the −2.5 area. In contrast, in the COSER data, this isogloss is situated just south of Toledo, including a large part of the province of Toledo in the −2 log-odds region. This supports that *haber* pluralization may be retroceding in these provinces. In contrast, the differences between the estimates that are predicted for Northern

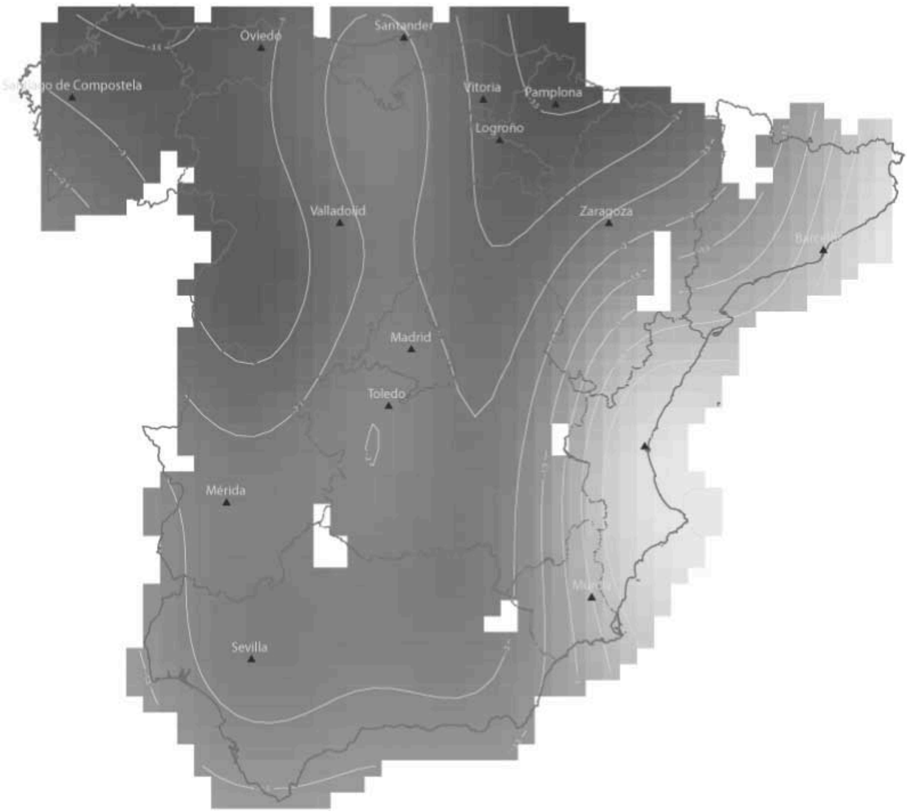

Figure 2. Predictions of a generalized additive model about the geographic distribution of *haber* pluralization in Peninsular Spanish: COSER and Twitter

Spain for the two datasets suggest that *haber* pluralization is advancing in the North, as the Twitter model generates higher log-odds for plural *haber* than the COSER model.

In sum, the results contributed by the models that were discussed in this section show that *haber* pluralization extends further westward in the Twitter data than in the COSER data. In addition, the data support that *haber* pluralization is retroceding in central Spain, but it appears to be advancing in northern and southern Spain. Let us now turn to the conclusions of this article.

6. Conclusions

In this article, I have examined the pluralization of presentational *haber* in a large corpus of naturally occurring language samples to investigate whether Peninsu-

