Research Article

Malte Rosemeyer*

Data-driven identification of situated meanings in corpus data using Latent Class Analysis

https://doi.org/10.1515/opli-2024-0029
received December 20, 2023; accepted September 25, 2024

Abstract: Identifying the meanings of grammatical elements in context is a major challenge for corpus-linguistic studies of grammatical variation. This study proposes a novel solution to this problem. I describe the situated meanings of grammatical elements as latent constructs, i.e., social concepts that cannot be observed directly but need to be inferred from the way that speakers behave. I use Latent Class Analysis (LCA) to create a data-driven typology of meanings for three modal periphrases in spoken Spanish and compare this typology to manual classification of the data in terms of modality. My findings show that (a) the situated meanings identified by the LCA do not directly correspond to the modal meanings that are commonly assumed to govern the variation between the three periphrases, and (b) the data-driven typology of meanings explains better the variation between these periphrases.

Keywords: situated meaning, interaction, modality, periphrasis, Latent Class Analysis

1 Introduction

Variation in the morphosyntactic format of utterances can frequently be explained in terms of meaning differences (Bybee 2010, 165). For instance, in Spanish, the periphrases *tener que* + infinitive 'have to', *deber* 'must' + infinitive, and *deber de* 'must' + infinitive can express deontic (1) or epistemic modal meanings (2). *Tener que* + infinitive is assumed to be more likely to be used with deontic readings than the *deber* + infinitive and especially *deber de* + infinitive. The reverse is true for epistemic readings.

(1) a. *Ten-go* *que* *cant-ar.*
have-PRS.IND.1SG that sing-INF
b. *Deb-o* *cant-ar.*
must-PRS.IND.1SG sing-INF
c. *Deb-o* *de* *cant-ar.*
must-PRS.IND.1SG of sing-INF
"I have to sing."

(2) a. *Tien-e* *que* *ser* *Juan.*
have-PRS.IND.3SG that be.INF Juan
b. *Deb-e* *ser* *Juan.*
must-PRS.IND.3SG be.INF Juan

*** Corresponding author: Malte Rosemeyer,** Institute for Romance Philology, Freie Universität Berlin, Habelschwerdter Allee 45, D-14195 Berlin, Germany, e-mail: malte.rosemeyer@fu-berlin.de
ORCID: Malte Rosemeyer 0000-0001-5348-561X

c. *Deb-e* *de* *ser* *Juan.*
must-PRS.IND.3SG of be.INF Juan
"That must be Juan."

The examples in (1)–(2) also serve to illustrate that linguistic expressions are frequently polysemous. Usage-based linguistics assumes that grammatical meaning reflects the language users' experience with particular situations (Zima 2021, 46–8, Desaguilier and Monneret 2023, 31–2). As a result, the meanings of linguistic expressions are anchored in their use in different situations and contexts (Diessel 2011, 837), a fact that can be described as polysemy.

In linguistic production data, the meanings of utterances usually need to be inferred by the analyst because direct access to the speakers' intuitions is impossible. This is a notorious problem for approaches that assume that identifying the meaning of utterances is pivotal for explaining variation between elements such as *tener que* and *deber de*. Two types of solutions to this problem are frequent in variationist linguistics. First, authors may establish a typology of meanings either on the basis of previous descriptions or qualitative descriptions of the linguistic elements in their corpus. Second, authors may refrain from establishing such a functional typology and simply analyze the variation in terms of contextual predictors that are assumed to be indicative of meaning differences. As I will show in this article, both methods face specific problems that diminish their usefulness in studying morphosyntactic variation.

I suggest a third solution to the polysemy problem, which does not involve using pre-defined meaning categories nor atomize meaning in terms of contextual predictors. Rather, I propose to use Latent Class Analysis (LCA; Lazarsfeld and Henry 1969) to establish a data-driven typology of grammatical meanings, an unprecedented approach in linguistics. I describe the situated meanings of grammatical elements as latent constructs. Latent constructs are non-observable variables that can be measured in terms of indicators that represent the underlying construct (Nylund-Gibson and Choi 2018). In the case of grammatical meanings, these indicators are features of the linguistic and non-linguistic context. This proposal is in line with approaches such as Conversation Analysis and Interactional Linguistics, which assume that speakers construe meaning in interaction on the basis of strategically employed linguistic and non-linguistic resources (Heritage 2008, Couper-Kuhlen and Selting 2018, Rossi 2020), but differs in its application of quantitative methodology.

I analyze variation in the Spanish verbal periphrases with *tener* and *deber* in order to show how LCA can be used to identify unobserved grammatical meanings based on their distribution in terms of a set of contextual predictors. My findings show that meaning categorization based on LCA is more useful in accounting for variation than manual classifications and that it also identifies meanings that have not been considered relevant by the literature. I use regression analysis to demonstrate the usefulness of this analysis in studying the role of social parameters and, in particular, socioeconomic status (SES). Consequently, the results from this article show LCA to be a statistical method that allows us to study the generation of meaning in context as a social practice of the speakers.

This article is organized as follows. In Section 2, I describe the analytical problems arising in the classification of meanings in corpus data. In Section 3, I propose to view situated meaning in corpus data as a latent construct and introduce LCA as an analytical tool. Sections 4 and 5 present the results from the case study. Section 4 describes the situated meanings of *tener* + infinitive and *deber* (*de*) + infinitive in a corpus of spoken Spanish on the basis of an LCA. In Section 5, I use the resulting latent classes to analyze the variation between these periphrases. Section 6 presents the conclusions.

2 Classifying meanings in corpus data

Many cases of variation in the morphosyntactic format of utterances can be explained in terms of differences in the grammatical meanings of these utterances. To give another example from Spanish, the periphrastic future construction (3a) is more likely to express futurity than the synthetic future construction (3b), which is more likely to be used to express epistemic readings (Rosemeyer and Sansiñena 2022).

(3) a. *María* *va* *a est-ar* *aquí* a *las* *ocho.*
María be.PRS.IND.3SG to.be-INF here at the eight
"María will be here at eight."
b. *María* *no* *vino* *al* *trabajo* *hoy.*
María NEG come.PST.PFV.IND.3SG to.the work today
Est-ará *enferma.*
be-FUT.IND.3SG sick
"Mary did not come in to work today. She must be sick."

This description analyzes variation in terms of what is assumed to be the semantics of these constructions. In doing so, it uses predefined categories that in fact derive from Aristotelian philosophy (Hintikka 1973, Johnson 2004, Malink 2006, 2011). However, as pointed out by Escandell-Vidal (2010, 12), the meaning of constructions such as the synthetic future needs to be understood in terms of its contribution to realizing very diverse types of linguistic actions, such as giving an order (4a), concede a point (4b), mark an inference from a preceding clause (4c), express a rhetorical question (4d), or make an offer (4e).

(4) a. *Est-arás* *a* *las* *ocho,* *joven.*
be-FUT.IND.2SG to the eight young.man
"You will be here at eight, young man."
b. *Hará* *calor,* *pero* *segu-iré* *us-ando* *mis*
make.FUT.IND.3SG heat but follow-FUT.IND.1SG use-GER my
botas.
boots
"(You may be right to assume that) it is hot, but I will still wear my boots."
c. *Si* *todavía* *lluev-e* *en* *Managua,*
if still rain-PRS.IND.3SG in Managua
est-ará *media* *inund-ada* *la* *ciudad.*
be-FUT.IND.3SG middle flood-PTCP the city
"If it still rains in Managua, the city should be half flooded."
d. *¡¿Pero* *tú* *qué* *sab-rás* *de* *los* *videojuegos?!*
but you what know-FUT.IND.3SG of the videogames
"What would YOU know about videogames?"
e. *¿Tom-arás* *un* *café?*
take-FUT.IND.2SG a coffee
"Would you like a coffee?"

The examples in (4) illustrate a split between semantic descriptions and the concrete and systematic ways in which speakers use the synthetic future in discourse in order to generate specific meanings. From the perspective of the language users themselves, these *situated meanings* (Linell 2009) are more relevant than the semantic opposition between futurity and epistemic modality. This fact is well-known in Conversation Analysis, which describes constructions such as the synthetic future as tools for common sense methods that solve specific problems in interaction (Bergmann 1981, Waynard and Clayman 2003), and has recently also been acknowledged in corpus-linguistic research (Reuneker 2023). Deppermann (2020, 235) describes meaning in interaction as situated, social, public, and construed in context. To quote Deppermann:

> In many cases, it is not enough to retrieve meaning from the mental lexicon; the meaning [of an utterance] often has to be construed for the specific circumstances and contexts of the interaction. While this necessity is obvious when considering referential processes, other facets of meaning are highly context-specific, too, and cannot be easily derived from context-free routines. (Deppermann 2020, 235; my translation)

These considerations explain why coding of corpus data in terms of concepts such as futurity and epistemic modality is difficult. Indeed, several of the examples in (4) are ambiguous in terms of this opposition. For

instance, (4e) can be interpreted as expressing futurity ('Will you take a coffee (if I ask you)?') or epistemic modality ('Is it the case that you take a coffee?'). This ambiguity appears to be unproblematic to the speakers in interaction, which indicates that modality might not be a relevant semantic category in the type of meaning negotiated by speakers in conversation. By contrast, ambiguity is a challenge to researchers trying to code meanings in corpus data, as "inconsistencies in manual coding and inaccuracies in measurement introduce a threat to the internal validity of our research by obscuring the signal that we seek to detect in our data" (Larsson et al. 2020, 237). At least three strategies for dealing with this problem can be identified in corpus-linguistic studies.

The first approach is to establish a coding category that collects all cases for which the intuitions of the linguist are unclear. For instance, in (4), we might code (4a) as expressing futurity, (4b) and (4d) as expressing epistemic modality, and (4c) and (4e) as unclear cases. While acknowledging ambiguity in interpretation is relevant from the perspective of semantic change (Heine 2002), this type of coding introduces a category that by definition escapes interpretation. In turn, no clear interpretation of any effect of such a variable on the variation at hand is possible. Moreover, this approach does not solve the problem of subjectivity of coding procedures, since unintelligibility is in itself a (gradual) measurement. Consequently, we would expect different coders to also differ in terms of their assignment of cases to the 'unclear' category. Some variationist studies, like for instance Nuyts and Byloo (2015) and Míguez (2021), hedge this problem by coding for all possible meanings for a token at the same time. This coding procedure would seem to imply that each of the meanings is equally possible for this token, creating other analytical problems.

A second approach is to use inter-rater reliability (IRR) procedures, as detailed by Plonsky and Derrick (2016). IRR measures the degree of agreement between several coders over the same stimulus from the corpus data. While IRR thus greatly enhances confidence in the relevance of the coding, it necessarily works with predefined meaning categories which are explained in the coding scheme. As a result, the situated meanings of the stimuli in question are still inaccessible to this methodology. For instance, Reuneker (2023) tests which of three classifications of conditionals taken from the linguistic literature produces the highest IRR in a group of raters. All of these classifications are based on introspective interpretations by the linguists who established them. While Reuneker's results indeed suggest significant differences in the IRR of the three classifications, he rightly concludes that from the perspective of the speakers themselves, the main function of conditionals may be argumentative (Reuneker 2023, 415–6). This description of conditionals, which is based on the actional potential in discourse, is not systematically tested in the classifications.

A third strategy consists of not coding for any meaning, but simply analyzing morphosyntactic variation in terms of contextual features which may or may not be related to the type of meaning expressed by the relevant utterances. This approach is common in studies inspired by probabilistic grammar (Bresnan 2007, Szmrecsanyi 2013, Mazzola et al. 2022, to name but a few). For instance, Mazzola et al. (2022) analyze variation in complementation patterns in Classical Spanish in terms of contextual predictors such as coreferentiality between arguments in the main and complement clauses. The fact that these predictors significantly affect the variation in question is then taken to motivate the interpretation that certain complementation strategies indicate a higher degree of semantic and syntactic integration of the two clauses. This data-driven approach to classification has the advantage that it is principally able to detect situated meanings on the basis of these predictors. When used with explorative methods such as conditional inference trees (Tagliamonte and Baayen 2012), it can even detect contextual predictors not previously considered in the literature. However, it faces two challenges. First, the interpretation of results strongly depends on the correctness of the premises regarding the relationship between predictors and situated meanings (e.g., coreferentiality as an indicator of semantic and syntactic integration). These premises are typically not tested systematically. Second, there is no principled way of assessing how many contextual predictors are necessary to distinguish situated meanings and predict variation. However, to test a large set of predictors and/or predictor levels, equally large data sets are necessary and not always available. As a result, most analyses in the probabilistic grammar framework only test a subset of variables assumed as specifically relevant.

3 Situated meaning as a latent construct

These problems notwithstanding, the data-driven approach to classifying meanings in corpus data is promising. Both Usage-Based Linguistics and Conversation Analysis describe meaning as an emergent property in interaction and propose that interactional meanings are best described in terms of contextual properties of the utterance. I consequently assume that (a) decontextualized meanings need to be described as meaning potentials and (b) meaning is not bound to the speaker's intention but emerges in interaction.

