languages

MDPI

Article

Categories and Frequency: Cognition Verbs in Spanish Subject Expression

Catherine E. Travis [1,*] and Rena Torres Cacoullos [2]

[1] School of Literature, Languages and Linguistics and ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra 2601, Australia
[2] Department of Spanish, Italian and Portuguese and Center for Language Science, Pennsylvania State University, University Park, PA 16802, USA; rena@psu.edu
* Correspondence: catherine.travis@anu.edu.au

Abstract: Are semantic classes of verbs genuine or do they merely mask idiosyncrasies of frequent verbs? Here, we examine the interplay between semantic classes and frequent verb-form combinations, providing new evidence from variation patterns in spontaneous speech that linguistic categories are centered on high frequency members to which other members are similar. We offer an account of the well-known favoring effect of cognition verbs on Spanish subject pronoun expression by considering the role of high-frequency verbs (e.g., *creer* 'think' and *saber* 'know') and particular expressions (*(yo) creo* 'I think', *(yo) no sé* 'I don't know'). Analysis of variation in nearly 3000 tokens of unexpressed and pronominal subjects in conversational data replicates well-established predictors, but highlights that the cognition verb effect is really one of 1sg cognition verbs. In addition, particular expressions stand out for their high frequency relative to their component parts (for *(yo) creo*, proportion of lexical type, and proportion of pronoun). Further analysis of 1sg verbs with frequent expressions as fixed effects reveals shared patterns with other cognition verbs, including an association with non-coreferential contexts. Thus, classes can be identified by variation constraints and contextual distributions that are shared among class members and are measurably different from those of the more general variable structure. Cognition verbs in variable Spanish subject expression form a *class anchored in lexically particular constructions*.

Keywords: linguistic categories; frequency measures; constructions; variation constraints; contextual distribution; cognition verbs; Spanish; subject pronoun expression

check for updates

Citation: Travis, Catherine E., and Rena Torres Cacoullos. 2021. Categories and Frequency: Cognition Verbs in Spanish Subject Expression. *Languages* 6: 126. https://doi.org/10.3390/languages6030126

Academic Editors: Esther L. Brown, Javier Rivas, Juana M. Liceras and Raquel Fernández Fuertes

Received: 2 April 2021
Accepted: 20 July 2021
Published: 27 July 2021

1. Introduction

What is the relation between categories and frequent items? Categories or classes are variously conceived, but there is growing support for exemplar categories with a high-frequency central member. Consistent with this view, categories have been shown to be gradient rather than discrete; to derive from experienced tokens rather than abstract features; and to have central and marginal, rather than equally uniform, members. Evidence comes from diachronic studies of new constructions, which generalize from specific exemplars (e.g., Bybee and Torres Cacoullos 2009); from acceptability judgements, which are influenced by similarity to frequent tokens (e.g., Bybee and Eddington 2006); and from the acquisition of argument structure constructions, which are easier to learn when the input is skewed toward a high-frequency member (e.g., Goldberg et al. 2004) (cf., Bybee 2010, Chapters 4 and 5).

In this paper, we add novel evidence from spontaneous speech, to demonstrate that usage-based categories are defined by variation patterns. Categories or classes are those sets of items that share *contextual constraints*—linguistic factors conditioning the selection of one variant over its alternative in discourse (Poplack and Torres Cacoullos 2015, pp. 268–270); and *contextual distributions*—the relative frequency with which those factors occur in discourse. The shared variation patterns defining a category in turn display some differences from the constraints and distributions of the more general variable structure. Variation patterns, furthermore, contribute evidence for *lexically anchored categories*, in which one or more forms stand out for their high frequency but share patterns with other members of the category considered in the aggregate.

Subject expression in Spanish and the unresolved discussion of whether verb class or lexical frequency makes a contribution is a handy arena for revisiting categories and their underpinnings. A widely reported language-particular constraint on Spanish subject expression is the favoring effect of cognition verbs such as 'think' and 'know', which show higher than average rates of subject pronouns (vs. unexpressed subjects). However, the coherence of the semantic classes apparently conditioning variable Spanish subject expression has been discussed in view of both skewed distributions toward a few high frequency verbs such as *creer* 'to think' and *saber* 'to know', and differences in subject pronoun rate among individual verbs within a class (e.g., Bayley et al. 2013; Erker and Guy 2012; Orozco and Hurtado 2021).

The study of morphosyntactic variation has established that lexical effects (which reflect memory storage of speakers' experience with words and phrases) can outweigh online effects of the morphosyntactic or phonetic features constituting the context in which speakers make choices between variants. For example, in complement clause mood selection in French, the identity of the matrix verb (a lexical effect) determines the presence of an embedded subjunctive more than contextual (or online) factors such as the variable presence of the complementizer *que* (Poplack 1992, pp. 255–56). However, acknowledging lexical effects need not detract from semantic classes. An example from English is the favoring effect of motion verbs in choice of the present tense (vs. *will* or *be going to*) as a future expression (e.g., *she's going on the day after Thanksgiving*). Just two verb types (*go* and *come*) represent half the number of tokens of the present as a future expression, but the favoring of the present holds for motion verbs overall, even when these two frequent verbs are set aside (Torres Cacoullos and Walker 2009b, pp. 334–35).

We also know that there are verb forms or collocations—verb-person-tense-polarity combinations—the very frequency of which justifies singling them out as lexically particular constructions but which show parallels in linguistic conditioning with the general variable structure. Consider complement-taking predicates in speech corpora of English (cf., Torres Cacoullos and Walker 2009a). *I think*, *I guess*, *I remember* and a handful of other subject-verb combinations make up a large proportion of complement-taking predicate tokens as well as large proportions of their respective lexical types (e.g., *I think* alone makes up a quarter of all the data and more than half the tokens of all forms of *think*). They also show lower than average rates of complementizer *that*, or higher rates of occurrence with no complementizer. However, the linguistic conditioning of variable *that* in frequent forms parallels its conditioning in the general main-and-complement clause structure (for example, in that lexical vs. pronominal complement clause subjects favor the presence of *that* across the board).

To address lexically anchored categories, here we show the relevance of lexically particular constructions such as *(yo) creo* 'I think' to Spanish subject expression. We adapt the variationist comparative method (Poplack and Meechan 1998, pp. 130–32; Torres Cacoullos and Travis 2019, p. 656; Torres Cacoullos and Walker 2009a, p. 31), to compare the subject expression patterns of frequently co-occurring combinations with those of other cognition verbs and with non-cognition verbs. Analyses using mixed effects logistic regression models, first with individual verb as a random effect and then also incorporating frequent verb forms as fixed effects, reveal similar conditioning for these constructions as for other cognition verbs. Also shared is an association with non-coreferential contexts. Contextual constraints and distributions thus provide evidence that cognition verbs in Spanish variable subject expression form a category anchored in lexically particular constructions.

2. Variable Subject Expression in Spontaneous Speech

The spontaneous speech data for this study come from face-to-face conversation from the Corpus of Conversational Colombian Spanish, collected in 1997 and 2004 in the city of Cali (CCCS, cf., Travis 2005, pp. 9–25). A total of 37 speakers were recorded, 24 women and 13 men, most in their 20s and 30s (age range: 24 to 60). Participants were primarily from the middle class, recruited through the social network of two research assistants, an undergraduate student and a professor at a university in Cali. A total of 30 recordings were made of two- to five-party conversations between couples, friends and family members. They took place during naturally arising interactions such as while eating dinner, cooking, doing homework, or waiting for friends, and ranged from 7 to 40 min long, with an average of 18 min. This provided a total of nine hours of speech and nearly 100,000 words for analysis.

Variable Spanish subject expression in speech concerns the choice between pronominal and unexpressed subjects as a grammatical means of referring to an accessible subject. Lexical noun phrases fall outside the envelope of variation, as they are a site for introducing new, or inaccessible, information (Travis and Torres Cacoullos 2018, p. 83). We focus on first person and third person singular subjects (1sg and 3sg), as the most frequent to occur in spontaneous speech data. The competing variants are pre-verbal subject pronouns and null, or unexpressed, subjects. Non-human and non-specific subjects are set aside as they are rarely realized by personal pronouns. Also outside the envelope of variation are post-verbal subject pronouns, which are subject to distinct linguistic conditioning, and *wh*-interrogatives, where the variation in this variety is between post-verbal and unexpressed subjects (on the variable context for Spanish subject expression, see Torres Cacoullos and Travis 2018, pp. 138–41).

We extracted all instances of variable 3sg subjects in the corpus, and a comparable number of 1sg subjects from a portion of the corpus, giving a total of 2802 tokens with an overall rate of expression of 41%. Example (1) illustrates this variability, with the relevant 1sg and 3sg subject instances marked in bold, and unexpressed subjects with a Ø in the Spanish original and the subject in parentheses in the translation on the right. (See Appendix A for the transcription conventions.)