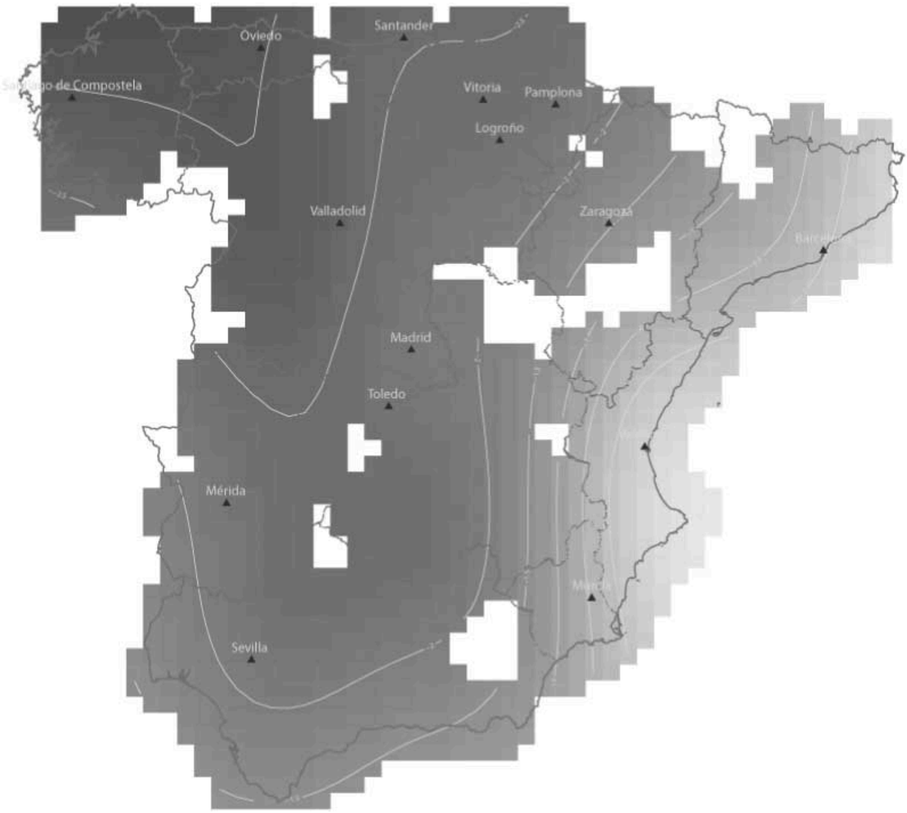

Figure 3. Predictions of a generalized additive model about the geographic distribution of *haber* pluralization in Peninsular Spanish: Twitter

lar data add further support to the hypothesis that *haber* pluralization is constrained by domain-general cognitive constraints on spreading activation. Also, I have examined whether *haber* pluralization may be spreading geographically in apparent time and whether or not its frequency increments over time in particular regions.

As for the first of these research questions, the results of this paper have shown that pluralized *haber* occurs more often with nouns that refer to typical action-chain heads, when negation is absent, and with tenses other than the preterit. These effects and their directionalities are stable across different varieties of Peninsular Spanish, while also being found in Caribbean Spanish. Therefore, the results of this paper add to a growing body of evidence that *haber* pluralization is indeed constrained by domain-general cognitive constraints on spreading activation. Given that the model relies on domain-general cognitive constraints, it seems plausible that it could also account for other cases of morphosyntactic vari-