The distinction between decontextualized and situated meanings is taken from Dialogic Syntax (Linell 2009). Linell assumes that:

> The meaning of a lexical item, that is, a word, is not a fixed set of semantic features all of which are always activated (that is, in all usage events involving the word). Instead, we seem to be faced with a 'meaning potential', which can be thought of as a structured set of semantic resources that are used in combination with contextual factors to prompt and give rise to situated meanings. It is part of meaning potential theory that potentials always, not just sometimes, interact with contextual factors. (Linell 2009, 330)

This concept is strongly reminiscent of the notion of coercion in theories of argument structure and construction grammar (Pustejovsky 1993, De Swart 1998, Michaelis 2004, Pustejovsky and Jezek 2008, Boas 2011) and has been successfully applied to the study of grammatical meanings. For instance, Gras and Sol Sansiñena (2015) analyze discourse-connective *que*-constructions in Spanish (5). They develop a typology of situated meanings of these constructions on the basis of (a) the type of contextual information that *que* refers to, (b) co-occurrence of linguistic resources such as predicate type and information structure, and (c) position in the conversational sequence. In doing so, they are able to identify a common decontextualized meaning to the use of *que* in all of the studied instances; *que* always has an indexical function pointing to a piece of information in the context.

(5)

Que	*me*	*gust-a*	*mucho*	*ir*	*al*	*cine*.
que	to.me	like-PRS.IND.3SG	a.lot	go.INF	to.the	cinema

"I love going to the cinema."

This type of description of situated meanings conforms to a central assumption of Conversation Analysis, namely that meaning is not bound to the speaker's intention (González-Lloret 2010, 61). Rather, participants construe meaning in interaction by reciprocal reference to sequential utterances. To quote Rossi (2020, 1):

> Interaction unfolds as a chain of initiating and responding actions. This chain is a source of internal evidence for the meaning of social behavior as it exposes the understandings that participants themselves give of what one another is doing.

From the theoretical perspective developed in this section of the article, situated meanings are expected to arise regularly from the combination of linguistic and contextual clues employed by the participants in the interaction. Crucially, these clues are proxies to meaning not only to the analyzing linguist but also to the speakers themselves. This means that situated meanings are social concepts that are not directly observable, but deduced from the interlocutors' behavior in interaction.

In line with this description, I contend that situated meaning can be described as a latent construct, a notion that is widely used in the context of psychology and the social sciences. Latent constructs are psychological or social concepts that cannot be observed directly. Rather, they are inferred from the way that people behave (Perron and Gillespie 2015, 93–118). Latent constructs are often used to explain complex phenomena and to make predictions about behavior. For example, Armstrong et al. (2011) describe the concept of psychological resilience as a latent construct that can be measured in terms of a person's performance regarding emotional self-awareness, expression, and self-control.

If situated meaning is a latent construct, this does not only suggest that they can be measured using proxies, but also that the sense-making function of each of these indicators should not be considered in isolation. Rather, situated meanings arise through specific combinations of these indicators. This means

that standard distributional analysis employed in probabilistic grammar and variationist linguistics can model the relationship between proxies and situated meanings only partially, and another type of statistical analysis is necessary. In order to do so, I propose to employ LCA, a statistical method that is able to model the combined effect of different proxies on latent constructs.

LCA is a statistical modeling technique used to identify unobserved groups or classes within a population based on their responses to a set of observed variables (Lazarsfeld 1950, Lazarsfeld and Henry 1969, Andersen 1982, Nylund-Gibson and Choi 2018). It was originally developed in order to describe the interrelatedness of items in sociological questionnaires (Green 1952, 71). In linguistics, LCA has recently been employed in first and second-language learning studies, where it is used to identify learner types or to generate profiles of bilingual skills of learners (Matthews and Bannard 2010, Ukoumunne et al. 2012, Halpin et al. 2021, Zhang et al. 2021, Black 2022, Gutiérrez et al. 2023).

As summarized in Nylund-Gibson and Choi (2018, 441–2), the main advantages of using LCA over similar statistical grouping methods such as factor and cluster analysis (Adli 2013), as well as the behavioral profiles approach based on co-occurrence tables (Gries and Divjak 2009, Glynn 2014) are that (a) LCA is geared towards the identification of latent classes, not 'superficial' correlations between variables in question and (b) LCA is model-based, permitting evaluation of the goodness of fit to the data.

For each individual in a population, LCA describes the probability that given a set of latent class indicators observed for this individual, the individual belongs to one of the emergent latent classes. In doing so, the model assumes local independence for each of the indicators of the latent classes. This means that "any association among the observed indicators is assumed to be entirely explained by the latent class variable, and once the latent class variable is modeled the indicators are no longer associated" (Nylund-Gibson and Choi 2018, 442). LCA yields two parameters: relative latent class size and conditional item probabilities (Nylund-Gibson and Choi 2018, 442). The latent classes identified by the model are mutually exclusive and exhaustive. In other words, each individual can only belong in one of the latent classes. The conditional item probabilities describe the probability for an individual to be included in each of the latent classes.

The results from LCA crucially hinge on the assumed set of latent classes. In order to determine the correct set of assumed latent classes, LCA employs a systematical process of model selection termed class enumeration (Nylund-Gibson and Choi 2018, 443–7). In this article, I will employ the forward-selection process recommended by Nylund-Gibson and Choi (2018, 443). Thus, the selection process starts from a one-class LCA, which serves as a comparative baseline for more complex models. The number of assumed classes is increased incrementally while observing to which degree this increase in complexity enhances the validity of the model. Model fit is measured using information criteria such as the Bayesian information criterion (BIC) and/or likelihood-based tests. Nylund-Gibson and Choi (2018, 443) recommend a conservative approach to class enumeration and avoid overspecification.

The application of LCA to corpus data in order to identify situated meanings requires treating each token of the studied construction as an individual, the situated meanings as the latent classes, and the contextual factors observed for each token as indicators of the respective latent classes. In the remainder of this article, I will illustrate this application for a case study in Spanish grammar, namely the Spanish verbal periphrases *tener que* + infinitive and *deber* + infinitive.

4 The situated meanings of Spanish modal verbal periphrases

Spanish employs many verbal periphrases to express temporal, aspectual, and modal grammatical meanings (Fernández de Castro 1990, Gómez Torrego 1999, Pusch and Wesch 2003, Martínez Gómez 2004, García Fernández 2012, Garachana Camarero 2017). There is variation between the periphrases *tener que* + infinitive and *deber* + infinitive, which can both be used not only to express deontic obligation (6) and necessity (7), but also epistemic necessity (8). *Tener que* and *deber* + infinitive furthermore compete with the quasi-synonymous variant *deber de* 'must of' + infinitive in these contexts.

(6) a. *Ten-go* *que* *devolv-er* *el* *dinero.*
have-PRS.IND.1SG that return-INF the money
b. *Deb-o* *devolv-er* *el* *dinero.*
must-PRS.IND.1SG return-INF the money
c. *Deb-o* *de* *devolv-er* *el* *dinero.*
must-PRS.IND.1SG of return-INF the money
" I have to to return the money."

(7) a. *Ten-go* *que* *estudi-ar* *para* *aprob-ar* *el* *examen.*
have-PRS.IND.1SG that study-INF to pass-INF the exam
b. *Deb-o* *estudi-ar* *para* *aprob-ar* *el* *examen.*
must-PRS.IND.1SG study-INF to pass-INF the exam
c. *Deb-o* *de* *estudi-ar* *para* *aprob-ar* *el* *examen.*
must-PRS.IND.1SG of study-INF to pass-INF the exam
"I need to study in order to pass the exam."

(8) a. *Mallorca* *tien-e* *que* *ser* *muy* *bonita*
Majorca have-PRS.IND.3SG that be.INF very beautiful
b. *Mallorca* *deb-e* *ser* *muy* *bonita.*
Majorca must-PRS.IND.3SG be.INF very beautiful
c. *Mallorca* *deb-e* *de* *ser* *muy* *bonita.*
Majorca must-PRS.IND.3SG of be.INF very beautiful
"Majorca must be beautiful."

Studies on the *tener que* - *deber* alternation indicate that *tener que* is more likely to express the deontic readings in (6)–(7), whereas *deber* is more likely to express the epistemic reading in (8) (Sirbu-Dumitrescu 1988, 141, Fernández de Castro 1999, 186, RAE 2009, Olbertz 2017, 5). *Deber de* is even more likely than *deber* to express epistemic readings (Balasch Rodríguez 2008, Blas Arroyo 2011, Eddington and Silva-Corvalán 2011). While these studies rely on introspection or manual classification of these readings in the data, other studies on modal periphrases have shown that the difference between these readings can be measured in terms of factors such as tense, predicate type, grammatical person, and diathesis (Blas Arroyo 2011, Rosemeyer 2017).

In a recent study, Thegel and Lindgren (2020) reinterpret this contrast in terms of a difference between the expression of subjective and intersubjective attitudes by the speaker. Thus, they claim that *tener que* typically expresses subjective meanings, in which the speaker alone is marked responsible for her attitude towards the proposition. In contrast, *deber* expresses intersubjective meanings, in which the attitude is attributed to a group of persons including the speaker (Thegel and Lindgren 2020, 5). They present a variationist analysis on the basis of concepts from probabilistic grammar, which shows the *tener que* - *deber* alternation to be sensitive to tense, polarity, grammatical person, and diathesis (in particular, the distinction between human and impersonal subjects). These parameters are taken to be indicative of the distinction between subjective and intersubjective meanings.

The significant overlap between the parameters used by Thegel and Lindgren (2020), on the one hand, and Blas Arroyo (2011) and Rosemeyer (2017), on the other hand, illustrates the difficulties of the probabilistic grammar approach when faced with the interpretation of the results. Indeed, from the theoretical perspective developed in this article, it is necessary to consider the interplay between the studied predictor variables in order to describe constructional variation in terms of semantic differences. These considerations allow me to formulate a first research question:

RQ1: To which degree does a treatment of situated meanings as latent constructs improve existing descriptions of the tener que - deber (de) alternation?

Thegel and Lindgren's reinterpretation of the semantic opposition between deontic and epistemic readings as a contrast between subjective and intersubjective readings moreover points to the fact that situated

meanings need to be interpreted as social actions (cf. the discussion in Section 2). The results from their article thus raise the question of whether variation between *tener que* + infinitive and *deber* + infinitive is best explained using the traditional description in terms of modal meanings or the contrast between subjective and intersubjective readings. This leads to the second research question:

RQ2: Is the variation between tener que + infinitive and deber (de) + infinitive *best explained in terms of modality or subjectivity?*

Finally, several studies have shown the relevance of social factors for the opposition between *deber* and *deber de* + infinitive. The results from the study by Blas Arroyo (2011) suggest that in comparison to *deber* + infinitive, the use of *deber de* + infinitive appears to be more likely in so-called intensifying contexts such as exclamatives or clefting (Blas Arroyo 2011, 25). This result can be seen to relate to the fact that the use of *deber de* is more likely in spontaneous, colloquial conversation (Blas Arroyo 2011, 26–7). This result was replicated using genre analysis by Rosemeyer (2017) for Renaissance and Modern Spanish. In his corpus dated between 1,500 and 2,015, the use of *deber de* is relatively more likely in comparison to *deber* in genres that instantiate a lower register (in particular, theater plays and narrative texts). Both Blas Arroyo (2011) and Rosemeyer (2017) suggest that these extralinguistic factors might actually be more relevant for the *deber* - *deber de* + infinitive alternation than modality. However, Rosemeyer's (2017) results also suggest an influence of normative language policy on the alternation. Already in the eighteenth century, prescriptive grammars of Spanish suggest that *deber de* should be used for the expression of epistemic modality and *deber* for deontic modality (Blas Arroyo 2014). It seems plausible to assume that social speaker characteristics moderate the typical situated meanings of *deber* and *deber de* in discourse, since speakers with a higher SES have frequently been found to be more likely to follow conservative prescriptivist rules in grammar (for some examples, consider Trudgill 1974, Poplack and Walker 1986, Holmes 1995). Therefore, the third research question is:

RQ3: To which extent is the variation between tener que + infinitive and deber (de) + infinitive *governed by social speaker characteristics?*

4.1 Data and coding procedures

I extracted all n = 1,233 cases of *tener que* + infinitive, n = 84 cases of *deber* + infinitive, and n = 52 cases of *deber de* + infinitive from the Peninsular Spanish section of the PRESEEA. Together, these tokens add to a final dataset of n = 1,369 cases.

PRESEEA is a dialectal corpus of semi-structured spoken sociolinguistic interviews in Spanish. The Peninsular Spanish section of the PRESEEA includes 88 interviews of about 916,000 words dated between 1988 and 2011, recorded in Alcalá de Henares, Granada, Madrid, Málaga, and Valencia. The PRESEEA is an adequate corpus for the investigation of the three research questions raised in the previous section because (a) it includes spoken, relatively informal, language and (b) it is a socially stratified corpus that represents variation in terms of age, gender, and SES (education and current job) of the speakers.