(1)
(re a health insurance policy Ángela is taking out for her husband, as a surprise for him)

1.	Ángela:	*... Ay,*	'... Oh,
2.		*qué pecado con Santi.*	I feel so bad about Santi.
3.		***Yo** quisiera que **él** supie=ra,*	**I**'d like him to know,
			(lit. 'that **he**$_{\text{SANTI}}$ knows')
4.		*porque --*	because --
5.		*... Pues,*	... Well,
6.		*para que **Ø** la [pueda usa=r].*	So that **(he**$_{\text{SANTI}}$**)** [can use it].'
7.	Sara:	[@@@]	[@@@]
8.	Ángela:	*... Qué hago?*	'... What should I do?'
9.	Sara:	*Y si le dices?*	'And if you tell him?
10.		*.. **Ø** Dice [que no] --*	.. **(He**$_{\text{SANTI}}$**)** will say [no] --
11.	Ángela:	*[No le va a gustar].*	'[It won't please him].
12.		*A él no le va a gustar.*	It won't please him.
13.		*... **Yo** después le digo.*	... **I**'ll tell him later.
14.		***Yo** creo que **yo** no me aguanto.*	**I** think that **I** won't be able to resist.
15.		*es que **yo** no me aguanto.*	It's that **I** won't be able to resist.'

(4 Seguro, 68–82)

3. Factors in Variable Subject Expression

To analyze the conditioning of the variation, we draw on the 40-year body of literature on subject expression in Spanish across different varieties and genres, which has identified similar linguistic constraints. Those that have received the most attention are subject person, accessibility, structural priming, tense-aspect-mood (TAM) and verb class (as reviewed, for example in Carvalho et al. 2015, pp. xiv–xv; Silva-Corvalán and Enrique-Arias 2017, pp. 172–87; Torres Cacoullos and Travis 2018, Ch. 5).

Subject person is often reported as the strongest constraint conditioning subject expression in Spanish when all persons are considered, with pronouns favored for 1sg over 3sg subjects. While the relatively higher rate has been attributed to the egocentric nature of the first person (e.g., Silva-Corvalán and Enrique-Arias 2017, p. 184), the difference between 1sg and 3sg diminishes if we consider also lexical subjects as a means of expression (Travis and Torres Cacoullos 2018, p. 78). What is important for the problem of defining the category under consideration here, as we will see, is that 1sg subjects show a greater tendency than 3sg to occur in environments that favor pronominal expression (in non-coreferential contexts and with cognition verbs).

A cross-linguistic effect is that of accessibility in accordance with the generalization that more accessible referents, that is, those which have been recently activated in the discourse, or represent given information, tend to be realized with less "coding material" (here, as unexpressed subjects), and less accessible referents with more "coding material" (here, pronouns) (Givón 1983a, p. 18). While accessibility has been operationalized in terms of distance from previous mention (e.g., in the papers in Givón 1983b), in Spanish, it is typically equated with coreferentiality, with pronominal subjects most likely to occur when there has been a switch in subject from the previous clause. This can be seen in example (1) above. In line 3, the subject referent (Santi) of the second clause is not coreferential with that of the preceding clause, and is expressed with the pronoun *él*, while in line 6, this same subject is retained, and is unexpressed.

A robust effect conditioning morphosyntactic variation in general is that of structural priming, as the tendency to repeat a previously used variant is observed in virtually every study that tests for it. For Spanish subject pronoun expression, priming was first examined across adjacent clauses (Cameron 1994), and it has more recently been demonstrated to occur most strongly between subjects with the same referent, even when separated by up to 10 clauses, in what is termed coreferential subject priming (Torres Cacoullos and Travis 2018, pp. 88-91). This phenomenon is illustrated in lines 14 and 15 in example (1) where, despite the coreferential contexts, the pronoun *yo* is repeated.

TAM is also often reported to have an effect, with subject pronouns tending to be favored in imperfective over perfective contexts. This is typically attributed either to ambiguity resolution (for example, 1sg and 3sg are ambiguous in the imperfect), or to the backgrounding function of some imperfective TAMs in discourse. Most consistent is the disfavoring effect of the perfective (preterit), which is tied to its greater tendency than imperfectives to be used in temporally sequential contexts in narratives and to occur with dynamic verbs (Torres Cacoullos and Travis 2018, pp. 97–101). Cognition verbs, on the other hand, occur proportionally more in the present tense than other verbs, as we discuss below, and it is this uneven distribution of TAMs across verb classes that is pertinent here, as present tense turns out to be a component of the cognition verb construction.

Of most interest for the relevance of lexically anchored categories is the widely reported constraint of verb class, with effects identified for dynamic verbs, which tend to favor unexpressed subjects more than stative verbs do (e.g., in example (1), *decir* 'to say' in lines 10 and 13 vs. *querer* 'to want' in line 3). Singled out in virtually all studies is the semantically rather than aspectually defined class of cognition verbs, which tends to favor expressed subjects the most (e.g., *que él supiera* 'that he knows' in line 3 and *yo creo* 'I think', in line 14). The favoring effect with cognition verbs has been attributed to the role of the pronoun to mark an utterance as the speaker's personal opinion (Aijón Oliva and Serrano 2010, p. 8), or "a higher level of speaker commitment" (Posio 2014, p. 14). Such pragmatic considerations, verified by quantitative patterns, may be part of what defines a construction (cf., Travis and Torres Cacoullos 2020; Vázquez Rozas and Enríquez Ovando 2020, pp. 225–26; see Section 7 below).

In sum, the same effects in Spanish subject expression have been repeatedly found, including for verb class. Nevertheless, the role of highly frequent verbs remains a topic of controversy and misunderstanding as to locus (frequency of what?) and direction (favoring or disfavoring?). What we highlight here is that lexical item and category need not be opposed. Rather, variation patterns reveal that lexically particular, frequent expressions act synergistically with the general verb class, in the case of cognition verbs, to favor subject pronouns.

4. Conditioning of Subject Pronoun Expression

We begin with an analysis of the constraints on subject expression in order to ascertain the impact of verb class alongside the set of predictors described above. To do this, we ran a series of regression analyses using generalized linear mixed effects models with the glmer() function in R (Bates et al. 2019; R Development Core Team 2019). Models were fit with subject pronoun realization (pronoun/zero) as the dependent variable, and person, accessibility, priming, TAM and verb class as independent variables. TAM was found not to be significant and was pruned from the model. We tested two-way interactions between each of the predictors, and of these, only subject person by accessibility was found to be statistically significant and included in the final model.

Speaker and verb (as lemma) were included as random intercepts. Including speaker as a random intercept is intended to ensure that the model considers individual differences so that inferences can be drawn beyond the study participants (see Guy 1980 on individual differences and the speech community). Including verb as a random intercept is intended to take account of lexical effects. It is, however, important to bear in mind that it is common in corpora for a large proportion of the data to be made up of items that occur only once. Such *hapax legomena* typically represent “roughly half the vocabulary size” by one account (Baayen 2001, p. 17). This is the case here, where 40% (122/294) of all verb lemmas present just one token.[1] Such low-frequency words cannot carry their own lexically specific probabilities, but rather must be associated with “lexicon-wide probabilities, which are based on data pooled across individual words” (Barth and Kapatsinski 2018, p. 103). From a modeling perspective, pooling such low-frequency words “avoid[s] making the random effect structure too sensitive to particularities” (Szmrecsanyi et al. 2016, p. 9). Accordingly, we pooled all *hapax legomena* into a single level in the random intercept for verb.

Table 1 provides the summary of the final model, with overall pronoun rates and token numbers for each linguistic context (level) presented in the first two columns; rows presenting the glmer model are shaded, and reference levels appear in unshaded rows. There are no surprises, as results are consistent with other studies—subject pronouns are favored by 1sg over 3sg subjects; by stative and cognition verbs over dynamic verbs, particularly so by cognition verbs; in non-coreferential over coreferential contexts; and for priming (measured as a coreferential mention in the previous 10 clauses), in the contexts of a previous pronoun and no previous mention (that is, in the absence of a prime) over the context of a previous unexpressed subject. In addition, as reflected in the significant interaction, the accessibility effect is weaker for 3sg than it is for 1sg. This is consistent with differential effects for 1sg vs. 3sg. As we have previously reported, the impact of distance from the previous mention is larger and becomes operative at shorter distances for 1sg than for 3sg subjects (Travis and Torres Cacoullos 2018, pp. 75–77).

Figure 1 presents the predicted rate of pronominal subjects by person and verb class, based on the output of the model in Table 1, and illustrates why a model including an interaction between person and verb class did not return a significant result for this interaction—the favoring of pronominal subjects most by cognition verbs holds for both 1sg and 3sg subjects, and the favoring of pronouns by 1sg over 3sg subjects holds for each verb class. But to understand the interplay between person and verb class, consider Figure 2, which breaks down each verb class by person. Here we observe that cognition verbs are overwhelmingly made up of 1sg subjects, which account for a full 88% of all instances, compared with 47% of dynamic, and 39% of stative, verbs. The converse also holds—over one fifth of 1sg subjects occur with cognition verbs, but under 3% of 3sg subjects do. Thus, though absence of a significant interaction in the model would indicate that 1sg and 3sg subjects pattern similarly with respect to verb class, this result should not be overinterpreted, given the relatively low number of 3sg cognition verbs (n = 39 vs. n = 296 for 1sg). Furthermore, other studies have reported a lack of a verb class effect for 3sg subjects (Shin 2014, p. 311).