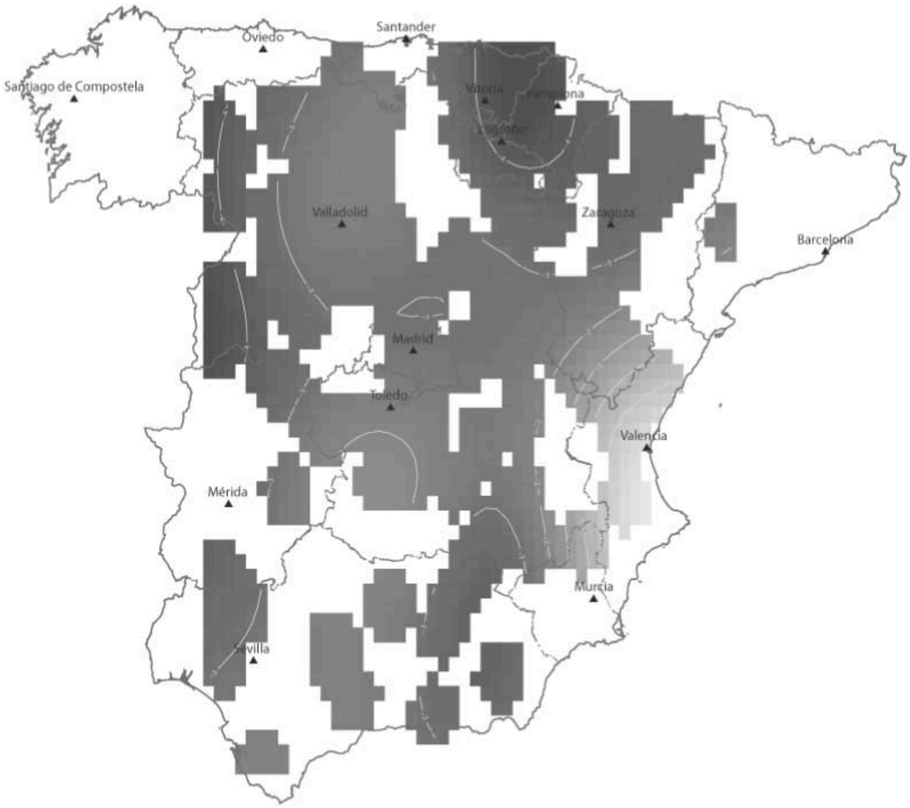

Figure 4. Predictions of a generalized additive model about the geographic distribution of *haber* pluralization in Peninsular Spanish: COSER

ation. Indeed, Claes and Johnson (under review) have recently shown that the model that was examined here also generates empirically correct predictions for the agreement variation that characterizes existential *there be* in British English. In addition, Claes (under review) has demonstrated that the same model can be extended to predict the variable use of subject personal pronouns in Cuban Spanish, whereas Geeraerts (2017) has argued for an extension of this model to lexical choice. This invites further research into these constraints and their effects on linguistic variation.

As for the second research question, which intended to investigate the geographic and the apparent-time distribution of *haber* pluralization in Peninsular Spanish, the results of the mixed-effects generalized linear model and the generalized additive mixed models support that pluralized *haber* appears to be progressing within individual regions as well as across regional varieties of Peninsular Spanish. This is supported by three lines of evidence. Firstly, the generalized linear mixed model estimates a higher overall probability of pluralized *haber* for the Twitter corpus, which can be taken to represent mainly young or middle-aged speakers of Peninsular Spanish. Secondly, the interaction between the regions and the corpora shows that, with the exception of the central region, *haber* pluralization has incremented its frequency in the entire country. Thirdly, the generalized additive mixed models support the same conclusions while also showing that the areas that mildly favor pluralized *haber* extend further westward. These data appear to suggest that *haber* pluralization is spreading from its epicenter in eastern Spain (particularly, the cities of Valencia, Barcelona, and Murcia) towards ever more distant varieties.

In sum, the conclusions of this article illustrate that combining more traditional linguistic resources such as spoken-language corpora with data culled from new social media may yield refreshing insights into the geographic and apparent-time distribution of morphosyntactic alternations. The sheer numbers of tokens that can be found on Twitter make it a particularly useful resource to the variationist who is interested in morphosyntactic variation, which is characterized by low occurrence rates. In this sense, this paper has also suggested a promising avenue for further research into the dialectal spread of morphosyntactic variation in Peninsular Spanish and beyond.

References

Bartón, Kamil. 2016. "MuMIn: Model Selection and Model Averaging Based on Information Criteria (AICc and alike)." Accessed May 2016. https://cran.r-project.org/web/packages/MuMIn/index.html.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2016. "lme4: Linear Mixed-Effects Models using 'Eigen' and S4." Accessed May 2016. https://cran.r-project.org/web/packages/lme4/index.html.

Bentivoglio, Paola, and Mercedes Sedano. 2011. "Morphosyntactic Variation in Spanish-Speaking Latin America." In *The handbook of Hispanic Sociolinguistics*, ed. by Manuel Díaz-Campos, 123–147. Oxford: Blackwell. https://doi.org/10.1002/9781444393446.ch8

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.

Blas-Arroyo, José-Luis. 2016. "Entre la Estabilidad y la Hipercorrección en un Antiguo 'Cambio desde Abajo': *Haber* Existencial en las Comunidades de Habla Castellonenses." *Lingüística Española Actual* 6 (1): 69–108.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference*. New York, NY: Springer.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

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. 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. https://doi.org/10.1075/resla.27.2.05cla

Claes, Jeroen. 2014d. *The Pluralization of Presentational Haber in Caribbean Spanish: A Study in Cognitive Construction Grammar and Comparative Sociolinguistics*. Antwerp: University of Antwerp PhD dissertation.

Claes, Jeroen. 2015. "Competing Constructions: The Pluralization of Presentational *Haber* in Dominican Spanish." *Cognitive Linguistics* 26 (1): 1–30. 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.