The data were coded manually in terms of (a) the difference between different modal meanings assumed in the literature and (b) ten contextual parameters that were assumed to be indicators of these meanings.

The manual coding of types of modal meanings was carried out using the standards described by Larsson et al. (2020). First, a coding scheme was established in which three types of modal meanings that *tener que* + infinitive and *deber (de)* + infinitive can express were distinguished. Deontic obligation readings imply that external circumstances force the referent of the subject to realize an action (9). Deontic necessity readings imply that the referent needs to realize an action to fulfil a desired goal (10). Finally, probability readings can be described in terms of epistemic necessity; i.e., the proposition is presented as hypothesis (11).

(9) The speaker talks about her daughter, who has a mental disability. (Madrid, 2002, MADR_M33_054)

tuve	***que***	***met-er=la***	*en*	*una*	*residencia*	*porque*
have.PST.PFV.IND.1SG	that	put-INF=her	in	a	residence	because
mi	*marido*	*est-uvo*	*enfermo*		*diez*	*años*
my	husband	be-PST.PFV.IND.3SG	sick		ten	years

"I had to put her in a residence because my husband was sick for 10 years"

(10) The speakers are discussing the difficulties of sleeping at an altitude of over 3,000 meters. (Granada, 2008, GRAN_H12_019)

Es	*que*	*el*	*cuerpo*	***se***	***tiene***	***que***	***acostumbr-ar***
be.PRS.IND.3SG	that	the	body	REFL	have.PRS.IND.3SG	that	get.used-INF

"The body needs to accustomed to it [the height]"

(11) The speaker is explaining why she never became a mother. (Granada, 2009,_GRAN_M11_040)

eso	***tiene***	***que***	***ser***	*una*	*tarea*	*muy*	*difícil*
this	have.PRS.IND.3SG	that	be.INF	a	task	very	difficult
educ-ar	*a*	*un*	*niño*				
educate-INF	to	a	child				

"It must be a very difficult task to educate a child"

From the *n* = 1,369 cases of *tener que* + infinitive and *deber (de)* + infinitive, a random selection of *n* = 100 was coded by both the author and a native speaker of Spanish in terms of this coding scheme. Inspection of the results revealed an Inter-Rater-Agreement score of 0.67. The commonly used Inter-Rater-Agreement scale by Landis and Koch (1977) considers scores between 0.61 and 0.80 to suggest 'substantial agreement' (Larsson et al. 2020, 243).

The data were also coded in terms of ten contextual predictors, which were hypothesized to indicate both the type of expressed situated meanings and the alternation between *tener que* + infinitive and *deber (de)* + infinitive on the basis of the previous literature on this topic (in particular, Blas Arroyo 2011, Rosemeyer 2017, Thegel and Lindgren 2020). These predictors describe properties of the subject (Person, SubjRef), verb (Aspect, Mood, Tense, and PredType), syntactic properties (Neg and Subordination), and pragmatic properties (Punctuation, InTurn). Table 1 summarizes these coding procedures, as well as the frequency distributions.

Before continuing to discuss the analysis itself, a brief discussion of the selection of these parameters as possible indicators of situated meanings is in order. In its original application for the analysis of sociological questionnaires, the selection of the possible indicators was relatively straightforward because it was limited by the elements in the questionnaire. When applied to corpus data, however, there does not appear to be an *a priori* limit to the linguistic indicators used in LCA. As mentioned above, in this article, I focused on indicators that previous literature believes to be proxies of the distinction between epistemic and deontic modal meaning precisely because this allows me to compare the relevance of modality for the description of the situated meanings identified by the LCA. This is not to say that my approach is exhaustive; in fact, adding more indicators may result in the identification of more situated meanings. Because an approach that includes (many) more indicators leads to increasingly complex models, it will necessitate more data and processing power. Finally, the description of the resulting latent classes will become more complex.

4.2 Analytical approach

I identified contextual profiles of the usage of *tener que* + infinitive and *deber (de)* + infinitive based on the ten variables using LCA. The analysis was carried out using the poLCA package (Linzer and Lewis 2011) in R (R Development Core Team 2024). The model estimated the posterior probability of membership in each class for each token in the dataset. Classes were interpreted on the basis of the average probability of each contextual predictor in a given class and compared to the manual coding of modal meanings of the tokens. Data and R code are publicly available on the OSF site for the project (Rosemeyer 2024).

Table 1: Coding procedures for contextual predictors

Predictor	Description	Levels	*n*
Person	Person morphology	1sg	108
		2sg	337
		3sg	426
		1pl	108
		2pl	16
		3pl	181
SubjRef	Type of subject referent	Animate	624
		Collective	212
		Impersonal	337
		Inanimate	196
Aspect	Verbal aspect	Imperfective	1,198
		Perfective	148
		Progressive	23
Mood	Verbal mood	Indicative	1,268
		Subjunctive	34
		Conditional	67
Tense	Verbal tense	Present	915
		Past	422
		Future	32
PredType	Predicate type	TelicAction	870
		AtelicAction	179
		State	320
Neg	Negation	False	1,250
		True	67
Subordination	Syntactic subordination (typology adopted from Huddleston and Pullum 2004)	MainClause	841
		ContentClause	373
		RelativeClause	155
Punctuation	Punctuation type (as annotated in the corpus)	Declarative	1,244
		Exclamative	63
		Question	62
InTurn	Position in turn	Initial	134
		Non-initial	1,235

The discrete class assignment of the selected five-class model to the data was subsequently used in a series of regression models that calculated the correlations between class membership and (a) selection of *tener que* + infinitive or *deber (de)* + infinitive, as well as (b) socioeconomic indicators for the speakers. These results are presented in Section 5.

4.3 Descriptive results

The first step of the analysis consisted of a LCA to identify the situated meanings of *tener que* + infinitive and *deber (de)* + infinitive in the dataset. I estimated models with up to ten classes without additional covariates. As proposed by Linzer and Lewis (2011, 6–7), a five-class model was chosen as the preferred model on the basis of the lowest BIC value (Table A1 in the appendix for model fit statistics). The discrete class assignment was used for the descriptive results showing the composition of classes. Figure 1 visualizes the composition of classes in terms of the ten contextual variables.

A first approach to the description of these latent classes is to analyze their relationship to the manual coding in terms of modality. Table 2 summarizes the distribution of modal meanings in terms of the latent classes. Inspection of relative frequencies suggests that the latent classes partially represent the difference between the modal readings. In particular, Class 1 appears to mostly contain tokens expressing deontic

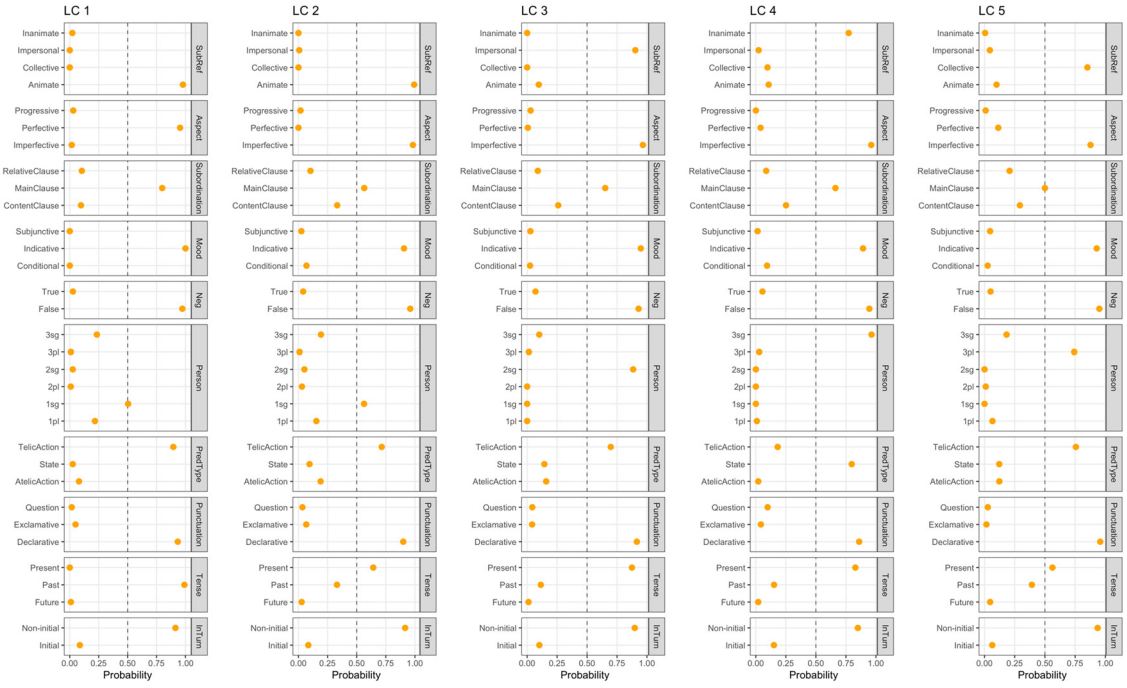

Figure 1: Posterior probability of predictor levels by latent class representing situated meanings.

obligation. While results for the other classes are less clear, deontic necessity readings are particularly frequent in Class 3 and epistemic necessity readings are relatively frequent in Class 4. Classes 2 and 5 mostly contain deontic obligation and necessity readings, with no clear specialization. The results of the LCA thus point to the necessity of considering more specialized situated meanings that can only partially be explained in terms of the modal meanings of *tener que* + infinitive and *deber (de)* + infinitive. Note also that the relative class sizes of the latent classes differ greatly. Classes 2 and 3 have the highest population share, followed by Classes 4 and 5. Class 1 has the lowest population share.

Let us now turn to the description of the latent classes in terms of situated meanings, starting with the classes that contain a high proportion of cases coded as expressing deontic obligation. While this qualitative analysis is inspired by notions from Conversation Analysis and Interactional Linguistics, it is important to mention that I cannot analyze the examples from my corpus in the same detail. Rather, I will try and derive the situated meanings on the basis of the description of the commonalities of the contextual configurations identified by the LCA.

Table 2: Distribution of modal meanings by latent classes in the dataset

Latent class	**Modal meaning (manual classification)**						**Sum *n* by latent class**
	Deontic obligation		**Deontic necessity**		**Epistemic necessity**		
	n	**%**	***n***	**%**	***n***	**%**	***n***
1	101	86	10	9	6	5	117
2	261	61	149	35	18	4	428
3	119	33	235	65	6	2	360
4	33	13	114	47	98	40	245
5	116	53	99	45	4	2	219
Sum *n* by modal meaning	630		607		132		**Total *n* 1,369**

Class 1 represents a contextual configuration in which the verb typically has first-person morphology and the subject refers to an animate referent. The verb is almost always inflected for perfective past and indicative mood and expresses a telic action. Compared with the other classes, the utterance is more likely to be coded as a declarative main clause, although there is a slightly higher proportion of exclamatives than in all other classes except Class 2. Both the probability for the utterance to be negated is lower, and for the utterance to be turn-initial is lower in Class 1 than in all other classes.

As a typical case for a Class 1 meaning, consider example (12).[1] In examples of this type, the speaker is describing a past action as motivated by exterior circumstances. The reasons for this delegation of responsibility can be due to the fact that admitting to the action frequently threatens the face of the speaker. Dropping out of school or sending your child to a residence, as in example (9), are actions that are viewed as undesirable. In other examples, the actions are simply unexpected in terms of the ongoing narration, likewise calling for attribution to external causes. While the meaning of *tener que* and *deber (de)* + infinitive in this class can easily be described as a modal meaning, namely deontic obligation, I define the situated meaning expressed by Class 1 tokens as 'delegation of responsibility'. Note that in terms of the description by Thegel and Lindgren (2020), this type of situated meaning expresses subjective modality, as the attitude in question is not shared by A and B but by A alone.

(12) A is explaining why he never finished school although he believes he would have been a good student (Málaga, 1992, MALA_H11_114)
A: *yo podía haber sido buen estudiante lo que pasa es que lo dejé*
"I could have been a good student, what happened is that I quit"
lo tuve que dejar
"I had to quit"
B: *por problemas familiares, no?*
"because of problems with the family, no?"

Class 2, the latent class with the highest population share, represents a contextual configuration that resembles Class 1 in various respects. In particular, the verb typically has first-person morphology and the subject consequently refers to an animate referent, although the posterior probability for first-person plural morphology is much lower than in Class 1. As in Class 3, the verb is more likely to express telic and sometimes atelic actions and to be inflected for past tense than in the other latent classes. However, in Class 1, it is much more likely to be inflected for present tense than in Class 2. Moreover, the verb in Class 3 configuration is typically inflected for the imperfective aspect. As in Class 1, negation is very unlikely in Class 2 configurations, the utterance is typically annotated as a declarative, and the utterance is unlikely to be turn-initial.