Table 1. Generalized linear (mixed) model predicting an expressed subject.

	% Pronoun	Overall *n*	Estimate	Std. Error	Z	*p*
(Intercept 3sg Dynamic verbs, non-Coref, No Previous mention)			−0.9542	0.1515	−6.298	<0.001
Subject Person—3sg	32%	1413				
Subject Person—1sg	50%	1389	0.7315	0.1206	6.065	<0.001
Verb Class—Dynamic	34%	1783				
Verb Class—Stative	47%	684	0.5618	0.1858	3.024	<0.01
Verb Class—Cognition	65%	335	0.9118	0.2349	3.882	<0.001
Accessibility—non-Coreferential	48%	1644				
Accessibility—Coreferential	31%	1158	−0.3671	0.1236	−2.971	<0.01
Priming—No Previous mention	43%	1324				
Priming—Previous pronoun	54%	637	0.2602	0.1084	2.401	<0.05
Priming—Previous unexpressed	27%	841	−0.6667	0.1044	−6.388	<0.001
Subject Person x Accessibility **3sg non-Coreferential**	36%	702				
3sg Coreferential	27%	711				
1sg non-Coreferential	56%	942				
1sg Coreferential	37%	447	−0.4117	0.1783	−2.309	<0.05

Positive coefficients are associated with a higher rate of pronominal expression. Overall pronoun rate 41% (1143/2802). 37 speakers, variance = 0.21 (SD = 0.46); 173 verb types, variance = 0.22 (SD = 0.47). Log likelihood: −1705.7; AIC: 3431.4; BIC: 3490.7.

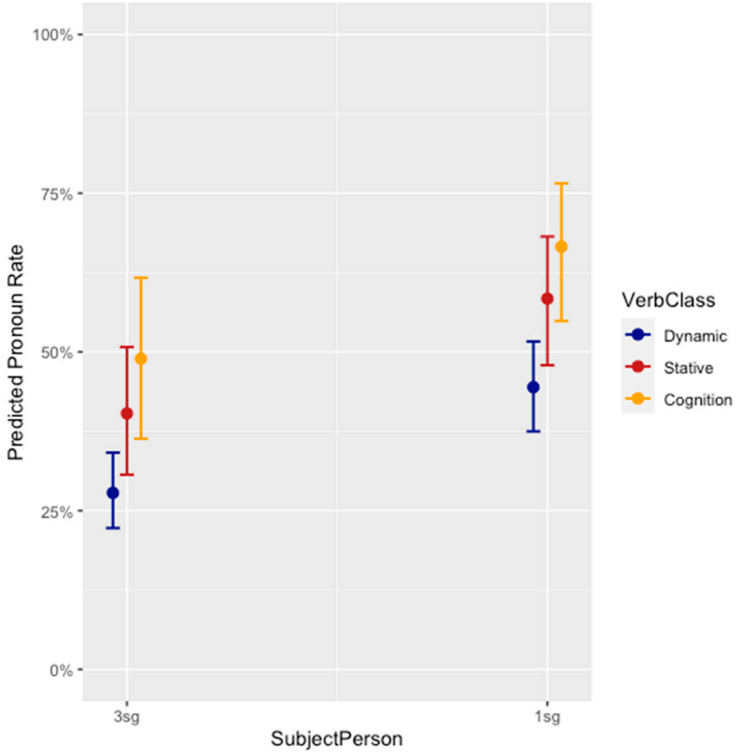

Figure 1. Predicted rate of subject expression for subject person by verb class, from model presented in Table 1. Pronominal subjects are favored with 1sg subjects and with cognition verbs.

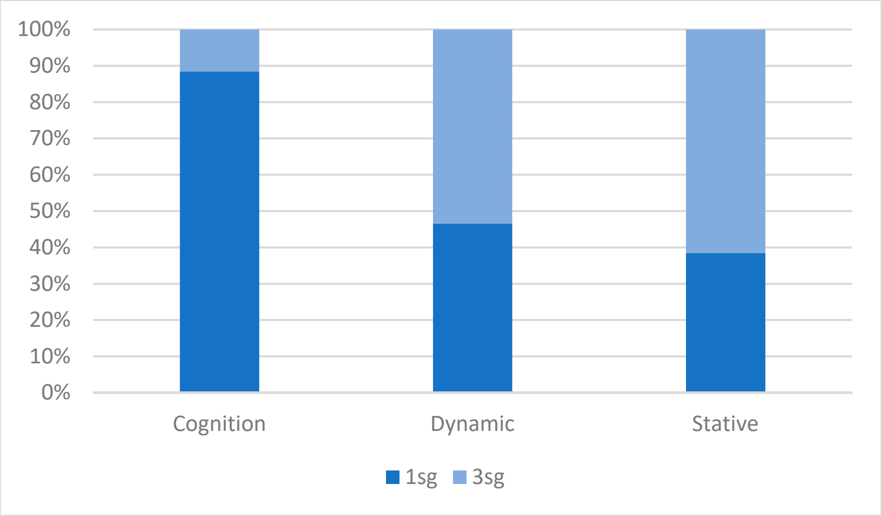

Figure 2. Data distribution for verb class by person. Cognition verbs are disproportionately 1sg.

The very robust finding reported in the Spanish subject expression literature for cognition verbs, then, is actually accounted for by cognition verbs *with 1sg subjects*, as observed by Torres Cacoullos and Travis (2018, p. 102). The favoring of subject pronouns with specifically 1sg cognition verbs suggests a 1sg cognition verb construction, which we can represent as [(*yo*) + COGNITION VERB$_{1SG}$]. Constructions, generally defined as pairings of form and meaning (e.g., Goldberg 2013), are operationalizable quantitatively as items tending to co-occur in particular contexts (Travis and Torres Cacoullos 2020, p. 140). Let us now explore the makeup of the 1sg cognition verb construction.

5. Straddling Lexical Types and Classes: Lexically Particular Constructions

What is the evidence for cognition verbs as a class within a [(*yo*) + COGNITION VERB$_{1SG}$] construction? Semantically, cognition verbs—also referred to as "knowledge" and "propositional attitude" predicates—express knowledge about, or attitude to, a proposition, and syntactically, they are characterized by their status as complement-taking predicates (Noonan 2007, pp. 124–30) (cf. also Givón 1984, p. 119). In actual usage, they also share a particular morphosyntactic profile in their tendency to occur with 1sg subjects as we have just seen, which has been related to the semantics of these verbs, since "the speaker must have access to the mental state to which the verb refers" (Weber and Bentivoglio 1991, p. 200). Cognition verbs are also mostly in the present tense (78%, 261/335 of the time, compared with just 50%, 897/1783, of dynamic verbs). A similar profile has been found in speech data from other varieties of Spanish (e.g., Shin 2014, p. 311; Torres Cacoullos and Travis 2018, pp. 101–2; Weber and Bentivoglio 1991, p. 203), and other languages, such as Swedish, Finnish, English and Mandarin (Dahl 2000, p. 5; Helasvuo 2014, p. 66; Scheibman 2001, p. 69; Tao 1996, p. 151). We can think of these semantic-morphosyntactic features as characterizing the class.

Despite these shared characteristics, the distribution of the verbs themselves is quite skewed, with a small number of verb types representing the majority of cognition verb tokens. For a view of this skewed distribution, we draw on a larger sample from the CCCS, which comprises all cognition verbs and all grammatical persons (n = 720). Figure 3 presents the distribution of this sample according to the most frequently occurring verb types and the most frequent forms in which those verbs occur, arranged by frequency of occurrence. Just two verbs constitute two thirds of all cognition verb tokens, *saber* 'to know' and *creer* 'to believe/think'.[2] Furthermore, two specific constructions from these verbs represent close to one third of all cognition verb tokens, *(yo) no sé* 'I don't know' and *(yo) creo* 'I think'. The third most frequent cognition verb, *pensar* 'to think', makes up only 10%, centered on two forms, *(yo) pensé* 'I thought' and *(yo) pienso* 'I think'. All other cognition verbs combined (a total of 12 verb types, e.g., *acordarse* 'to remember' (*n* = 40), *imaginarse* 'to imagine' (*n* = 47), *darse cuenta* 'realize' (*n* = 23), *entender* 'understand' (*n* = 23)) make up just one quarter of the data; these are collapsed in Figure 3, but grouped by person.[3]