Claes, Jeroen. 2017. "La Pluralización de *Haber* Presentacional en el Español Peninsular: Datos de Twitter." *Sociolinguistic Studies* 11 (1): 41–64.

Claes, Jeroen. under review. "A Cognitive Sociolinguistic Model of Morphosyntactic Alternations: A Case Study of Subject Pronoun Expression in Cuban Spanish." *Review of Cognitive Linguistics*.

Claes, Jeroen and Daniel Ezra Johnson. under review. "Cognitive Linguistics and the Predictability of Effects: Agreement in English and Spanish Presentational Constructions." *Sociolinguistics Studies*.

Croft, William. 2003. *Typology and Universals*. Cambridge, MA: Cambridge University Press.

D'Aquino Ruiz, Giovanna. 2004. " *Haber* Impersonal en el Habla de Caracas. Análisis Sociolingüístico." *Boletín de Lingüística* 21: 3–26.

Davies, Mark. 2002-. *Corpus del Español. 100 million words (1200s-1900s)*. Accessed January 2016. http://www.corpusdelespanol.org/.

Dell, Gary S. 1986. "A Spreading-Activation Theory of Retrieval in Sentence Production." *Psychological Review* 92 (3): 283–321. https://doi.org/10.1037/0033-295X.93.3.283

DeMello, George. 1991. "Pluralización del Verbo *Haber* Impersonal en el Español Hablado Culto de Once Ciudades." *Thesaurus* 46: 445–471.

Díaz Campos, Manuel. 2003. "The Pluralization of *haber* in Venezuelan Spanish: A Sociolinguistic Change in Real Time." *IU Working Papers in Linguistics* 3 (5): 1–13.

Du Bois, John W. 2003. "Argument structure: Grammar in use." In *Preferred Argument Structure: Grammar as Architecture for Function*, ed. by John W. Du Bois, Lorraine E. Kumpf, and William Ashby, 11–60. Amsterdam/Philadelphia, PA: John Benjamins. https://doi.org/10.1075/sidag.14.04dub

Ellis, Nick C., and Fernando Ferreira-Junior. 2009. "Constructions and Their Acquisition: Islands and the Distinctiveness of Their Occupancy." *Annual Review of Cognitive Linguistics* 7: 187–220. https://doi.org/10.1075/arcl.7.08ell

Fernández-Ordóñez, Inés. 2005-. "Corpus Oral y Sonoro del Español Rural (COSER)." Accessed May 2016. http://www.corpusrural.es/.

Geeraerts, Dirk. 2017. "Entrenchment as Onomasiological salience." In *Entrenchment and the Psychology of Language Learning: How we Reorganize and Adapt Linguistic Knowledge*, ed. by Hans-Jörg Schmid, 153–174. Berlin/Boston, MA: De Gruyter.

Gentry, Jeff. 2016. "twitteR: R Based Twitter Client." Accessed August 2016. https://cran.r-project.org/web/packages/twitteR/index.html.

Gili-Gaya, Samuel. 1980. *Curso Superior de Sintaxis Española*. Barcelona: Vox.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: Chicago University Press.

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, Adele E. 2011. "Corpus Evidence of the Viability of Statistical Preemption." *Cognitive Linguistics* 22 (1): 131–153. https://doi.org/10.1515/cogl.2011.006

Gómez-Molina, José-Ramón. 2013. "Pluralización de *Haber* Impersonal en el Español de Valencia (España)." *Verba* 40: 253–284.

Kahle, David, and Hadley Wickman. 2016. "ggmap: Spatial Visualization with ggplot2." Accessed August 2016. https://cran.r-project.org/web/packages/ggmap/ggmap.pdf.

Keenan, Edward. 1976. "Towards a Universal Definition of Subject." In *Subject and topic*, ed. by Charles N. Li, 305–333. New York, NY: Academic Press.

Labov, William. 1994. *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford: Blackwell.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Volume 2: Descriptive Application*. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 2007. "Cognitive Grammar." In *The Oxford Handbook of Cognitive Linguistics*, ed. by Dirk Geeraerts and Hubert Cuyckens, 421–462. Oxford: Oxford University Press.

Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Langacker, Ronald W. 2010. "Cognitive Grammar." In *The Oxford Handbook of Linguistic Analysis*, ed. by Bernd Heine and Heiko Narrog, 87–110. Oxford: Oxford University Press.

Lastra, Yolanda, and Pedro Martín-Butragueño. 2016. "La concordancia de *haber* Existencial en la Ciudad de México." *Boletín de Filología* 51 (2): 121–145.

Llorente, Antonio Maldonado de Guevara. 1980. "Consideraciones sobre el Español Actual." *Anuario de Letras* 18: 5–61.

Lorenzo, Emilio. 1971. *El Español de Hoy: Lengua en Ebullición*. Madrid: Gredos.

Myachykov, Andriy, and Russel S. Tomlin. 2015. "Attention and Salience." In *Handbook of Cognitive Linguistics*, ed. by Ewa Dabrowska and Dagmar Divjak, 31–52. Berlin/New York, NY: De Gruyter Mouton.

Nakagawa, Shinichi, and 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. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Paredes-García, Florentino. 2016. "La pluralización del Verbo *Haber* Existencial en Madrid: ¿Etapas Iniciales de un Cambio Lingüístico?" *Boletín de Filología* 51 (2): 209–234.

Pato, Enrique. 2016. “La Pluralización de *Haber* en Español Peninsular.” In *En Torno a Haber: Construcciones, Usos y Variación desde el Latín hasta la Actualidad*, ed. by Carlota de Benito Moreno and Álvaro Octavio de Toledo, 357–391. Berlin: Peter Lang.

Prince, Ellen. 1992. “The ZPG letter: Subjects, Definiteness, and Information-status.” In *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, ed. by William C. Mann and Sandra A. Thompson, 295–326. New York, NY: John Benjamins. https://doi.org/10.1075/pbns.16.12pri

Quilis, Antonio. 1983. *La Concordancia Gramatical en la Lengua Española Hablada en Madrid*. Madrid: Consejo Superior de Investigaciones Científicas.

R Core Team. 2016. “R: A Language and Environment for Statistical Computing.” *Vienna: R Foundation for Statistical Computing*. https://www.R-project.org/.

Real Academia Española and Asociación de Academias de la Lengua Española. 2009. *Nueva Gramática de la Lengua Española*. Madrid: Espasa-Calpe.

Samper-Padilla, José Antonio and Clara Eugenia Hernández-Cabrera. 2012 “En Torno a los Usos Personales de *Haber* en el Español de Las Palmas de Gran Canaria”. In *Cum Corde et in Nova Grammatical: Estudios Ofrecidos a Guillermo Rojo*, ed. by Tomás Jiménez Juliá, Belén López Meirama, Victoria Vázquez Rozas, and Alexandre Veiga, 743–754. Santiago de Compostela: University of Santiago de Compostela.

Schmid, Hans-Jörg, and Helmut Küchenhoff. 2013. “Collostructional Analysis and Other Ways of Measuring Lexicogrammatical Attraction: Theoretical Premises, Practical problems and Cognitive Underpinnings.” *Cognitive Linguistics* 24 (3): 531–577. https://doi.org/10.1515/cog-2013-0018

Speelman, Dirk. 2014. “Logistic Regression: A Confirmatory Technique for Comparisons in Corpus Linguistics.” In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, ed. by Dylan Glynn and Justyna A. Robinson, 487–533. Amsterdam/Philadelphia, PA: John Benjamins. https://doi.org/10.1075/hcp.43.18spe

Szmrecsanyi, Benedikt. 2008. *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin/Boston, NY: De Gruyter.

Toutanova, Kristina, Christopher D. Klein, Dan Manning, and Yoram Singer. 2003. “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.” *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*. North American Chapter of the Association for Computational Linguistics. 252–259.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and Rolf Harald Baayen. 2014. “Lexical Differences between Tuscan Dialects and Standard Italian: Accounting for Geographic and Socio-Demographic Variation Using Generalized Additive Mixed Modeling.” *Language* 90 (3): 669–692.

Wood, Simon N. 2016. “mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation.” Accessed May 2016. https://cran.r-project.org/web/packages/mgcv/index.html.

Author's address

Jeroen Claes
KU Leuven
21 Blijde-Inkomststraat - bus 3308
3000 LEUVEN
Belgium

jeroenclaes@gmail.com