Given that the contextual configuration of Class 2 and Class 1 readings is similar, the readings are also expected to be similar in various respects. However, a close look does show that there are differences in these situated meanings. The *tener que* + infinitive token in (13) exemplifies Class 2 readings.

(13) Interaction near the end of the interview (Alcalá, 1998, ALCA_M23_010)
A: *¿ahora te vas?*
"Are you leaving now?"
B: *bueno no cuando termine*
"well, no, when I am finished"
A: *¡ah! porque yo te iba a decir que iba a Simago/que te dejaba en Simago que ya te pillaba muy cerquita*
"ah! because I wanted to tell you that I will pass by Simago/I could drop you off at Simago and that is very close to your place"

1 Given that the linguistic properties of the examples used to describe the latent classes are evident in the description itself and that an extended context is necessary to understand them, I will leave out the interlinear glosses for these examples. Furthermore, I will illustrate all latent classes using *tener que* + infinitive constructions to maximize comparability.

B: *¡ah! no no/luego voy andando*
"ah! no no/I will walk"
A: *porque tengo el coche ahí*
"because I have the car there"
B: ***es que me tengo que ir un poco más tarde pero***
"the thing is that I have to leave a bit later but"

The main difference to Class 1 readings appears to be that Class 2 readings typically describe prospective actions. The link between deontic readings and futurity is well-known in Spanish and other languages (Bybee et al. 1994, 184–5, Schäfer-Prieß 1999, Hernández Díaz 2017). Indeed, in an example such as (13), *tener que* + infinitive seems to be quasi-synonymous with a future construction such as *me voy a ir un poco más tarde* 'I will leave a bit later'. When used with imperfective past tense morphology, the action is prospective with regard to the deictic center. In example (14), the action encoded with *tener que* + infinitive is prospective as regards the described situation in the past.

(14) The speaker is telling about her move to a village (Madrid, 2002, MADR_M22_030)
B: *bueno estuvimos muy poquito porque ya* ***nos teníamos que volver***
"well we only stayed for a little while because we already had to return"

Both in examples (13) and (14), the action described by *tener que* + infinitive is used to motivate a previous move by the speaker. Thus, in (13), the fact that E has to leave a bit later is used to explain why B declined A's invitation to give her a lift. As in (12), *tener que* thus redresses a face-threatening act. In (14), the action encoded with *tener que* gives the reason for the fact that B and her family only stayed a little while at the village. I define the situated meaning expressed by Class 2 tokens as 'motivation of another move'.

Crucially for my argument, examples such as (13) and (14) are difficult to code in terms of the distinction between deontic obligation and necessity, as is reflected by the fact that Class 2 contains a significant share of cases coded as expressing either meaning (Table 2). Indeed, whether or not the subject referent needs to realize an action due to external or internal circumstances is irrelevant as regards the situated meaning of motivating a previous move.

The contextual configuration of Class 5 presents an even more balanced distribution in terms of the coding of modal meanings. In Class 5, the verb is typically inflected for third person plural and the subject refers to a collective referent. The verb is inflected for indicative present or imperfective past tense and expresses a telic action. The utterance is likely to be coded as a non-negated declarative main clause in non-initial turn position. Example (15) is prototypical for Class 5 configurations.

(15) The speakers are talking about television. B has just asked A whether he thinks that TV moderators should speak norm-oriented Spanish. (Málaga, 1992, MALA_H11_114)
A: *digamos que son personajes públicos, ¿no?*
"let's say that they are public persons, no?"
entonces ***tienen que dar*** *una imagen/su imagen*
"so they should give an image/their image"
B: *entonces según la norma*
"so following the norm"

In (15), A uses the *tener que* + infinitive construction to make a suggestion. This suggestion is motivated in terms of a type of intersubjective necessity; the speaker believes that it is in the best interest of society that TV moderators appear serious and thus speak norm-oriented Spanish. Previous studies describe this type of meaning as 'modal necessity' (López Izquierdo 2008, Hernández Díaz 2017, 210–1) or 'weak obligation' (Bybee et al. 1994, 186–7). In line with Bybee et al.'s description, weak obligation readings represent situations in which the fact that an obligation is not fulfilled, the consequences are not as serious as in strong obligation

cases. This is clearly the case in Class 5 readings. I define the situated meaning expressed by Class 5 tokens as a 'suggestion of a future collective action'.

Example (16), belonging to Class 5, shows that the situated meaning of suggesting a future collective action is not necessarily bound to the use of third-person morphology. The speaker A suggests that society should abolish the lottery because it is in their best interest. The first-person morphology is used to express inclusive reference, i.e., refer to both the speaker and society as a whole. Examples (15) and (16) furthermore demonstrate that suggestions of a future collective action are clearly intersubjective readings; in both cases, the speaker assumes that "the attitude in question is shared between the speaker and a larger group of people" (Thegel and Lindgren 2020, 5).

(16) The speakers are talking about the lottery. I just mentioned that many people play the lottery a lot in Spain. (Málaga, 1999, MALA_H33_714)
A: *pero ¡vamos!/que eso es peligroso ¿eh?*
"but hey this is dangerous, eh?"
eso es un vicio
"this is a vice"
y como tal vicio yo creo que ***tenemos que eliminarlo***
"and as a vice I believe that we should eliminate it"

Let us now turn to the description of Class 3, the second largest latent class, which moreover contains a relatively large population of tokens coded as expressing deontic necessity. In Class 3, the verb is typically inflected for the second-person singular and the subject expresses impersonal reference. The verb is inflected for indicative present tense and expresses a telic action. The utterance is likely to be coded as a declarative main clause in a non-initial turn position. Negation is more likely in Class 3 than in all of the other latent classes. As a typical example for this class, consider (17).

(17) A and B are talking about fishing trout. While A goes fishing a lot, B does not appear to be familiar with fishing. A just explained that trouts are very clever and hard to catch even if you wait all day with the fishing rod. (Granada, 2008, GRAN_H32_032)
A: *no vas a pescar ninguna*
"you will not catch any trout"
te tienes que ir *a otro sitio*
"you need to go to another place"
y ***tienes que ir*** *como sorprendiéndolas para pillarlas*
"and you have to make an effort to surprise them to catch them"

In Class 3, impersonal reference is expressed using the so-called *tú impersonal*, whose use is much more frequent in spoken Spanish than the use of the more formal *uno* 'one' (DeMello 2000, Guirado 2011, Kluge 2016, Posio 2017). As described by Kluge (2016, 502), the use of the second person invites speaker B to insert themselves in the position of the person fishing for trouts. This use of the second person clearly expresses an intersubjective meaning very much similar to Class 5 readings; the steps for catching a trout described in (17) are necessary and 'reasonable' (Thegel and Lindgren 2020, 13) not only for the speaker but also for the hearer and indeed, everyone.

The situated meaning instantiated by Class 3 configurations is also similar to the one in Class 5. In (17), speaker A makes a suggestion about a type of action that is required in all cases where one is trying to catch a trout. The genericity of the statement is indicated by the use of present tense morphology. Class 3 readings thus express weak obligation in the sense of Bybee et al. (1994, 186–7). In contrast to Class 5 readings, however, Class 3 readings are generic and do not express future reference. Furthermore, Class 5 readings assume that the action is useful and beneficial to the entire community, whereas Class 3 does not. I define the situated meaning expressed by Class 3 tokens as a 'suggestion of an impersonal action'.

As was described above, Class 3 tokens are more likely to be negated than all other latent classes. Example (18) illustrates that negation does not seem to affect the situated meaning represented by Class 3 configurations; A simply expresses a suggestion about an action that one does not need to realize in the generic situation of traveling in the city. Examples of this type are described as realizing an intersubjective meaning by Thegel and Lindgren (2020, 14–5).

(18) B asked A about his opinion about the metro and A is detailing the advantages of the metro. (Granada, 2008, GRAN_H32_032)
A: *pues coges tu/metro/y te vas a Albolote/y* ***no tienes que coger el coche***
"so you take your metro and you go to Albolote and you do not need to go by car"

Finally, let us discuss the type of situated meaning expressed by Class 4 configurations, which differ from the other latent classes in various respects. Recall that Class 4 unites a majority of tokens coded as expressing epistemic necessity. In Class 4 contexts, the verb typically has third-person singular morphology and the subject refers to an inanimate referent. The verb is almost always inflected for the present tense and very frequently expresses a state. Although the indicative mood is extremely frequent, Class 4 is more likely to involve conditional morphology than the other latent classes. As in the other latent classes, the utterance is frequently coded as a declarative main clause; however, in Class 4 configurations, the probability of use of question marks is higher than in the other latent classes. Likewise, the utterance to be turn-initial is higher in Class 4 than in all other classes.

Example (19) is typical for this type of contextual configuration. The speaker uses the *tener que* construction in order to postulate a hypothesis inferred from the previous context. Thus, based on the fact that the village inhabitants asked the helpers to leave, she infers that they are likely to be very proud about their achievements. This situated meaning, which I will call 'postulation of a hypothesis', consequently has to be described as a subjective meaning, since it represents an inference drawn by the speaker, which is not necessarily shared by the other participants in the interaction.

(19) The speakers are talking about A's travel to India. A narrates how she was shown around a group of self-sufficient villages in which, after having received development assistance, the inhabitants themselves administrate all aspects of daily life. (Granada, 2009, GRAN_M11_040)[2]
A: *y los mismos pueblos le estaban diciendo que se podían ir ya de allí*
"and the villages themselves told them [the helpers] that they could get out of there now"
y que podían ayudar a otras gentes
"and that they could help other people"
entonces eso para ellos ***tiene que ser un orgullo***
"so they must be very proud"
llevan munchos años
"they have been doing this for many years"

As shown by example (19), the hypotheses postulated in Class 4 contexts frequently concern mental attitudes because these attitudes are inaccessible to observation and need to be derived via inference. As a result, they typically serve evaluative functions. This explains why *tener que* and *deber* constructions in Class 4 configurations are more likely to be used at the beginning of a turn than in other latent classes. Consider example (20).

2 The adverb *muncho* 'very' is a dialectal variant of the adverb *mucho*. Use of *muncho* is frequent in Andalusian Spanish (Pato 2013).

(20) The speakers are talking about how things were much better before. They just discussed that weddings were much better in earlier times. (Alcalá de Henares, 1998, ALCA_M22_028)
B: *igual que el tema de que de lo de los hijos que antes nos íbamos de casa mucho más temprano y ahora fíjate tú*
"the same with the issue of the children, before we moved out much earlier and now mind you"
A: *sí sí/no ahora ya* ((laugh))
"yes yes/not anymore"
B: *hay gente casi con treinta años que no se van nunca ¿no?*
"there are people who are thirty and never leave, right?"
A: *horrible*
"horrible"
B: *eso es por el te*
"that's because of"
A: ***tiene que ser horrible*** *vamos*
"that must really be horrible"

A's utterance *tiene que ser horrible* is a reaction to B's utterance *hay gente casi con treinta años que no se van nunca, ¿no?,* which is designed to elicit negative evaluation by A. In terms of the terminology from Conversation Analysis, this relationship can be described in terms of preference structure (Stukenbrock 2013, 234). The utterance *tiene que ser horrible* is a preferred reaction to B's utterance and is consequently used turn-initially.

Finally, the fact that Class 4 configurations are more likely to be annotated as an information request can be explained in terms of the situated meanings expressed in these contexts. Due to the epistemic uncertainty involved in the postulation of a hypothesis, the speaker frequently asks the hearer for confirmation of this hypothesis. This correlation, which was already observed in a previous study on future tense and epistemic modality in Spanish (Rosemeyer and Sansiñena 2022), is exemplified in example (21).

(21) A has just told B that he has attended a basketry workshop. (Málaga, 1991, MALA_M21_008)
A: *para hacer co cestos de esos de*/*de mimbre*
"to make baskets of those of/of wicker"
B: *sí*
"yes"
A: *canastillos*
"small baskets"
B: *entonces eso* ***tiene que estar bonito/*** ¿no?
"so that must be nice/right?"
A: *sí*
"yes"

Table 3 summarizes the description of the five latent classes in terms of situated meanings and the parameter of subjectivity in my data. In the next sections of this article, I will show how the relevance of social parameters for this data-driven description of situated meanings and how using LCA significantly improves the description of the variation between *tener que* + infinitive and *deber (de)* + infinitive.

5 Situated meanings and the variation between *tener que* and *deber (de)* + infinitive

Having established a data-driven identification of the situated meanings of *tener que* and *deber* (*de*) + infinitive in my corpus data, I now proceed to assess the relevance of these situated meanings, as well as the SES of the

Table 3: Summary of description of latent situated meanings

Latent class	Situated meaning	Subjectivity
Class 1	Delegation of responsibility	Subjective
Class 2	Motivation of another move	Subjective
Class 3	Suggestion of an impersonal action	Intersubjective
Class 4	Postulation of a hypothesis	Subjective
Class 5	Suggestion of a future collective action	Intersubjective

speakers, for the selection of the three periphrases. This section of the article will demonstrate that by treating situated meanings as latent classes, an important problem in current variationist approaches, namely confusion between the status of variables as dependent or predictor variables, can be overcome. Moreover, it will show the data-driven description of situated meanings to be superior to the introspective classification of meanings in terms of the degree of explained variation.