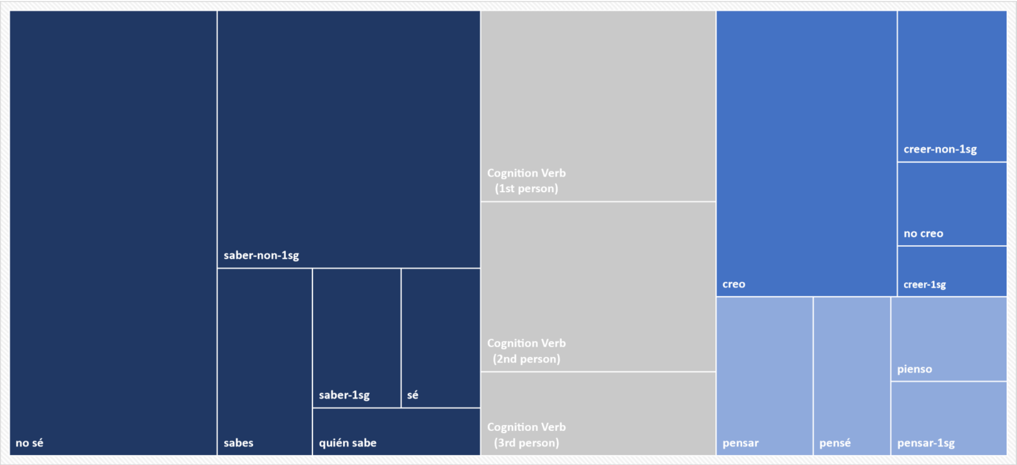

Figure 3. Distribution of cognition verbs by most frequent verb types and forms (*n* = 720). The lexical makeup of cognition verbs is skewed: two thirds of cognition verb tokens are accounted for by two verb types, *saber* 'to know' and *creer* 'to think', and one third by two expressions, *(yo) no sé* 'I don't know' and *(yo) creo* 'I think'.

saber 'to know (n = 340):	*(yo) no sé* 'I don't know', *(tú) sabes* 'you know', *(yo) sé* 'I know', *quién sabe* 'who knows'
creer 'to think/believe' (n = 135):	*(yo) creo* 'I think', *(yo) no creo* 'I don't think'
pensar 'to think' (n = 75):	*(yo) pensé* 'I thought', *(yo) pienso* 'I think'
Cognition verbs (other) (n = 170)	

Cross-linguistically, 1sg cognition verb combinations are recognized to have a specialized meaning, for example in English, related to "degrees of certainty or commitment to a proposition" (Noonan 2007, p. 125). This is consistent with the proposal that a main function of complement-taking predicate phrases is to "frame a clause in subjective epistemic terms" (Thompson 2002, p. 138). Thus, *(yo) no sé* 'I don't know' can be used to express lack of knowledge, but also as a discourse marker, for example to soften a statement as in (2) (Rivas and Brown 2009; Travis 2006, pp. 93–95). And though *(yo) creo* 'I think' derives from a verb meaning 'to believe', both it and *(yo) pienso* 'I think' are said to be "basic methods of expressing the epistemic-evidential stance of speakers", with *(yo) creo* being preferred in modern-day Spanish across varieties (in contrast to French, where, despite the existence of cognate verbs, *je pense* is the form that has won out for this epistemic use) (Vázquez Rozas 2015, pp. 579–580).

(2)
(re friends who made a poor business choice)

1.	Ángela:	*Pero es que tan --*	'But they are such --
2.		*.. ay,*	.. oh,
3.		***yo no sé,***	**I don't know,**
4.		*tan bobos.*	such silly people.'

(3 Familia, 959–962)

Just what, then, is the relationship between these lexically particular constructions and the set of lower frequency items? Does evidence of a class of cognition verbs remain once we take into account the behavior of these highly frequent expressions?

6. Classes and Lexically Particular Constructions: The Test of Variation Patterns

We test the status of cognition verbs as a class in relation to variable subject expression with 1sg subjects, singling out lexically particular constructions. One source of evidence comes from rates of subject pronoun expression. Here, we also bring to bear another source of evidence, which comes from variation patterns: *linguistic constraints* and *contextual distribution of the data*.

We first consider rates of subject pronoun expression. Figure 4 presents these for each of the 1sg forms identified in Figure 3, for 1sg other cognition verbs combined, and for 1sg non-cognition verbs. As can be seen, the favoring of subject pronouns is widely shared across the class. The subject pronoun rate for cognition verb forms ranges from 59% for *(yo) creo* to 92% for *(yo) pienso*, and as a set, 1sg other cognition verbs have a rate of 61%, substantially higher than 1sg non-cognition verbs at 46% (with the one exception of *(yo) no creo* 'I don't think', at 40%). We can verify, then, that the favoring of subject pronouns is not idiosyncratic behavior of frequent verbs, but generally holds across 1sg cognition verb items.

Furthermore, subject pronoun rates may be associated less with the lexical verb, and rather with particular verb-tense-subject-polarity combinations, most notably, *(yo) creo* and *(yo) no sé*. This too we verify in Figure 4, which shows that individual lexical types do not show uniformly high rates. For *saber* 'to know', positive polarity *(yo) sé* 'I know' has among the highest rates, at 85%, while *(yo) no sé* 'I don't know' is in the middle, at 66%. The converse is so for *creer* 'to think/believe', for which *(yo) creo* 'I think' has a higher rate than *(yo) no creo* 'I don't think' (59% vs. 40%).

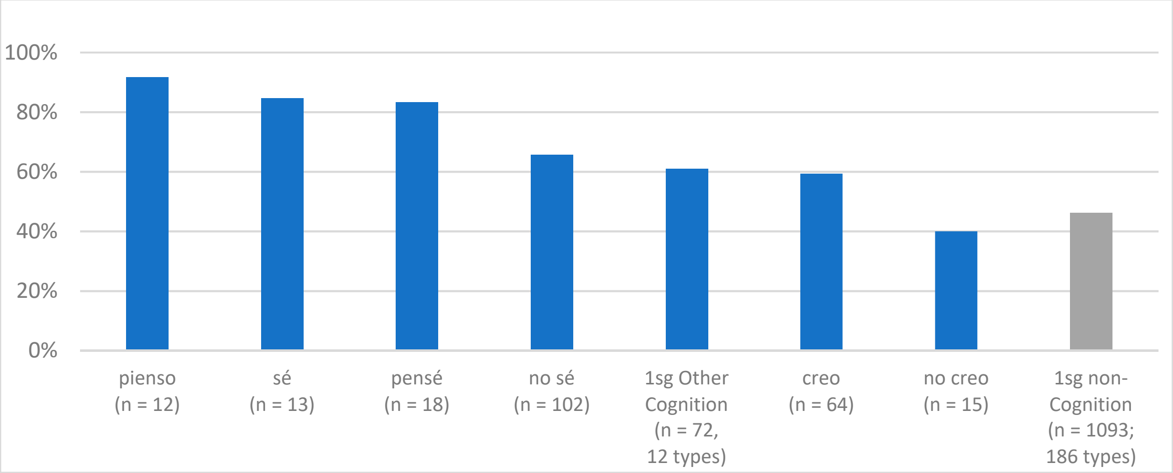

Figure 4. Rates of subject expression for verbs with 1sg subjects (forms with > 10 tokens). The favoring of subject pronouns generally holds across all cognition verb items; subject expression rates vary for individual lexical types according to particular constructions.

Next, we consider variation patterns. We propose to test the relevance of the cognition verb class and lexically particular constructions to subject expression by zooming in on the linguistic conditioning of variability. To do this, we ran a second generalized linear mixed effect model, this time on 1sg subjects only, again with subject pronoun realization (pronoun/zero) as the dependent variable; with accessibility, priming, and verb class as independent variables; and with speaker and verb as random intercepts (as previously, pooling verbs that occur only once). To compare cognition verbs and specific constructions, we reconfigure the predictor of verb class as one of verb class–construction, treating the two most frequently occurring forms—*(yo) creo* and *(yo) no sé*—as separate levels in this predictor, collapsing dynamic and stative verbs into one level of non-cognition verbs, and comparing with other cognition verbs as the reference level. In this way, instead of a blanket random effect for verb to account for lexical idiosyncrasies, we incorporate the most frequent forms as fixed effects in the model to directly test their relationship with other cognition verbs.

Table 2 presents the final model summary. Evidence for the role of lexically particular expressions in contouring the more general construction is seen in that, first, even when we separate out *(yo) creo* and *(yo) no sé*, the shared patterning of cognition verbs holds: non-cognition verbs have a significantly lower rate of pronoun expression than other cognition verbs. And second, there is no significant difference between other cognition verbs and neither *(yo) creo* nor *(yo) no sé*.[4] These results thus support cognition verbs as a class that is distinct from non-cognition verbs, with *(yo) creo* and *(yo) no sé* as members.

Table 2. Generalized linear (mixed) model predicting an expressed subject with 1sg verbs.

