Studies in Language Companion Series
234

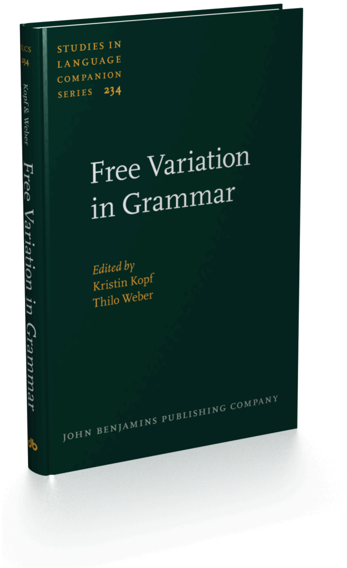
CHAPTER 11

Variation of deontic constructions in spoken Catalan

An exploratory study

Roser Giménez García | Universitat de Barcelona
| Laboratorio SQ - Lingüistas Forenses
Sheila Queralt | Laboratorio SQ - Lingüistas Forenses
F. Xavier Vila | Universitat de Barcelona

https://doi.org/10.1075/slcs.234.11gim



STUDIES IN LANGUAGE COMPANION SERIES 234

Free Variation in Grammar

Edited by
Kristin Kopf
Thilo Weber

JOHN BENJAMINS PUBLISHING COMPANY

Pages 315–349 of
Free Variation in Grammar: Empirical and theoretical approaches
Edited by Kristin Kopf and Thilo Weber
[**Studies in Language Companion Series**, 234]
2023. vi, 352 pp.

John Benjamins Publishing Company

CHAPTER 11

Variation of deontic constructions in spoken Catalan

An exploratory study

Roser Giménez García,[1,2] Sheila Queralt[2] & F. Xavier Vila[1]
[1] Universitat de Barcelona | [2] Laboratorio SQ - Lingüistas Forenses

Different areas of linguistic research have given different meanings to the notion of free variation. This paper reflects on this term and variationist linguistics. We focus on the variation between prescriptive and nonstandard deontic verbal constructions in Catalan. Through a variationist lens, we use decision trees to explore 1,060 tokens of infinitive constructions with *caldre, haver de, tenir de, tenir que* and *haver-hi que*. The discussion of results is broadened to show that variationist linguistics can dismiss but not prove the existence of free variation, a notion we argue is nevertheless relevant in linguistics, whether as a fuel for more empirical studies of language use or as a label for linguistic phenomena for which no explanation is (yet) known.

Keywords: variationist linguistics, free variation, Catalan, Spanish, deontic modality, verbal periphrasis, decision trees

1. Introduction

In variationist linguistics, it has been claimed that if a linguistic variable cannot distinguish between social groups, then it must be in free variation (Ellis 1999: 463, Labov 1966). However, since the underlying assumption of this approach (i.e. that language is systematic and rule-governed; Labov 1966) would in theory preclude free variation, the general aim of this paper is to reflect on whether the variationist approach to linguistic variation is really fit to prove the (non)existence of free variation. To do so, it specifically aims to explore the variation between deontic verbal constructions in Catalan, taking into consideration linguistic and sociolinguistic variables. The sample analysed comes from a longitudinal research project. This facilitates controlling for situational and individual factors, since the speakers and the communicative situation remain stable across time.

https://doi.org/10.1075/slcs.234.11gim

The rest of the paper is structured as follows. Section 2 deals with free variation in linguistics. Section 3 introduces the variation between five deontic constructions in contemporary Central Catalan. Section 4 outlines the methodological decisions of our study. Section 5 presents the main results, discussed in Section 6. Section 7 broadens the discussion to the implications of the study as to whether free variation can be empirically tested through variationist linguistics. Section 8 summarises our conclusions.

2. Free variation in language

Various areas of linguistic research have conceptualised free variation differently. For example, in Optimality Theory (OT), only grammatical principles are absent in variation phenomena described as instances of free variation, but these may be affected by sociolinguistic or performance variables, among others. This is because, in OT, "[t]he grammar is deterministic, in the sense that each input is mapped onto a single output – the most harmonic candidate for a constraint hierarchy" (Kager 2004: 404).

This understanding of free variation contrasts with those in psycholinguistics and sociolinguistics. Whereas in psycholinguistics this term refers to "non-systematic variation in an individual language user", in variationist linguistics it is defined as "non-systematic variation within a speech community", and the factors considered in these studies to explain variation may be social and situational as well as linguistic (Ellis 1999: 463). This paper uses the latter definition.