5.1 Analytical approach

The analysis of the variation proposed in this article entails a major change to the standard practice of describing variation adopted in variationist analysis and probabilistic grammar. The description of situated meanings as social practices by the speakers developed in Section 3 implies that the use of a specific linguistic form is motivated by the speaker's intent to express a certain situated meaning. Consequently, it is conceptually inconsistent to treat the variation between use of *tener que* + infinitive and *deber (de)* + infinitive as the dependent variable in a multivariate analysis. Rather, the periphrases are recruited by the speakers as contextual predictors of the situated meanings, in the same way as the other predictors that were used in Section 4.3 to establish the five latent classes. I consequently used a multivariate statistical approach in which the contribution of the use of *tener que* + infinitive and *deber (de)* + infinitive to the five latent classes was measured.

Moreover, this approach entails a new perspective on how to model the role of socioeconomic indicators for the variation between *tener que* + infinitive and *deber (de)* + infinitive. The third research question, developed in Section 4, asked to which extent the variation between *tener que* + infinitive and *deber* (*de*) + infinitive is governed by social speaker characteristics. If the use of *tener que* + infinitive and *deber (de)* + infinitive is a contextual predictor of the situated meanings modeled as five latent classes, this research question needs to be answered in terms of the manner in which the use of these periphrases to express the situated meanings is moderated by socioeconomic speaker status. In other words, it is necessary to investigate SES in interaction with *tener que* + infinitive and *deber (de)* + infinitive as predictors of latent classes. This can be schematically illustrated as in (22), where '~' means 'predicted by' and ':' designates an interaction effect between the two predictor variables.

(22) Latent Class ~ Periphrasis: Socioeconomic Status

Note that in (21), SES is modeled as a single predictor, although SES is measured using several variables such as income, education, etc. Just like the situated meanings, the classification of speakers into groups representing SES was operationalized using LCA. This is standard practice in social demography (consider, for instance, Lowthian et al. 2021, Hammami et al. 2022). The results from this LCA will be presented in the next section.

5.2 Latent speaker classes

SES of the participants was measured using five variables coded in the PRESEEA corpus for the speakers, which are described in Table 4, namely level of education, speaker age, sex, income, and role in the conversation. The SpeakerAge variable was coded using the distribution quartiles.

Table 4: Predictors of latent speaker classes

Predictor	Levels	*n*
SpeakerEducation	Low	22
	Mid	24
	High	55
SpeakerAge	20–29	24
	30–41	24
	42–59	31
	60–83	19
SpeakerSex	Woman	54
	Man	44
SpeakerIncome	Low	40
	High	58
SpeakerRole	Participant	68
	Interviewer	30

Note that this information is not given to all of the speakers in the corpus. For the analysis presented in the remained of this section, I consequently had to eliminate the data from n = 28 speakers, corresponding to n = 132 tokens of *tener que* + infinitive and *deber (de)* + infinitive in our dataset, from the data. The final dataset thus consisted of n = 1,237 tokens of *tener que* + infinitive and *deber (de)* + infinitive, produced by n = 98 speakers.

In the same manner as in Section 4, I identified latent speaker profiles based on five variables using LCA. The model estimated the posterior probability of membership in each class for each speaker in the speaker dataset. I again estimated models with up to ten classes without additional covariates. Based on the lowest BIC value, a two-class model was chosen as the preferred model (Table A2 in the appendix for model fit statistics). The discrete class assignment was used for the descriptive results showing the composition of classes. Figure 2 visualizes the composition of the two speaker classes in terms of the five contextual variables.

The two identified latent classes clearly distinguish speakers in terms of SES. Class 1 speakers generally have a lower degree of education and a lower income than Class 2 speakers. Class 1 speakers are overwhelmingly participants, whereas a majority of the speakers assigned to Class 2 are interviewers. This fact probably explains the age and sex effects: Class 2 speakers are more likely to be women in the age group between 42 and 59 than Class 1 speakers, whereas, in line with the aim of the corpus to be representative in terms of social factors, the group of Class 1 speakers is more balanced.

5.3 Multivariate results

In this section, I report results from two multinomial logistic regression analyses measuring the correlation between situated meanings expressed by the three periphrases under study (*tener que* + infinitive, *deber* + infinitive, *deber de* + infinitive) and two predictor variables: (a) the formal difference between the three periphrases and (b) the posterior probability for a speaker to be assigned to the second speaker class, which represents a higher SES. I was able to model SES as a numerical variable representing the posterior probability of assignment of the second speaker class due to the fact that only two latent classes were identified by the LCA in Section 5.2. This operationalization increases the statistical resolution for the SES variable. I also tested for an interaction effect between the two predictor variables. The coding of the predictor variables is described in Table 5.

I was interested in the extent to which a description of situated meanings as latent constructs improves existing descriptions of the *tener que* - *deber* (*de*) alternation. Consequently, I calculated two multinomial regression models that only differ in terms of the dependent variable. The dependent variable in Model 1 was the manual coding of the data in terms of the distinction between deontic obligation, deontic necessity, and

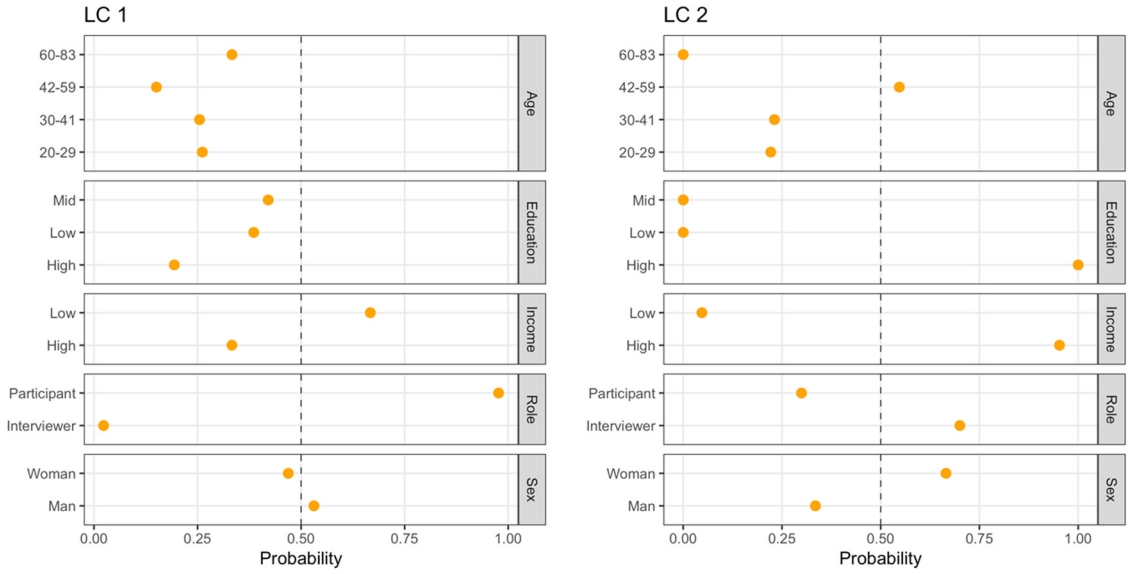

Figure 2: Posterior probability of predictor levels by Latent Speaker Class.

Table 5: Description of predictor variables used in the multinomial logistic regression models

Variable	Description	Levels
Periphrasis	Periphrasis type	*tener que* + infinitive *deber* + infinitive *deber de* + infinitive
SES	Socioeconomic status measured in terms of the probability of a speaker to belong to Latent Speaker Class 2 as opposed to Latent Speaker Class 1	Numerical variable, with a range from 0 (lowest probability) to 1 (highest probability)

epistemic necessity. The dependent variable in Model 2 was the latent classes representing situated meanings identified in Section 4.

The use of multinomial logistic regression analysis (Orme and Combs-Orme 2009, Ch. 3, Levshina 2015, 277–89) was necessary because the dependent variable was modeled in terms of a discrete assignment of either the coded modal meanings (Model 1) or one of the five latent classes representing situated meanings (Model 2) to each case in the data. The analysis was performed in R (R Development Core Team 2024), using the *nnet* package (Venables and Ripley 2002). No model selection process was conducted due to the fact that only two predictor variables were analyzed. For both models, all levels of the dependent variables were tested as reference levels. The full results for the two regression analyses are given in Tables A3 and A4.

Model 1, with the dependent variable Interpretation, found significant main effects for the predictor variables Periphrasis and SES. However, the model failed to find statistically significant differences between the use of *tener que* and *deber de*, as well as *deber* and *deber de*, in terms of the dependent variable Interpretation. In other words, the model only successfully explains the choice between *tener que* and *deber*.

As evident in Figure 3, the model predicts deontic obligation and epistemic necessity readings to be expressed using *tener que*. In contrast, *deber* is most likely to express deontic necessity readings. Note that *tener que* is actually more likely to express epistemic necessity than *deber*, an effect that reached statistical significance.

Model 1 also predicts epistemic and deontic necessity readings to be more likely the higher the probability for the speaker to belong to Latent Speaker Class 2, representing a higher SES. This effect is significantly stronger for deontic than for epistemic necessity readings. Finally, the interaction effect between Periphrasis and SES did not reach statistical significance irrespective of the reference level of the dependent variable.

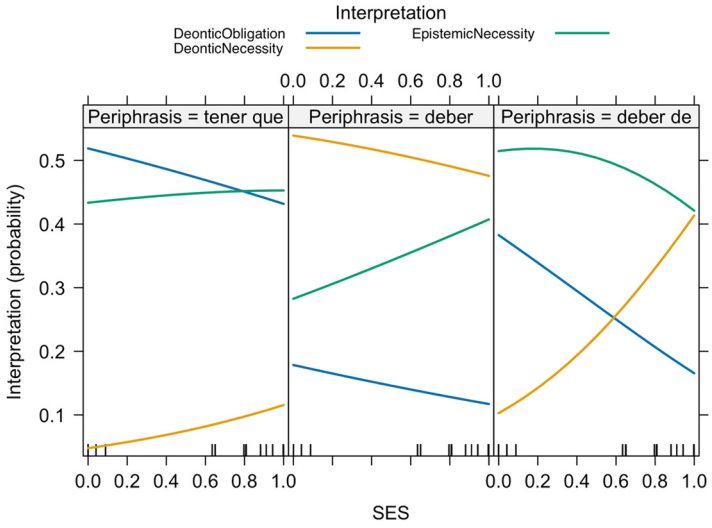

Figure 3: Marginal effects plot of the main effects of Model 1 (multinomial logistic regression with the dependent variable Interpretation).

Model 2, which measured the correlation between the latent classes measuring situated meanings, found significant main effects for the predictor variables Periphrasis and SES, as well as the interaction term. Figure 4 visualizes these results as marginal effects.

Regarding the main effects, Model 2 finds *tener que* to be particularly likely to express Class 2 ('Motivation of another move') and Class 3 readings ('Suggestion of an impersonal action'). The use of *deber* and *deber de* is significantly less likely in these contexts. While *tener que* is not particularly likely to be used to express Class 1 readings ('Delegation of responsibility'), it is significantly more likely to do so than *deber* and *deber de*. As to *deber*, the findings from Model 2 show this periphrasis to be particularly likely to express meanings classified as Class 4 ('Postulation of a hypothesis'). Finally, *deber de* is most likely used to express Class 5 meanings ('Suggestion of a future collective action'), although it is also significantly more likely than *tener que* to be used to express Class 4 meanings.

The main effect of SES is not easily apparent in Figure 4 because of the significant interaction with Periphrasis. Model 2 predicts that the higher the probability that the speaker belongs to Latent Speaker Class

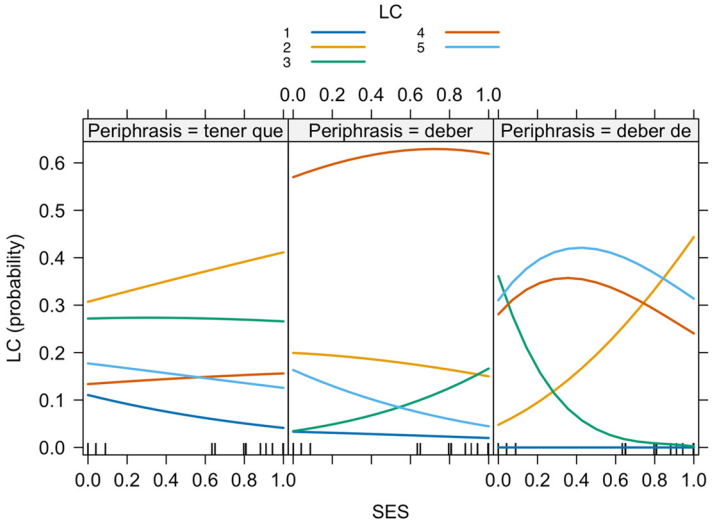

Figure 4: Marginal effects plot of Model 2 (multinomial logistic regression with the dependent variable LC).