	% Pronoun	**Overall *n***	**Estimate**	**Std. Error**	***Z***	***p***
(Intercept Cognition verbs, non-Coreferential, No Previous mention)			0.7797	0.354	2.202	<0.05
Verb—Other Cognition	67%	130				
Verb—*(yo) creo*	59%	64	0.3349	0.4864	0.688	=0.49
Verb—*(yo) no sé*	66%	102	−0.567	0.5067	−1.119	=0.26
Verb—non-Cognition	46%	1093	−0.9461	0.3447	−2.745	<0.01
Accessibility—non-Coreferential	56%	942				
Accessibility—Coreferential	37%	447	−0.7542	0.1347	−5.597	<0.001
Priming—No Previous mention	55%	566				
Priming—Previous pronoun	59%	423	0.1594	0.1475	1.081	=0.28
Priming—Previous unexpressed	35%	400	−0.7445	0.1516	−4.911	<0.001

Positive coefficients are associated with a higher rate of pronominal expression. Overall pronoun rate 50% (697/1389). 25 speakers, variance = 0.10 (SD = 0.32); 141 verb types, variance = 0.64 (SD = 0.80). Log likelihood: −872.4; AIC: 1762.9; BIC: 1810.0.

We now come to contextual distribution, which must be recognized as a component of variation patterns in spontaneous speech. Consider the numbers presented in the "Overall *n*" column in Table 2, which show that 1sg verbs occur twice as often in non-coreferential vs. coreferential contexts (compared with 1.4 times as often for 1sg and 3sg combined, seen in Table 1). Might higher pronoun rates for 1sg cognition verbs merely reflect disproportionate occurrence in non-coreferential contexts? It was not possible to include an interaction for accessibility by verb class–construction in the model reported in Table 2 because of the skewed data distributions, which we turn to shortly. But a visualization of the effect of accessibility broken down by verb class–construction in Figure 5 shows that the higher pronoun rate in non-coreferential than coreferential contexts holds across cognition and non-cognition verbs, and that the favoring with cognition verbs holds across coreferential and non-coreferential contexts. Thus, the favoring of subject pronouns is a genuine effect for 1sg cognition verbs, that applies irrespective of coreferentiality (cf., Posio 2013, p. 283). Though the effect may appear to be stronger for non-cognition verbs, the vastly smaller token numbers for the other three levels in this predictor (see Table 2) must be taken into account in comparing the differences in the confidence intervals returned in the model.

Close examination of the contextual distribution for verb class–construction according to accessibility reveals that there is a linguistically significant relationship of "dependence" between the two predictors (Sankoff 1988, p. 986), in that cognition verbs are disproportionately used in non-coreferential contexts. This can be seen in Figure 6, which presents the proportion of the data occurring in coreferential vs. non-coreferential contexts for the four levels of the verb class–construction predictor in Table 2 and Figure 5: the proportion of the data occurring in non-coreferential contexts is lowest for 1sg non-cognition verbs at 64%, substantially lower than any of the cognition verbs categories, which are at 79% for 1sg other cognition verbs, 76% for *(yo) no sé*, and as high as 86% for *(yo) creo*. Some of the few instances of *(yo) creo* in a coreferential context are seen in line 14 in example (1) above. In contrast, 50% (711/1413) of 3sg subjects occur in non-coreferential contexts (not shown here). (On the distinct distribution for 1sg and 3sg by distance and differences in the workings of accessibility according to grammatical person, see Travis and Torres Cacoullos 2018, pp. 79–81).

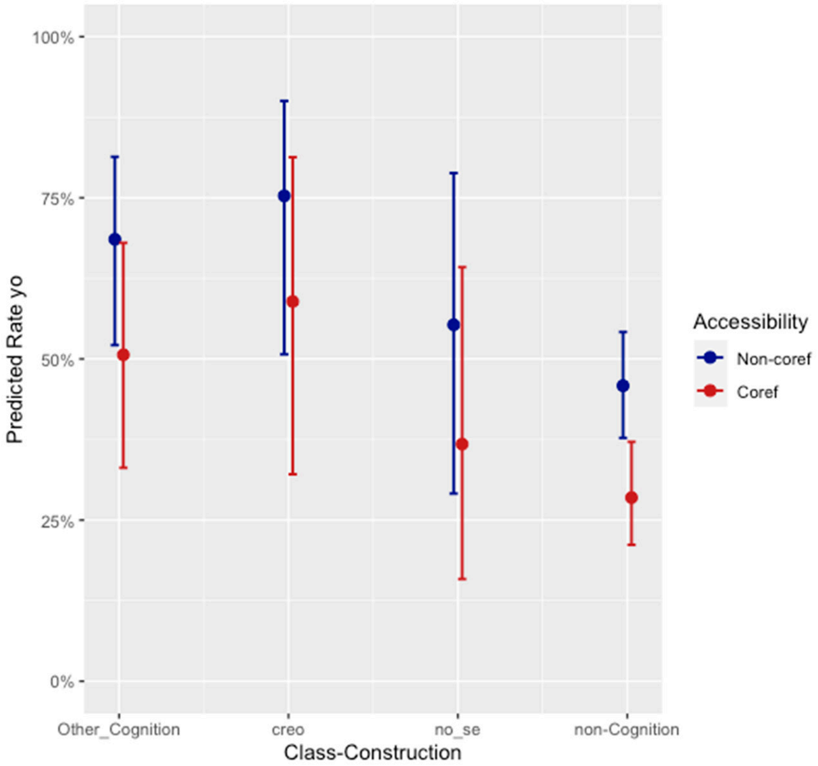

Figure 5. Predicted rate of subject expression for verb class-construction by accessibility, from model presented in Table 2. Pronominal subjects are favored in non-coreferential contexts for other cognition verbs, for the lexically particular constructions *(yo) creo* and *(yo) no sé*, and for non-cognition verbs.

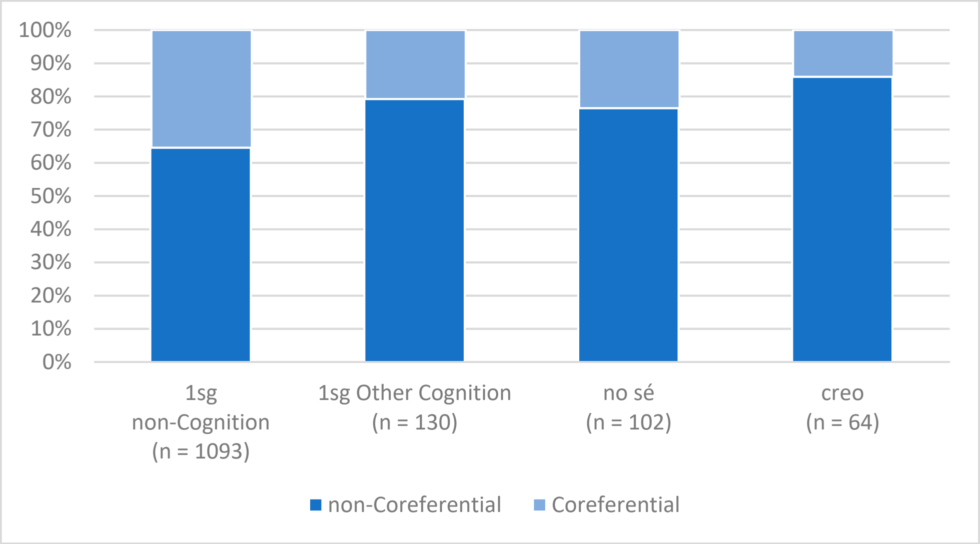

Figure 6. Data distribution for verb class-construction by accessibility. Cognition verbs (especially *(yo) creo*) occur disproportionately in non-coreferential contexts.

A greater preponderance of 1sg cognition verbs in non-coreferential contexts may be a general tendency, not specific to this dataset. A similar distribution is observed in the New Mexico Spanish-English Bilingual corpus and in the Santa Barbara Corpus of Spoken American English,[5] and appears to hold for Peninsular Spanish and European Portuguese spoken data (Posio 2013, pp. 283–84). Thus, contextual distribution is itself part of the variation patterns characterizing 1sg cognition verbs as a class, and *(yo) creo* and *(yo) no sé* as members of that class.

Contextual distribution may have a cumulative effect in enhancing the impact of favoring contexts. Here, the association with non-coreferential contexts may contribute to the higher pronoun rate overall for 1sg cognition verbs. This is because variation is conditioned not only by online, context-dependent factors (such as accessibility and priming for subject expression), but also by "usage history", reflecting speakers' cumulative prior experience with a form's contextual distribution (Bybee 2010, p. 43). Especially relevant is frequency of occurrence in a favorable context, according to which, high frequency of occurrence in a context that favors one variant over another may, via a cumulative effect, promote the choice of that variant across the board (Brown 2004; Bybee 2002).

The effect of frequency of occurrence in a favorable context has been observed in both phonology and morphosyntax. An example from phonology is variable word-initial [s] realization in New Mexican Spanish (Brown 2004). Reduction to [h] is favored in the phonetic environment of a preceding non-high vowel (an online, context-dependent factor). In this favoring environment, such as when following *no* 'no', a word like *señora* 'lady' is more likely to reduce than its masculine counterpart, *señor* 'gentleman'. This can be explained by the more frequent occurrence of the former in this favorable preceding non-high vowel context, often following feminine articles *la* 'the' and *una* 'an', compared with the corresponding *el* and *un* for the masculine *señor* (a storage, experience-dependent factor). Thus, word-initial [s] reduction is impacted by a word's overall frequency of occurrence in contexts that favor reduction.

An example from morphosyntax is the variable pluralization of Spanish *haber* 'there is/are' when the single argument is a plural noun (prescriptively, existential *haber* is always singular, the opposite of English) (Brown and Rivas 2012). One of the factors favoring plural verb morphology with *haber* is preponderance of the plural noun form in subject role. For example, among animate nouns, pluralization is more likely with *maestros* 'teachers' than with *abogados* 'lawyers' and among inanimate nouns, with *chismes* 'gossip tales' than with *ventanas* 'windows'. The first of each pair occurs more often than the second as a subject, thus more frequently agreeing with plural verbal morphology. This effect of grammatical relation probabilities is another example of how contextual distribution functions as a cumulative usage-based factor that impacts selection of variants in online production. For variable subject expression, the higher rate of occurrence in non-coreferential contexts may result in an overall higher rate of pronominal vs. unexpressed subjects for 1sg cognition verbs, which holds across the class—for *(yo) creo*, *(yo) no sé* and the set of less frequent verbs (Brown 2020).

7. Unravelling Frequency Effects: Conventionalized Chunks

The patterns of variation we examined above have allowed us to establish the internal coherence of the class of verbs in the [(*yo*) + COGNITION VERB$_{1SG}$] construction. Though some frequent forms make up the bulk of tokens of the class, those lexically specific constructions and the set of other cognition verbs exhibit shared patterns of favoring subject pronoun expression (Table 2) and association with non-coreferential contexts (Figure 6).

At the same time, frequency propels the conventionalization, or chunking, of these lexically particular constructions. Let us consider *(yo) creo*, which has received a lot of attention as the most frequent manifestation of cognition verbs and as strongly favoring subject pronoun expression across varieties of Latin American (e.g., Erker and Guy 2012, p. 539) and Peninsular Spanish (e.g., Aijón Oliva and Serrano 2010; Posio 2015, p. 67). For a broad overview, we go to the oral portion of the "Genre/Historical" sub-corpus of the Corpus del Español (Davies 2002).

Two frequency measures are pertinent to lexically particular expressions. Most obvious is *overall token frequency*. In the Corpus del Español, the form *creo* is the most frequent 1sg verb form, occurring nearly twice as often as the next most frequent form *sé* (*creo n* = 9215, *sé n* = 5885) (cf., Travis and Torres Cacoullos 2012, pp. 739–740). Complementary to overall token frequency is *relative frequency*, or the frequency of an expression relative to the component parts that make it up. The special status of *(yo) creo* is evident in its frequency relative to both the verb and the pronoun.

Figure 7 shows, first, that *(yo) creo* (*n* = 4165, in the darker shade) represents a large proportion of its corresponding lexical type, a full 84% of all instances of *creer* (pie chart on the left). In accounting for such a large proportion of *creer*, *(yo) creo* may be accessed independently as a unit, given that relative frequency affects degree of compositionality. Such chunking can be seen synchronically, for example, in that derived forms that are more frequent than the base word are more likely to be accessed whole (without being decomposed into affix + base), such that *impatient*, which is more frequent than *patient*, is more likely to be accessed directly than *imperfect*, which is less frequent than *perfect* (Hay 2001, pp. 1047, 1061). Diachronically, relative frequency is likewise important, for example, in the creation of complex prepositions such as *a pesar de* 'in spite of', which has become more frequent than *pesar* (originally, 'sorrow') (Torres Cacoullos 2006) (see Bybee 2010, pp. 138–46 on 'in spite of').

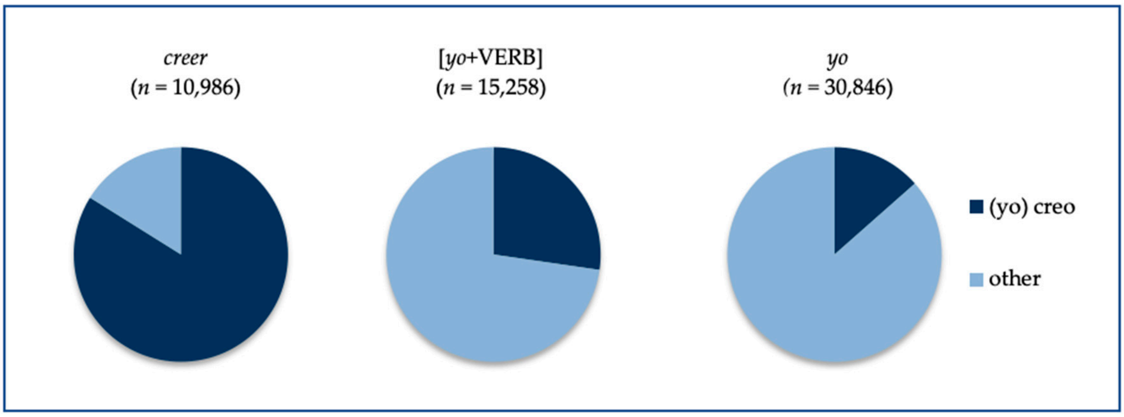

Figure 7. Relative proportion of data represented by *(yo) creo* in the oral portion of the Corpus del Español (Genre-Historical Corpus, Davies 2002). The relative frequency of *(yo) creo* with respect to its component parts is high: 88% of *creer*, 27% of [*yo*+VERB], and 14% of *yo*.

Second, the frequency of the string *yo creo* relative to other instances of the 1sg subject pronoun *yo* reveals a strong association between *creo* and *yo*. *Yo creo* represents over one quarter of the instances of *yo* immediately followed by a verb (27%, second chart in Figure 7). This proportion stands out in particular when we consider that the next most frequent verbs are the present perfect auxiliary *haber* 'to have' and the light verb *tener* 'to have', each of which occurs approximately one sixth as often as *yo creo* (just over 700 times).[6] Even considering all instances of *yo*, *yo creo* still represents a substantial proportion, 14% (third chart), followed by *yo no* (*n* = 3387). Compare this with the most frequent items to follow *I* in English conversation, *am* and *don't*, each accounting for around 10% of all instances of *I* (*n* = 414), and leading to contraction (*I'm*) and phonetic reduction (especially in *I don't know*) (Bybee and Scheibman 1999, pp. 590–92).

A consequence of its token and relative frequency would be to promote access of *yo creo* as a chunk. With chunking, the component parts of an expression become less analyzable and more independent of other instances of the same units (Bybee 2010, pp. 33–56). Thus, *(yo) creo* may be processed directly as a chunked unit, rather than through the paradigm of the verb *creer* or as the combination of pronoun and verb.

Chunking and conventionalization of *(yo) creo* as a lexically particular construction has developed over time. In pre-modern Spanish texts *creo* was neither frequent nor did it favor *yo* (Ramos 2016, p. 120; Vázquez Rozas 2015, p. 594). There has also been a generalization in meaning of the construction (Vázquez Rozas and Enríquez Ovando 2020). While *(yo) creo* is still used to mean 'believe in something', most frequent is the construction with a clausal complement in which *creo* has a general meaning of 'think'

(Posio 2014, p. 7). Even as a complement-taking predicate, *(yo) creo* may function more as an epistemic adverbial than a main clause, as can be seen in the two tokens of *creo* in lines 1 and 3 in (3) (Travis 2006, pp. 97–98). The loss of specific meaning features is accompanied by morphosyntactic decategorialization seen in its occurrence as a parenthetical (as in (4)), and internal fixedness, seen in the rarity of intervening elements such as adverbs (Posio 2014, pp. 10–11).

(3)

1.	Nury:	*Porque* ***creo*** *que ella se casa con él,*	'because **(I) think** she's going to marry him,
2.		*y se viene a vivir acá.*	and come and live here.
3.		.. ***Creo.***	.. **(I) think.**
4.		*No estoy segura.*	(I)'m not sure.'

[11 Estudios, 825–828]

(4)

1.	Rocío:	*y a eso está en la cevichería en Guapi,*	'and that's what it costs in the seafood restaurant in Guapi,
2.		***yo creo.***	**I think**.'

[18 Tumaco, 1340–1341]

Lexically particular constructions may differ across speech communities. While *(yo) creo* appears to be a pan-Hispanic phenomenon, the same is not so for other expressions. For example, in these data from Cali, Colombia, *(yo) no sé* 'I don't know' stands out, both for its frequent occurrence and for its favoring of subject expression. While the same is so in New Mexican Spanish (Torres Cacoullos and Travis 2018, pp. 169–70), in other varieties, *no sé* tends to occur without a subject pronoun (e.g., Cameron 1992, p. 102; Erker and Guy 2012, p. 539) (cf. Rivas and Brown 2009 for comparison of *no sé* across three varieties of Spanish). Similarly, *sabes* 'you know' has a low rate of subject pronoun expression in these data (just 10%, 3/29), but studies of other dialects have noted *tú sabes*, with the pronoun, as a fixed expression (Bayley et al. 2013, p. 25; Claes 2011, p. 196). Thus, lexically particular constructions "represent the conventional way of expressing an idea" (Bybee 2010, p. 81), and they conventionalize differently in accordance with community norms.