2, which represents the highest SES, the more likely Classes 2, 3, and 4 readings are. As evident in Figure 4, this effect is moderated by the used type of periphrasis; it is significantly stronger for Class 2 readings if the chosen periphrasis is *deber de* + infinitive. In other words, the effect for Class 2 readings is mostly due to *deber de* + infinitive.

5.4 Discussion of results

In this section of the article, I discuss the findings from the multivariate analysis in terms of the three research questions developed in Section 4. RQ 1 asked to which extent the analysis of situated meaning as a latent construct improves descriptions of the *tener que - deber (de)* alternation. A direct comparison between Model 1, which uses manually coded modal meanings as the dependent variable, and Model 2, which uses latent situated meanings, in terms of the fit to the data is impossible due to the different dependent variables. However, there are at least two reasons to believe that Model 2 is superior to Model 1. First, Model 1 only successfully explains the choice between *tener que* and *deber*, whereas Model 2 finds significant effects for the choice between all three periphrases and consequently better explains the choice of these periphrases as indicators of situated meanings. Second, the significant interaction effect between Periphrasis and SES implies that Model 2 successfully models the interplay between social factors and periphrasis choice in the expression of situated meanings. The same interaction effect does not reach statistical significance for Model 1.

The comparison between Model 1 and Model 2 also contributes to answering RQ 2, which asked whether the variation between *tener que*, *deber* and *deber de* + infinitive is best explained in terms of modality or subjectivity, with mixed results. Given that model 2 found the use of *deber de* + infinitive to be most likely with intersubjective Class 5 readings, whereas the use of *deber* + infinitive is extremely likely with subjective Class 4 readings, it appears that intersubjectivity explains the use of *deber de* + infinitive to a greater degree than the use of *deber* + infinitive. While *tener que* appears to be less specialized in terms of its meaning potential, it is relatively more likely to be used to express the subjective Class 1 and 2 readings. Given that manual classification of the data in terms of modal readings turned out to not be able to explain the use of all three periphrases, the findings presented in this article can be taken to partially support Thegel and Lingren's (2020) analysis of the variation between the periphrases in terms of subjectivity.

Finally, the analysis developed in this article allows us to study the extent to which variation in the three periphrases is governed by social speaker characteristics (RQ 3). Both Models 1 and 2 found that the opposition between two latent speaker classes, which was interpreted in terms of a difference in SES, significantly influences the types of meaning expressed by the periphrases under study. Higher SES speakers are significantly more likely to use the three periphrases to suggest impersonal actions (Class 3 readings) and postulate hypotheses (Class 4) than lower SES speakers. While the exact reasons for this finding need further study, a possible explanation is that lower SES is generally associated with a lower degree of agency, understood in terms of concepts such as independence, ambition, and dominance (Shanahan 2000, Evans 2002, Abele and Wojciszke 2013, 2014). Consequently, the fact that lower SES speakers are more likely to characterize situations as independent of their actions may be the result of their real or perceived lower sense of agency.

The analysis also showed this effect of SES on expressed meanings to be moderated by periphrasis type. In particular, the higher the SES, the more likely speakers are to use *deber de* + infinitive to motivate another move (Class 2 reading). Given that Class 2 readings were described as subjective, this result does not conform to the hypothesis, proposed in Section 4, that higher SES speakers conform to linguistic norms to a greater degree. Recall that the function of Class 2 readings was described in terms of the mitigation of a face-threatening act. Examples such as (23) and (24), uttered by a speaker classified as belonging to Latent Speaker Class 2, seem to suggest that higher SES speakers have conventionalized the use of *deber de* + infinitive as this politeness resource to a greater degree than lower SES speakers. Note that both examples involve a *verbum dicendi*, which is typical for Class 2 readings.

(23) A and B are talking about A's attitude toward aging. (Granada, 2006, _GRAN_H33_015)
B: *¿qué opina sobre el conflicto generacional?*
"what is your opinion about the generational conflict?"
A: *que existió/existe y existirá*
"that existed, exists, and will exist"
B: *¿en qué sentido?*
"in which sense?"
A: *y* ***debo de decir*** *pues que/uno es joven/y/cuando uno es joven/tiene la obligación de oponerse a todo* (pause*) cree usted que yo me voy a oponer ahora/a mi edad*
"and I have to say, well, that some people are young and when someone is young they have the obligation to oppose everything (pause) do you believe that I will oppose people of my age?"

(24) B and A are talking about I's hobbies. (Granada, 2006, _GRAN_H33_015)
B: *¿qué aficiones tiene?*
"what hobbies do you have?"
A: *pues//pues tengo/vamos//mis aficiones favoritas son* (pause) *es hace un poco de deporte* ***yo debo de reconocer*** *también que que mi profesión a veces es un poco sendentario*
"well, well I have. my favorite hobbies are (pause) it's a bit of sports I also have to admit that my profession is sometimes a bit sedentary"

6 Conclusions

In this article, I have proposed a data-driven approach to the identification of the situated meanings of the three Spanish verbal periphrases *tener que*, *deber* and *deber de* + infinitive. Situated meaning was defined as a latent construct, which has a social reality but can only be measured in terms of contextual properties of the utterance in question. By using LCA, five situated meanings expressed by the periphrases under study were identified. Multivariate analysis uncovered systematic differences between the three periphrases in terms of the preferred expression of these situated meanings. The analysis also showed differences in the SES of speakers to predict the types of expressed situated meanings. The situated meanings identified using LCA proved superior to manual classification of the data in terms of modality types in various respects. In particular, the data-driven classification of situated meanings allowed us to demonstrate that the opposition between the three studied periphrases is governed less by modality types than by the difference between subjective and intersubjective meanings.

LCA allows studying the generation of meaning in context as a social practice of the speakers (Heritage 2008, Couper-Kuhlen and Selting 2018, Rossi 2020). I consequently consider LCA to be an analysis that is particularly well-suited to a combination of quantitative and qualitative descriptions of corpus data. As a result, the use of LCA can significantly advance usage-based linguistics in terms of the *desideratum* to describe language variation in terms of meaning differences.

Despite these advantages, a number of challenges can be identified that could be addressed in future studies. First, the methodology employed in this study does not solve the problem of objectively and rigorously determining which contextual proxies are relevant for the identification of latent situated meanings. As was detailed in Section 4.1, there is no *a priori* limit to the linguistic indicators used in LCA on the basis of corpus data. While the selection of the indicators in this study was inspired by the possibility of comparing the relevance of modality for the description of situated meanings, it is possible that inclusion of more and other indicators changes the results from the LCA. Second, with n = 1,369 tokens, the dataset used for this study is relatively small. It is to be expected that the number of identified situated meanings increases with larger datasets, allowing more detailed descriptions. This limitation also relates to the third challenge, which concerns the description of the relationship between the different situated meanings. In particular, the latent classes identified by LCA are not ordered hierarchically or in terms of similarity. A more detailed analysis may uncover more systematic correspondences between these classes than the proposed criterion of subjectivity.

Acknowledgements: I am grateful to Giulia Mazzola, Ferdinand von Mengden, Silvina Espíndola Moschner, Barbara Vetter, and two anonymous reviewers for comments on an earlier version of this manuscript. This article is profited greatly from discussions with the audiences at the XXXIX Encontro Nacional da Associação Portuguesa de Linguística (Covilhã, 26 October 2024) and the Congreso internacional de lenguas iberorrománicas (Helsinki, 10 January 2024).

Funding information: The author acknowledges support by the Open Access Publication Fund of Freie Universität Berlin.

Author contributions: The author confirms the sole responsibility for the conception of the study, presented results, and manuscript preparation.

Conflict of interest: The author states no conflict of interest.

Data availability statement: The datasets generated during and analysed during the current study are available in the OSF repository, https://doi.org/10.17605/OSF.IO/AP3ZK.

References

Abele, Andrea E. and Bogdan Wojciszke. 2013. "The Big Two in Social Judgment and Behavior." *Social Psychology* 44 (2): 61–2. doi: 10.1027/1864-9335/a000137.

Abele, Andrea E. and Bogdan Wojciszke. 2014. "Communal and Agentic Content in Social Cognition: A Dual Perspective Model." In *Advances in Experimental Social Psychology*, edited by James M. Olson and Mark P. Zanna, 195–255. New York: Academic Press.

Adli, Aria. 2013. "Syntactic Variation in French Wh-questions: A Quantitative Study from the Angle of Bourdieu's Sociocultural Theory." *Linguistics* 51 (3): 473–515.

Andersen, Erling B. 1982. "Latent Structure Analysis: A Survey." *Scandinavian Journal of Statistics* 9 (1): 1–12. http://www.jstor.org/stable/4615848.

Armstrong, Andrew R., Roslyn F. Galligan, and Christine R. Critchley. 2011. "Emotional Intelligence and Psychological Resilience to Negative Life Events." *Personality and Individual Differences* 51 (3): 331–6. doi: 10.1016/j.paid.2011.03.025.

Balasch Rodríguez, Sonia. 2008. "*Debe* (*de*) *ser*: evolución de la variación." In *Selected Proceedings of the 4th Workshop on Spanish Sociolinguistics*, edited by Maurice Westmoreland and Juan Antonio Thomas, 109–19. Somerville, MA: Cascadilla Proceedings Project.

Bergmann, Jörg R. 1981. "Ethnomethodologische Konversationsanalyse." In *Dialogforschung*, edited by Peter Schröder and Hugo Steger, 9–52. Düsseldorf: Schwann.

Black, Kristin E. 2022. "Variation in Linguistic Stance: A Person-centered Analysis of Student Writing." *Written Communication* 39 (4): 531–63. doi: 10.1177/07410883221107884.

Blas Arroyo, José Luis. 2011. "*Deber* (*de*) + infinitivo: ¿un caso de variación libre en español? Factores determinantes en un fenómeno de alternancia sintáctica"." *Revista de Filología Española* 91 (1): 9–42.

Blas Arroyo, José Luis. 2014. "Prescripción y praxis: una aproximación variacionista sobre la alternancia deber y deber de + infinitivo en la historia del español." *Neuphilologische Mitteilungen* 4: 647–87.

Boas, Hans C. 2011. "Coercion and Leaking Argument Structures in Construction Grammar." *Linguistics* 49 (6): 1271–1303. doi: 10.1515/ling.2011.036.

Bresnan, Joan. 2007. "Is syntactic Knowledge Probabilistic? Experiments with the English Dative Alternation." In *Roots: Linguistics in Search of Its Evidential Base*, edited by Sam Featherston and Wolfgang Sternefeld, 75–96. Berlin, New York: De Gruyter.

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

Bybee, Joan L., Revere D. Perkins, and William Pagliuca. 1994. *The Evolution of Grammar. Tense, Aspect and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Couper-Kuhlen, Elizabeth and Margret Selting. 2018. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge: Cambridge University Press.

De Swart, Henriëtte. 1998. "Aspect Shift and Coercion." *Natural Language and Linguistic Theory* 16: 347–85.

DeMello, George. 2000. "'Tú' impersonal en el habla culta." *Nueva Revista de Filología Hispánica* 48 (2): 359–372. doi: 10.24201/nrfh.v48i2.2564.

Deppermann, Arnulf. 2020. "Interaktionale Semantik." In *Semantiktheorien II. Analysen von Wort- und Satzbedeutungen im Vergleich*, edited by Jörg Hagemann and Sven Staffeldt, 235–78. Tübingen: Stauffenburg.

Desaguilier, Guillaume and Philippe Monneret. 2023. “Cognitive Linguistics and a Usage-based Approach to the Study of Semantics and Pragmatics.” In *The Handbook of Usage-Based Linguistics*, edited by Manuel Díaz-Campos and Sonia Balasch, 31–53. Blackwell: Wiley.

Diessel, Holger. 2011. “Review article of ‘Language, usage and Cognition’ by Joan Bybee.” *Language* 87: 830–44.

Eddington, David and Carmen Silva-Corvalán. 2011. “Variation in the use of *deber* and *deber de* in written and oral materials from Latin America and Spain.” *Spanish in Context* 8 (2): 257–71.

Escandell-Vidal, Victoria. 2010. “Futuro y evidencialidad.” *Anuario de Lingüística Hispánica* 26: 9–34.

Evans, Karen. 2002. “Taking control of their lives? Agency in young adult transitions in England and the New Germany.” *Journal of youth studies* 5 (3): 245–69.

Fernández de Castro, Félix. 1990. *Las perífrasis verbales en español: comportamiento sintáctico e historia de su caracterización*. Oviedo: Publicaciones del Departamento de Filología Española.

Fernández de Castro, Félix. 1999. *Las perífrasis verbales en el español actual*. Madrid: Gredos.

Garachana Camarero, Mar. 2017. “Los límites de una categoría híbrida. Las perífrasis verbales.” In *La gramática en la diacronía. La evolución de las perífrasis modales en español*, edited by Mar Garachana Camarero, 35–80. Madrid, Frankfurt: Iberoamericana, Vervuert.

García Fernández, Luis. 2012. *Las perífrasis verbales en español*. Madrid: Castalia.

Glynn, Dylan. 2014. “The many uses of *run*.” In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, edited by Dylan Glynn and Justyna A. Robinson, 117–44. Amsterdam, Philadelphia: John Benjamins.

Gómez Torrego, Leonardo. 1999. “Los verbos auxiliares. Las perífrasis verbales de infinitivo.” In *Gramática descriptiva de la lengua española*, edited by Ignacio Bosque and Violeta Demonte, 3323–88. Madrid: Espasa Calpe.

González-Lloret, Marta. 2010. “Conversation Analysis and Speech Act Performance.” In *Speech Act Performance: Theoretical, Empirical and Methodological Issues*, edited by Alicia Martínez-Flor and Esther Usó-Juan, 57–74. Amsterdam, Philadelphia: John Benjamins.

Gras, Pedro and María Sol Sansiñena. 2015. “An Interactional Account of Discourse-Connective *que*-constructions in Spanish.” *Text & Talk* 35 (4): 505–29.

Green, Bert F. 1952. “Latent Structure Analysis and its Relation to Factor Analysis.” *Journal of the American Statistical Association* 47 (257): 71–6. doi: 10.2307/2279978.

Gries, Stefan Th and Dagmar Divjak. 2009. “Behavioral Profiles: A Corpus-based Approach to Cognitive Semantic Analysis.” In *New Directions in Cognitive Linguistics*, edited by Vyvyan Evans and Stéphane Pourcel, 57–75. Amsterdam, Philadelphia: John Benjamins.

Guirado, Kristel. 2011. “Uso impersonal de *tú* y *uno* en el habla de Caracas y otras ciudades.” *Círculo de Lingüística Aplicada a la Comunicación* 47: 3–27.

Gutiérrez, Nuria, Valeria M. Rigobon, Nancy C. Marencin, Ashley A. Edwards, Laura M. Steacy, and Donald L. Compton. 2023. “Early Prediction of Reading Risk in Fourth Grade: A Combined Latent Class Analysis and Classification Tree Approach.” *Scientific Studies of Reading* 27 (1): 21–38. doi: 10.1080/10888438.2022.2121655.

Halpin, Emily, Nydia Prishker, and Gigliana Melzi. 2021. “The Bilingual Language Diversity of Latino Preschoolers: A Latent Profile Analysis.” *Language, Speech, and Hearing Services in Schools* 52 (3): 877–88. doi: 10.1044/2021_LSHSS-21-00015.

Hammami, Nour, Inese Gobina, Justė Lukoševičiūtė, Michaela Kostičová, Nelli Lyyra, Genevieve Gariepy, Kastytis Šmigelskas, Adriana Baban, Marta Malinowska-Cieślik, and Frank J. Elgar. 2022. “Socioeconomic Inequalities in Adolescent Health Complaints: A Multilevel Latent Class Analysis in 45 Countries.” *Current Psychology* 1: 1–12. doi: 10.1007/s12144-022-03038-6.

Heine, Bernd. 2002. “On the Role of Context in Grammaticalization.” In *New Reflections on Grammaticalization*, edited by Ilse Wischer and Gabriele Diewald, 83–101. Amsterdam, Philadelphia: Benjamins.

Heritage, John. 2008. “Conversation Analysis as Social Theory.” In *The New Blackwell Companion to Social Theory*, edited by Blackwell, 300–20. Oxford: Turner, Bryan S.

Hernández Díaz, Axel. 2017. “Las perífrasis con el verbo *haber* + *infinitivo*. De los valores expresados por estas formas.” In *La gramática en la diacronía. La evolución de las perífrasis verbales modales en español*, edited by Mar Garachana Camarero, 197–228. Madrid, Frankfurt: Iberoamerica, Vervuert.

Hintikka, Jaakko. 1973. *Time & Necessity. Studies in Aristotle’s Theory of Modality*. Oxford: Oxford University Press.

Holmes, Janet. 1995. “Two for/t/: flapping and glottal stops in New Zealand English.” *Te Reo* 38: 53–72.

Huddleston, Rodney and Geoffrey K. Pullum. 2004. “The Classification of Finite Subordinate Clauses.” In *An International Master of Syntax and Semantics: Papers Presented to Aimo Seppänen on the Occasion of his 75th Birthday*, edited by Gunnar Bergh, Jennifer Herriman, and Mats Mobärg, 103–16. Göteborg: Acta Universitatis Gothoburgensis.

Johnson, Fred. 2004. “Aristotle’s Modal Syllogisms.” In *Handbook of the History of Logic, Vol. 1*, edited by Dov M. Gabbay and John Woods, 247–307. Amsterdam: Elsevier.

Kluge, Bettina. 2016. “Generic Uses of the Second Person Singular – How Speakers Deal with Referential Ambiguity and Misunderstandings.” *Pragmatics* 26 (3): 501–22. doi: 10.1075/prag.26.3.07klu.

Landis, J. Richard and Gary G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1): 159–74. doi: 10.2307/2529310. http://www.jstor.org/stable/2529310.

Larsson, Tove, Magali Paquot, and Luke Plonsky. 2020. “Inter-rater Reliability in Learner Corpus Research: Insights from a Collaborative Study on Adverb Placement.” *International Journal of Learner Corpus Research* 6 (2): 237–51. doi: 10.1075/ijlcr.20001.lar.

Lazarsfeld, Paul. 1950. “The Logical and Mathematical Foundation of Latent Structure Analysis.” In *Measurement and Prediction*, edited by Samuel A. Stouffer, Louis Guttman, Edward A. Suchmann, Paul Lazarsfeld, Shirley A. Star, and John A. Clausen, 362–412. Princeton: Princeton University Press.

Lazarsfeld, Paul and Neil Henry. 1969. *Latent Structure Analysis*. New York: Houghton Mifflin.
Levshina, Natalya. 2015. *How To Do Linguistics With R*. Amsterdam, Philadelphia: John Benjamins.
Linell, Per. 2009. *Rethinking Language, Mind and World Dialogically: Interactional and Contextual Theories of Human Sense-Making*. Charlotte, NY: Information Age Publishing.
Linzer, Drew A. and Jeffrey B. Lewis. 2011. "poLCA: An R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software* 42 (10): 1–29. doi: 10.18637/jss.v042.i10.
López Izquierdo, Marta. 2008. "Las perífrasis modales de necesidad." In *Actas del VII Congreso Internacional de Historia de la Lengua Española: Merida (Yucatán), 4–8 septiembre de 2006*, vol. 1, edited by José G. Moreno de Alba and Concepción Company Company, 789–806. Madrid: Arco/Libros.
Lowthian, Emily, Nicholas Page, G. J. Melendez-Torres, Simon Murphy, Gillian Hewitt, and Graham Moore. 2021. "Using Latent Class Analysis to Explore Complex Associations between Socioeconomic Status and Adolescent Health and Well-Being." *Journal of Adolescent Health* 69 (5): 774–81. doi: 10.1016/j.jadohealth.2021.06.013.
Malink, Marko. 2006. "A Reconstruction of Aristotle's Modal Syllogistic." *History and Philosophy of Logic* 27: 95–141.
Malink, Marko. 2011. "Organon." In *Aristoteles-Handbuch: Leben – Werk – Wirkung*, edited by Christof Rapp and Klaus Corcilius, 75–84. Stuttgart: Metzler.
Martínez Gómez, Esther. 2004. "Las perífrasis verbales en español." *Revista Electrónica de Estudios Filológicos* 7. https://www.um.es/tonosdigital/znum7/estudios/kdelasperifrasis.htm.
Matthews, Danielle and Colin Bannard. 2010. "Children's Production of Unfamiliar Word Sequences is Predicted by Positional Variability and Latent Classes in a Large Sample of Child-directed Speech." *Cognitive Science* 34 (3): 465–88. doi: 10.1111/j.1551-6709.2009.01091.x.
Mazzola, Giulia, Bert Cornillie, and Malte Rosemeyer. 2022. "Asyndetic Complementation and Referential Integration in Spanish. A Diachronic Probabilistic Grammar Account." *Journal of Historical Linguistics* 12 (2): 194–240. doi: 10.1075/jhl.20031.maz.
Michaelis, Laura A. 2004. "Type Shifting in Construction Grammar: An Integrated Approach to Aspectual Coercion." *Cognitive Linguistics* 15: 1–67.
Míguez, Vítor. 2021. "The Diachrony of Galician *certamente* and *seguramente*: A Case of Grammatical Constructionalization." In *Modality and Diachronic Construction Grammar*, edited by Martin Hilpert, Bert Cappelle, and Ilse Depraetere, 123–48. Amsterdam: John Benjamins.
Nuyts, Jan and Pieter Byloo. 2015. "Competing Modals: Beyond (inter)subjectification." *Diachronica* 32 (1): 34–68. doi: 10.1075/dia.32.1.02nuy.
Nylund-Gibson, Karen and Andrew Young Choi. 2018. "Ten Frequently Asked Questions about Latent Class Analysis." *Translational Issues in Psychological Science* 4: 440–61. doi: 10.1037/tps0000176.
Olbertz, Hella. 2017. "Periphrastic Expressions of Non-epistemic Modal Necessity in Spanish: A Semantic Description." *Web Papers in Functional Discourse Grammar* 90: 1–23.
Orme, John G. and Terri Combs-Orme. 2009. *Multiple Regression with Discrete Dependent Variables*. Oxford: Oxford University Press.
Pato, Enrique. 2013. "Sobre la forma *muncho*." *ELUA: Estudios de Lingüística. Universidad de Alicante* 27: 329–42. doi: 10.14198/ELUA2013.27.12. https://revistaelua.ua.es/article/view/2013-n27-sobre-la-forma-muncho.
Perron, Brian E. and David F. Gillespie. 2015. *Key Concepts in Measurement*. Oxford: Oxford University Press.
Plonsky, Luke and Deidre J. Derrick. 2016. "A Meta-analysis of Reliability Coefficients in Second Language Research." *The Modern Language Journal* 100 (2): 538–53. doi: 10.1111/modl.12335.
Poplack, Shana and Douglas Walker. 1986. "Going through (L) in Canadian French." In *Diversity and Diachrony*, edited by David Sankoff, 173–98. Amsterdam, Philadelphia: John Benjamins.
Posio, Pekka. 2017. "Entre lo impersonal y lo individual: Estrategias de impersonalización individualizadoras en el español y portugués europeos." *Spanish in Context* 14 (2): 209–29. doi: 10.1075/sic.14.2.03pos.
Pusch, Claus and Andreas Wesch. 2003. "Verbalperiphrasen in den (ibero-)romanischen Sprachen/Perífrasis verbals en les llengües (ibero-)romàniques/Perífrasis verbales en las lenguas (ibero-)románicas." In *Beihefte zu Romanistik in Geschichte und Gegenwart*. Hamburg: Buske.
Pustejovsky, James. 1993. "Type Coercion and Lexical Selection in Semantics and the Lexicon." In *Semantics and the Lexicon*, edited by James Pustejovsky, 73–94. Dordrecht: Kluwer.
Pustejovsky, James and Elisabetta Jezek. 2008. "Semantic Coercion in Language: Beyond Distributional Analysis." *Rivista di Linguistica* 20 (1): 181–214.
R Development Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Last accessed March 5, 2024. http://www.R-project.org.
RAE. 2009. *Nueva gramática de la lengua española/2. Sintaxis II*. Madrid: Espasa Calpe.
Reuneker, Alex. 2023. "Assessing Classification Reliability of Conditionals in Discourse." *Argumentation* 37 (3): 397–418. doi: 10.1007/s10503-023-09614-9.
Rosemeyer, Malte. 2017. "La historia de las perífrasis *deber/deber de* + INF: variación, norma y géneros textuales." In *La gramática en la diacronía. La evolución de las perífrasis verbales modales en español*, edited by Mar Garachana, 147–95. Madrid, Frankfurt a.M.: Iberoamericana, Vervuert.
Rosemeyer, Malte. 2024. "Data-driven Identification of Situated Meanings in Corpus Data Using Latent Class Analysis." *Supplementary Materials*. Last accessed July 7, 2024. https://osf.io/ap3zk.

Rosemeyer, Malte and María Sol Sansiñena. 2022. “How Sentence type Influences the Interpretation of Spanish Future Constructions.” *Functions of Language* 29 (1): 116–41. doi: 10.1075/fol.00040.ros.

Rossi, Giovanni. 2020. “Conversation Analysis (CA).” In *The International Encyclopedia of Linguistic Anthropology*, edited by James Stanlaw, 1–13. Hoboken: Wiley.