Recognition of lexically specific constructions helps us understand why token frequency as such does not have a uniform effect (cf., Bayley et al. 2013; Erker and Guy 2012). It has been suggested that frequency operates in interaction with other factors, so that "high frequency either activates or amplifies" other factors. For example, in the case of Spanish subject pronoun expression, person and verb class effects appear only among frequent verbs (Erker and Guy 2012, p. 545). This result, however, likely reflects the behavior of the lexically particular constructions, which are defined precisely by subject person and verb class. Moreover, there is no usage-based reason for an "expectation of consistent favoring of pronoun occurrence" by high frequency (Erker and Guy 2012, p. 539). High frequency promotes reductive sound change but has a "conserving effect" for regularization and analogical change (Bybee 2010, pp. 24, 75) and, therefore, with lexically particular constructions, conventionalization may go in either direction, of elevated *or* depressed pronoun rates.

8. Conclusions

We conclude that lexically particular construction and general class effects are synergistic: highly frequent, particular expressions contribute to shaping general patterns, as the center of classes to which they attract members with shared semantic-morphosyntactic characteristics. These shared characteristics are seen in quantitative variation patterns in discourse, which are constituted by both contextual constraints and contextual distributions.

Lexically particular constructions are pertinent to the linguistic conditioning of variation and thus should be taken into account for interpretation of results (regardless of whether our statistical models include lexical item as a random effect; Torres Cacoullos and Travis 2019, p. 686). Decades of study of variable Spanish subject pronoun expression have established broad agreement on the conditioning factors. These are replicated here, but when we consider the relationships between predictors, it is clear that cognition verbs are overwhelmingly used in the first person singular, such that the widely reported cognition verb effect is really one of 1sg cognition verbs. While discussions on the priority of frequent verbs as opposed to semantic classes have recognized strikingly different frequencies of lexical types, there are also strikingly different frequencies of particular verb-tense-subject-polarity combinations. The role of such particular expressions is overlooked in analyses of subject expression focusing on either cognition verbs as a class or on specific frequent verb types.

Variation patterns provide a measure of category status. If there is a cognition class, patterns of 1sg subject expression will be shared across cognition verbs, distinguishing them from other verbs, and this will apply to high-frequency lexically particular instances as it will to other members. Here, we have shown that *(yo) creo* and (*yo) no sé* have shared patterns with other cognition verbs, including the favoring of subject pronouns and an association with non-coreferential contexts. At the center of the class is *(yo) creo*: on the strength of high token and relative frequencies in addition to the favoring of subject pronoun expression, it is best considered a chunked unit which, nevertheless, contours the class. Thus, semantic classes of verbs are centered on high-frequency members. For Spanish variable subject expression, cognition verbs form a category anchored in 1sg lexically particular constructions.

Author Contributions: This paper derives from data collected as part of a PhD completed by Catherine Travis in 2002 at La Trobe University, Australia. Both authors contributed equally to the writing of this paper, and both authors have read and agreed to the published version of the manuscript.

Funding: The compilation of the Corpus of Conversational Colombian Spanish was funded by an Australian Postgraduate Research Award granted to Catherine Travis (1996–2001), and the Latin American and Iberian Institute of the University of New Mexico (2002). Data coding and analysis were supported by a joint grant to both authors from the National Science Foundation (BCS-1019112/1019122), and the ARC Centre of Excellence for the Dynamics of Language (CE140100041).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the Faculty of the Humanities of La Trobe University (protocol approval #111/96, approved 1996).

Informed Consent Statement: Informed consent was obtained from all participants in this study.

Data Availability Statement: The data presented in this study may be made available on request from the corresponding authors. The data are not publicly available to accord with the informed consent guidelines provided to the participants.

Acknowledgments: We are grateful to the anonymous reviewers and to the editors of this volume, Esther Brown and Javier Rivas, for valuable feedback on this paper, and to Danielle Barth for advice on the statistical analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Examples are reproduced verbatim from the transcripts; information in parentheses following each example provides the recording number and name, and the numbers of lines presented. All names given are pseudonyms.

Transcription Conventions (Du Bois et al. 1993)

Carriage return	new Intonation Unit	=	lengthened syllable
.	final intonation contour	[ ]	overlapped speech
,	continuing intonation contour	..	short pause (0.5 s)
?	appeal intonation contour	...	medium pause (0.5–0.7 s)
--	truncated intonation contour	@	one syllable of laughter

Notes

[1] A similar proportion of verb types occur only once in other subject expression data sets: in the New Mexico Spanish-English Bilingual corpus (Torres Cacoullos and Travis 2018, Chapters 2 and 3), 47% (217/457), and in the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000–2005), 56% (142/255) (see Torres Cacoullos and Travis 2018, p. 10 for a summary of these datasets). The same skewing does not generally apply to speakers, however. In the CCCS, just three of the 37 speakers produce only one token (though 20 speakers, or over one half, produce under 30 tokens).

[2] In contrast, for dynamic verbs, the most frequent, *decir* 'to say', represents 17% of the total number of tokens (n = 307), followed by *hacer* 'to do' at 6% (n = 107); for statives, the most frequent are *ser* 'to be' at 30% (n = 203), *tener* 'to have' at 24% (n = 162), and *estar* at 18% (n = 123) (with the 1sg and 3sg CCCS dataset used for the studies in this paper).

[3] These verbs also occur in some frequent combinations—*(yo) no me acuerdo* 'I don't remember' (n = 17) and the discourse marker *imaginate* 'imagine!', for which the *voseo* verb form is also part of the construction (n = 22 *voseo* vs. 2 *tuteo*) (cf., Travis 2006, p. 101).

[4] A releveled model with *(yo) creo* as the reference level indicates that it is not significantly distinct from *(yo) no sé* ($\beta = 0.19$, $p = 0.58$); it is marginally significantly distinct from non-cognition verbs ($\beta = -0.50$, $p = 0.07$). (To ease model convergence, in this analysis we included a random intercept for speaker only, and not for verb.)

[5] In the New Mexico Spanish-English Bilingual corpus, 70% (448/640) of 1sg cognition verbs occur in non-coreferential contexts compared with 48% (1264/2656) of 1sg non-cognition verbs and 45% (1013/2275) of all 3sg verbs. In the Santa Barbara Corpus of Spoken American English, 79% (79/100) of 1sg cognition verbs occur in non-coreferential contexts vs. 46% (161/347) for 1sg non-cognition verbs and 46% (247/540) for 3sg.

[6] *Tener* 'to have' occurs as a possessive verb, as part of Verb-Noun units, and with *que* + V as a modal of obligation, 'have to V'.

References

Aijón Oliva, Miguel Ángel, and María José Serrano. 2010. El hablante en su discurso: Expresión y omisión del sujeto de *creo*. *Oralia* 13: 7–38.

Baayen, Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Press.

Barth, Danielle, and Vsevolod Kapatsinski. 2018. Evaluating linguist mixed-effects models of corpus-linguistic data in light of lexical diffusion. In *Mixed-Effects Regression Models in Linguistics: Quantitative Methods in the Humanities and Social Sciences*. Edited by Dirk Speelman, Kris Heylen and Dirk Geeraerts. Cham: Springer, pp. 99–116.

Bates, Douglas, Martin Mächler, Ben Bolker, Steve Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, and John Fox. 2019. lme4: Linear Mixed-Effects Models Using 'Eigen' and S4. R package, version 1.1–121. Available online: https://cran.r-project.org/web/packages/lme4/index.html (accessed on 1 January 2021).

Bayley, Robert, Kristen Greer, and Cory Holland. 2013. Lexical frequency and syntactic variation: A test of a linguistic hypothesis. *University of Pennsylvania Working Papers in Linguistics* 19: 21–30.

Brown, Esther L. 2004. The Reduction of Initial /s/ in New Mexican Spanish: A Usage-Based Approach. Ph.D. thesis, Department of Spanish and Portuguese, University of New Mexico, Albuquerque, NM, USA.

Brown, Esther L. 2020. *The Long-Term Accrual in Memory of Contextual Conditioning Effects*. Paper presented at the Center for Language Science Speaker Series. State College: Center for Language Science, October 9.

Brown, Esther L., and Javier Rivas. 2012. Grammatical relation probability: How usage shapes analogy. *Language Variation and Change* 24: 317–41. [CrossRef]

Bybee, Joan, and David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82: 323–55. [CrossRef]

Bybee, Joan, and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37: 575–96. [CrossRef]

Bybee, Joan, and Rena Torres Cacoullos. 2009. The role of prefabs in grammaticization: How the particular and the general interact in language change. In *Formulaic Language, Vol. 1: Distribution and Historical Change*. Edited by Roberta L. Corrigan, Edith Moravcsik, Hamid Ouali and Kathleen Wheatley. Amsterdam: John Benjamins, pp. 187–217.

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14: 261–90. [CrossRef]

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Cameron, Richard. 1992. Pronominal and Null Subject Variation in Spanish: Constraints, Dialects, and Functional Compensation. Unpublished Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Cameron, Richard. 1994. Switch reference, verb class and priming in a variable syntax. *Papers from the Regional Meeting of the Chicago Linguistic Society: Parasession on Variation in Linguistic Theory* 30: 27–45.

Carvalho, Ana M., Rafael Orozco, and Naomi Lapidus Shin. 2015. Introduction. In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspectiv*. Edited by Ana M. Carvalho, Rafael Orozco and Naomi Lapidus Shin. Georgetown: Georgetown University Press, pp. xiii–xxvi.

Claes, Jeroen. 2011. Constituyen las Antillas y el Caribe continental una sola zona dialectal Datos de la variable expresión del sujeto pronominal en San Juan de Puerto Rico y Barranquilla, Colombia. *Spanish in Context* 8: 191–212. [CrossRef]

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7: 37–77. [CrossRef]

Davies, Mark. 2002. Corpus del Español: 100 million Words, 1200s–1900s. Available online: http://www.corpusdelespanol.org (accessed on 20 July 2021).

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse*. Edited by Jane Edwards and Martin Lampert. Hillsdale: Lawrence Erlbaum Associates, pp. 45–89.

Du Bois, John W., Wallace L. Chafe, Charles Myer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. *Santa Barbara Corpus of Spoken American English, Parts 1–4*. Philadelphia: Linguistic Data Consortium.

Erker, Daniel, and Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88: 526–57. [CrossRef]

Givón, T., ed. 1983a. Topic continuity in discourse: An introduction. In *Topic Continuity in Discourse: A Quantitative Cross-Linguistic Study*. Amsterdam: John Benjamins, pp. 1–41.

Givón, T., ed. 1983b. *Topic Continuity in Discourse: A Quantitative Cross-Linguistic Study*. Amsterdam: John Benjamins.

Givón, T. 1984. *Syntax: A Functional-Typological Introduction*. Amsterdam and Philadelphia: John Benjamins, vol. 1.

Goldberg, Adele E. 2013. Constructionist approaches. In *The Oxford Handbook of Construction Grammar*. Edited by T. Hoffman and G. Trousdale. Oxford: Oxford University Press, pp. 15–31.

Goldberg, Adele E., Devin M. Casenhiser, and Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 13: 289–316. [CrossRef]

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Language in Time and Space*. Edited by William Labov. New York: Academic Press, pp. 1–36.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39: 1041–70. [CrossRef]

Helasvuo, Marja-Liisa. 2014. Agreement or crystallization: Patterns of 1st and 2nd person subjects and verbs of cognition in Finnish conversational interaction. *Journal of Pragmatics* 63: 63–78. [CrossRef]

Noonan, Michael. 2007. Complementation. In *Language Typology and Syntactic Description: Complex Constructions*, 2nd ed. Edited by Timothy Shopen. Cambridge: Cambridge University Press, vol. 2, pp. 52–150.

Orozco, Rafael, and Luz Marcela Hurtado. 2021. A variationist study of subject pronoun expression in Medellín, Colombia. *Languages* 6: 5. [CrossRef]

Poplack, Shana, and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* 2: 127–38. [CrossRef]

Poplack, Shana, and Rena Torres Cacoullos. 2015. Linguistic emergence on the ground: A variationist paradigm. In *The Handbook of Language Emergence*. Edited by Brian MacWhinney and William O'Grady. Malden: Wiley-Blackwell, pp. 267–91.

Poplack, Shana. 1992. The inherent variability of the French subjunctive. In *Theoretical Analysis in Romance Linguistics*. Edited by Christiane Laeufer and Terrell A. Morgan. Amsterdam: John Benjamins, pp. 235–63.

Posio, Pekka. 2013. The expression of first-person-singular subjects in spoken Peninsular Spanish and European Portuguese: Semantic roles and formulaic sequences. *Folia Linguistica* 47: 253–91. [CrossRef]

Posio, Pekka. 2014. Subject expression in grammaticalizing constructions: The case of *creo* and *acho* 'I think' in Spanish and Portuguese. *Journal of Pragmatics* 63: 5–18. [CrossRef]

Posio, Pekka. 2015. Subject pronoun usage in formulaic sequences: Evidence from Peninsular Spanish. In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*. Edited by Ana M. Carvalho, Rafael Orozco and Naomi Lapidus Shin. Washington, DC: Georgetown University Press, pp. 59–78.

R Development Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, Available online: http://www.R-project.org (accessed on 1 January 2021).

Ramos, Miguel. 2016. Continuity and change: First person singular subject pronoun expression in earlier Spanish. *Spanish in Context* 13: 103–27. [CrossRef]

Rivas, Javier, and Esther L. Brown. 2009. *No sé* as a discourse marker in Spanish: A corpus-based approach to a cross-dialectal comparison. In *A Survey on Corpus-Based Research: Panorama de Investigaciones Basadas en Corpus*. Edited by Aquilino Sánchez and Pascual Cantos. Murcia: AELINCO, pp. 631–45.

Sankoff, David. 1988. Variable rules. In *Sociolinguistics: An International Handbook of the Science of Language and Society*. Edited by Ulrich Ammon, Norbert Dittmar and Klaus J. Mattheier. Berlin: Walter de Gruyter, vol. 2, pp. 984–97.

Scheibman, Joanne. 2001. Local patterns of subjectivity in person and verb type in American English conversation. In *Frequency and the Emergence of Linguistic Structure*. Edited by Joan Bybee and Paul J. Hopper. Amsterdam: John Benjamins, pp. 61–89.

Shin, Naomi Lapidus. 2014. Grammatical complexification in Spanish in New York: 3sg pronoun expression and verbal ambiguity. *Language Variation and Change* 26: 303–30. [CrossRef]

Silva-Corvalán, Carmen, and Andrés Enrique-Arias. 2017. *Sociolingüística y Pragmática del Español*, 2nd ed. Washington, DC: Georgetown University Press.

Szmrecsanyi, Benedikt, Douglas Biber, Jesse Egbert, and Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28: 1–29. [CrossRef]

Tao, Hongyin. 1996. *Units in Mandarin Conversation: Prosody, Discourse and Grammar*. Amsterdam and Philadelphia: John Benjamins.

Thompson, Sandra A. 2002. 'Object Complements' and conversation: Towards a realistic account. *Studies in Language* 26: 125–63. [CrossRef]

Torres Cacoullos, Rena, and Catherine E. Travis. 2018. *Bilingualism in the Community: Code-Switching and Grammars in Contact*. Cambridge: Cambridge University Press.

Torres Cacoullos, Rena, and Catherine E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57: 653–92. [CrossRef]

Torres Cacoullos, Rena, and James A. Walker. 2009a. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47: 1–43. [CrossRef]

Torres Cacoullos, Rena, and James A. Walker. 2009b. The present of the English future: Grammatical variation and collocations in discourse. *Language* 85: 321–54. [CrossRef]

Torres Cacoullos, Rena. 2006. Relative frequency in the grammaticization of collocations: Nominal to concessive *a pesar de*. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*. Edited by Timothy L. Face and Carol A. Klee. Somerville: Cascadilla Proceedings Project, pp. 37–49.

Travis, Catherine E. 2005. *Discourse markers in Colombian Spanish: A study in polysemy*. Berlin and New York: Mouton de Gruyter.

Travis, Catherine E. 2006. Subjetivización de construcciones: Los verbos 'cognitivos' en el español conversacional. In *Serie Memorias del VIII Encuentro Internacional de Lingüística en el Noroeste*. Edited by Rosa María Ortiz Ciscomani. Hermosillo, Sonora and Mexico: UniSon, vol. 2, pp. 85–109.

Travis, Catherine E., and Rena Torres Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics* 23: 711–48. [CrossRef]

Travis, Catherine E., and Rena Torres Cacoullos. 2018. Discovering structure: Person and accessibility. In *Questioning Theoretical Primitives in Linguistic Inquiry (Papers in Honor of Ricardo Otheguy)*. Edited by Naomi Lapidus Shin and Daniel Erker. Amsterdam and Philadelphia: John Benjamins, pp. 67–90.

Travis, Catherine E., and Rena Torres Cacoullos. 2020. The role of pragmatics in shaping linguistic structures. In *The Routledge Handbook of Spanish Pragmatics*. Edited by Dale Koike and J. César Félix-Brasdefer. London and New York: Routledge, pp. 129–47.

Vázquez Rozas, Victoria, and Araceli Enríquez Ovando. 2020. *(Yo) creo* en el español de la Ciudad de México y de Galicia: Diferencias de gramaticalización. In *Evidencialidad: Determinaciones léxicas y construccionales*. Edited by Ricardo Maldonado and Juliana de la Mora. Ciudad de México: UNAM/UAQ, pp. 199–239.

Vázquez Rozas, Victoria. 2015. Dialogue and epistemic stance: A diachronic analysis of cognitive verb constructions in Spanish. *eHumanista/IVITRA* 8: 577–99.

Weber, Elizabeth G., and Paola Bentivoglio. 1991. Verbs of cognition in spoken Spanish: A discourse profile. In *Discourse Pragmatics and the Verb: The Evidence from Romance*. Edited by Suzanne Fleischman and Linda R. Waugh. London: Routledge, pp. 194–213.