Schäfer-Prieß, Barbara. 1999. “Lateinische und romanische Periphrasen mit ‘haben’ und Infinitiv: zwischen ‘Obligation’, ‘Futur’ und ‘Vermutung’.” In *Reanalyse und Grammatikalisierung in den Romanischen Sprachen*, edited by Jürgen Lang and Ingrid Neumann-Holzschuh, 97–109. Tübingen: Niemeyer.

Shanahan, Michael J. 2000. “Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective.” *Annual review of sociology* 26 (1): 667–92.

Sirbu-Dumitrescu, Domnita. 1988. “Contribución al estudio de la semántica de los verbos modales en español con ejemplos del habla de Madrid.” *Hispania* 71 (1): 139–48.

Stukenbrock, Anja. 2013. “Sprachliche Interaktion.” In *Sprachwissenschaft. Grammatik - Interaktion - Kognition*, edited by Peter Auer, 217–59. Berlin, Heidelberg: Springer.

Szmrecsanyi, Benedikt. 2013. “Diachronic Probabilistic Grammar.” *English Language and Linguistics* 19 (3): 41–68. doi: 10.17960/ell.2013.19.3.002.

Tagliamonte, Sali and Harald Baayen. 2012. “Models, Forests and Trees of York English: Was/were Variation as a Case Study for Statistical Practice.” *Language Variation and Change* 24 (2): 135–78.

Thegel, Miriam and Josefin Lindgren. 2020. “Subjective and Intersubjective Modality: A Quantitative Approach to Spanish Modal Verbs.” *Studia Neophilologica* 92 (1): 124–48. doi: 10.1080/00393274.2020.1724822.

Trudgill, Peter. 1974. “Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography.” *Language in Society* 3: 215–46.

Ukoumunne, Obioha Chukwunyere, Melissa Wake, John Carlin, Edith L. Bavin, Jarrad Lum, Jemma Skeat, Joanne W. Williams, L. Conway, E. Cini, and S. Reilly. 2012. “Profiles of Language Development in Pre-school Children: A Longitudinal Latent Class Analysis of Data from the Early Language in Victoria Study.” *Child: Care, Health and Development* 38 (3): 341–9. doi: 10.1111/j.1365-2214.2011.01234.x.

Venables, William and Brian Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer.

Waynard, Douglas W. and Steven E. Clayman. 2003. “Ethnomethodology and Conversation Analysis.” In *Handbook of Symbolic Interactionism*, edited by Larry T. Reynolds and Nancy J. Herman-Kinney, 173–202. Walnut Creek, CA: Altamira Press.

Zhang, Shuai, Esther Odilia Breuer, Matthias Grünke, and R. Malatesha Joshi. 2021. “Using Spelling Error Analyses to Examine Individual Differences in German Students from Diverse Linguistic Backgrounds: A Latent Class Approach.” *Journal of Learning Disabilities* 55 (2). doi: 10.1177/00222194211059820.

Zima, Elisabeth. 2021. *Einführung in die gebrauchsbasierte Kognitive Linguistik*. Berlin, New York: De Gruyter.

Appendix

Table A1: Model fit statistics for the LCA identifying situated meanings

Number of classes	Log likelihood	AIC	BIC
2	−8946.7	17983.5	18218.5
3	−8547.6	17231.3	17586.4
4	−8294.5	16770.9	17246.1
5	**−8164.5**	**16557.0**	**17152.3**
6	−8118.9	16511.9	17227.3
7	−8080.4	16480.8	17316.3
8	−8044.8	16455.6	17411.2
9	−8025.3	16462.6	17538.3
10	−7992.3	16442.7	17638.5

Table A2: Model fit statistics for the LCA identifying speaker classes

Number of classes	Log likelihood	AIC	BIC
2	**−376.9**	**787.9**	**831.8**
3	−359.9	771.9	839.2
4	−353.1	776.3	866.7
5	−348.4	784.8	898.5
6	−344.4	794.9	931.9
7	−342.4	808.8	969.0
8	−341.5	824.9	1008.5
9	−340.1	840.2	1047.0
10	−338.9	855.9	1085.9

Table A3: Results from Model 1, the multinomial logistic regression analysis measuring the correlation between Interpretation (manual coding of coding in terms of modality types, dependent variable), Periphrasis and SES (predictor variables)

		Reference level = Deontic obligation						**Reference level = Deontic necessity**						**Reference level = Epistemic necessity**					
Model 1		**Deontic necessity**			**Epistemic necessity**			**Deontic obligation**			**Epistemic necessity**			**Deontic obligation**			**Deontic necessity**		
Variable	**Level**	Coeff	SE	*P*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	**−2.4**	**0.2**	**<0.001**	−0.2	0.1	<0.05	**2.4**	**0.2**	**<0.001**	**2.2**	**0.2**	**<0.001**	**0.2**	**0.1**	**<0.05**	**−2.2**	**0.2**	**<0.001**
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)			(reference level)			(reference level)		
	deber	**3.5**	**0.5**	**<0.001**	0.6	0.5	>0.05	**−3.5**	**0.5**	**<0.001**	**−2.9**	**0.4**	**<0.001**	−0.6	0.5	>0.05	**2.9**	**0.4**	**<0.001**
	deber de	1.1	0.6	>0.05	0.5	0.4	>0.05	−1.1	0.6	>0.05	−0.6	0.6	>0.05	−0.5	0.4	>0.05	0.6	0.6	>0.05
SES	—	**1.1**	**0.3**	**<0.001**	0.2	0.2	>0.05	**−1.1**	**0.3**	**<0.001**	**−0.8**	**0.3**	**<0.01**	−0.2	0.2	>0.05	**0.8**	**0.3**	**<0.01**
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	−0.8	0.9	>0.05	0.6	0.9	>0.05	0.8	0.9	>0.5	1.3	0.7	>0.5	−0.6	0.9	>0.05	−1.3	0.7	>.05
	deber de: SES	1.2	1.2	>0.05	0.4	1.0	>0.05	−1.2	1.2	>0.5	−0.8	1.0	>0.5	−0.4	1.0	>0.05	0.8	1.0	>0.05

Abbreviations: Coeff = coefficient, SE = standard error, p = p value. Significant effects in bold. Model evaluation statistics: Akaike Information Criterion = 2217.76; McFadden's pseudo *R2* = 0.05.

Table A4: Results from Model 2, the multinomial logistic regression analysis measuring the correlation between latent classes representing situated meanings (dependent variable), Periphrasis and SES (predictor variables)

		Reference level = LC 1											
Model 2		**LC 2**			**LC 3**			**LC 4**			**LC 5**		
Variable	**Level**	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	**1.0**	**0.1**	<**0.001**	**0.9**	**0.1**	<**0.001**	0.2	0.1	>0.05	**0.5**	**0.1**	<**0.001**
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)		
	deber	0.8	0.9	>0.05	−0.9	1.3	>0.05	**2.7**	**0.9**	<**0.01**	1.1	1.0	>0.05
	deber de	**10.4**	**0.6**	<**0.001**	**12.6**	**0.3**	<**0.001**	**13.0**	**0.3**	<**0.001**	**12.9**	**0.3**	<**0.001**
SES	—	**1.3**	**0.4**	<**0.001**	**0.9**	**0.4**	<**0.01**	**1.1**	**0.4**	<**0.01**	0.6	0.4	>0.05
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	−1.1	1.9	>0.05	1.1	2.1	>0.05	−0.6	1.8	>0.05	−1.4	2.0	>0.05
	deber de: SES	**3.5**	**1.4**	<**0.05**	−3.5	3.6	>0.05	1.2	1.4	>0.05	1.9	1.3	>0.05
		Reference level = LC 2											
Model 2		**LC 1**			**LC 3**			**LC 4**			**LC 5**		
Variable	**Level**	Coeff	SE	*P*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	**−1.0**	**0.1**	<**0.001**	−0.1	0.1	<0.05	**−0.8**	**0.1**	<**0.001**	**−0.6**	**0.1**	<**0.001**
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)		
	deber	−0.8	0.9	>0.05	−1.6	0.9	>0.05	**1.9**	**0.4**	<**0.001**	0.4	0.6	>0.05
	deber de	−10.9	263.4	>0.05	**2.1**	**0.8**	<**0.01**	**2.6**	**0.8**	<**0.01**	**2.4**	**0.8**	<**0.01**
SES	—	**−1.3**	**0.4**	<**0.001**	−0.3	0.2	>0.05	−0.1	0.2	>0.05	**−0.6**	**0.3**	<**0.01**
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	1.1	1.9	>0.05	2.2	1.3	>0.05	0.5	0.8	>0.05	−0.4	1.2	>0.05
	deber de: SES	−2.4	20.7	>0.05	−6.9	4.5	>0.05	−2.3	1.3	>0.05	−1.6	1.2	>0.05
		Reference level = LC 3											
Model 2		**LC 1**			**LC 2**			**LC 4**			**LC 5**		
Variable	**Level**	Coeff	SE	*P*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	**−0.9**	**0.1**	<**0.001**	**0.1**	**0.1**	<**0.05**	**−0.7**	**0.1**	<**0.001**	**−0.4**	**0.1**	<**0.001**
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)		
	deber	0.9	1.3	>0.05	1.6	0.9	>0.05	**3.5**	**0.9**	<**0.001**	**1.9**	**0.9**	<**0.5**
	deber de	−12.4	202.8	>0.05	**−2.1**	**0.8**	<**0.01**	0.5	0.4	>0.05	0.3	0.4	>0.05
SES	—	**−0.9**	**0.4**	<**0.01**	0.3	0.2	>0.05	0.2	0.3	>0.05	−0.3	0.3	>0.05
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	−1.1	2.1	>0.05	−2.2	1.3	>0.05	−1.7	1.2	>0.05	−2.6	1.5	>0.05
	deber de: SES	0.7	1.1	>0.05	6.9	4.9	>0.05	4.7	4.8	>0.05	5.4	4.8	>0.05
		Reference level = LC 4											
Model 2		**LC 1**			**LC 2**			**LC 3**			**LC 5**		
Variable	**Level**	Coeff	SE	*P*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	−0.2	0.1	>0.05	**0.8**	**0.1**	<**0.001**	**0.7**	**0.1**	<**0.001**	**0.3**	**0.1**	<**0.05**
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)		
	deber	**−2.7**	**0.9**	<**0.01**	**−1.9**	**0.4**	<**0.001**	**−3.5**	**0.9**	<**0.001**	**−1.5**	**0.5**	<**0.01**
	deber de	−13.6	290.0	>0.05	**−2.6**	**0.8**	<**0.01**	−0.5	0.4	>0.05	−0.2	0.4	>0.05
SES	—	**−1.1**	**0.4**	<**0.01**	0.1	0.2	>0.05	−0.2	0.3	>0.05	−0.5	0.3	>0.05
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	0.6	1.8	>0.05	−0.5	0.8	>0.05	1.7	1.2	>0.05	−0.9	1.1	>0.05
	deber de: SES	−1.1	10.9	>0.05	2.3	1.2	>0.05	−4.7	4.8	>0.05	0.7	1.1	>0.05
		Reference level = LC 5											
Model 2		**LC 1**			**LC 2**			**LC 3**			**LC 4**		
Variable	**Level**	Coeff	SE	*P*	Coeff	SE	*p*	Coeff	SE	*p*	Coeff	SE	*p*
(Intercept)	—	**−0.5**	**0.1**	<**0.001**	**0.6**	**0.1**	<**0.001**	**0.4**	**0.1**	<**0.001**	**−0.3**	**0.1**	<**0.05**

(Continued)

Table A4: *Continued*

		Reference level = LC 5											
Model 2		**LC 1**			**LC 2**			**LC 3**			**LC 4**		
Periphrasis	*tener que*	(reference level)			(reference level)			(reference level)			(reference level)		
	deber	−1.1	1.0	>0.05	−0.4	0.6	>0.05	**−1.9**	**0.9**	**<0.05**	**1.5**	**0.5**	**<0.01**
	deber de	−13.2	265.2	>0.05	**−2.4**	**0.8**	**<0.01**	−0.3	0.4	>0.05	0.2	0.4	>0.05
SES	—	−0.6	0.4	>0.05	**0.6**	**0.3**	**<0.01**	0.3	0.3	>0.05	0.5	0.3	>0.05
Periphrasis: SES	*tener que:* SES	(reference level)			(reference level)			(reference level)			(reference level)		
	deber: SES	1.4	2.0	>0.05	0.4	1.2	>0.05	2.6	1.5	>0.05	0.9	1.1	>0.05
	deber de: SES	−1.3	14.4	>0.05	1.6	1.2	>0.05	−5.4	4.8	>0.05	−0.7	1.1	>0.05

Abbreviations: Coeff = coefficient, SE = standard error, p = p value. Significant effects in bold. Model evaluation statistics: Akaike Information Criterion = 3686.38; McFadden's pseudo $R2$ = 0.04.