Even though (or perhaps precisely because) the starting point in variationist linguistics is that "[s]yntactic variation at the level of the individual speaker and the community is not chaotic and distributed randomly but is governed by social rules (Labov 1972, 1994 and many others)" (Cornips 2015: 153), free variation has fuelled many variationist studies. In this area of sociolinguistics, as Joos (2012: 185–186) puts it, *free variation:*

> [...] mean[s] merely 'not yet accounted for'. It is the technical label for whatever clearly does not need to be accounted for during the current operation in analysis; and to assume that it will never need to be accounted for in later operations would be a serious misunderstanding.

In this sense, it is only intended as a temporary tag attached to certain linguistic phenomena while scholars seek explanations for them. This is the meaning that Stokes (2011: 129) seems to give to free variation when summarising Espuny's (1998: 284) finding that a speaker changes from the Catalan infinitive periphrasis *haver de* to *tenir que* for no apparent contextual reason: "There seems to be no reason for the shift from one to the other, perhaps indicating that there is free variation between the two forms."

This use of free variation has been crucial in variationist linguistics, as it signals issues that need further empirical attention. In our case, it was partly this suggestion that two deontic periphrases might display free variation in Catalan, alongside the frequent (but superficial) allusions to speakers' use of nonstandard variants (*tenir que, haver-hi que*) in the descriptive literature, that led us to employ this topic as a tool to reflect on variationist linguistics and free variation.

Free variation has been used in other areas of linguistics. Kiesling (2011: 8) graphically describes its use in structuralist and some generativist studies on phonology as a "dustbin" for phonemes whose phonetic characters could not be categorically predicted. The categorisation of phenomena for which no explanation is known constitutes a very important first step towards the construction of knowledge. This meaning of free variation has allowed structuralists and generativists to focus on areas of language which can be used to develop their programs.

Thus, in variationist linguistics, the concept of free variation is helpful in distinguishing linguistic phenomena that can and cannot be explained by (socio)linguistic constraints. Complementarily, its use in formalist approaches is relevant for the refinement of models of grammar. Nevertheless, neither perspective makes it possible to demonstrate the existence of free variation in language. Variationism might rule out free variation when a dataset correlates with independent variables, but it can hardly argue for the existence of free variation when it assumes that language is governed by rules that can be discovered by observing empirical evidence. Simultaneously, traditional models of grammar have dismissed linguistic variation and use as something other than their object of study and have focused instead on decontextualised structures (Chomsky 1965: 3).

However, the need for complementary approaches to converge, i.e. for linguistic use to inform theoretical models, has been repeatedly advocated for (e.g. Adger and Trousdale 2007: 274; Adli, García García and Kaufmann 2015: 14–15; Seiler 2015: 259–260). Several proposals on how to account for variation within Minimalism and other theoretical enterprises have emerged (e.g. Adger 2016; Baechler and Pröll 2019; Bader 2020). In what follows, variation between verbal constructions in Catalan is used to reflect on variationist linguistics and free variation, a notion shared by numerous approaches to linguistics.

3. Deontic verbal constructions in Catalan

Central Catalan varieties have various linguistic mechanisms that express deontic modality through verb forms. Some, like *haver de* + infinitive (from now on, *haver de*), are referred to as verbal periphrases in prescriptive grammars, whereas *caldre* + infinitive (here onwards, *caldre*) is termed a non-periphrastic construction (IEC 2016: 951). We therefore use the cover term *verbal constructions* for all these variants or allostructions (i.e. "(truth-)semantically equivalent but formally distinct manifestations of a more abstractly represented construction"; Cappelle 2009: 187).

The inflected verb forms in these constructions are all followed by an infinitive – *deure* + infinitive, *caldre* + infinitive, *haver de* + infinitive and *tenir de* + infinitive (from here on, *tenir de*) – and may be used interchangeably in many linguistic contexts, although some syntactic differences exist between them (see Section 3.1). Nevertheless, *deure* + infinitive is nowadays rarely used to express deontic modality in most Catalan varieties (Cabanes Fitor 1996; IEC 2016: 951): in informal use, speakers choose (in principle) among five semantically equivalent forms, the latter three constructions above and another two infinitive periphrases, which originate from the centuries-long linguistic contact between Catalan and Spanish: *tenir que* + infinitive and *haver-hi que* + infinitive (henceforth, *tenir que* and *haver-hi que*).

Although the latter are described as "unacceptable" by the prescriptive Catalan grammar (IEC 2016: 951), "[t]here is abundant anecdotal evidence" of their use, especially of *tenir que* (Stokes 2015: 461). Such evidence includes Rigau's (1998: 80) brief mention that *haver-hi que* would be gaining ground in contexts where *caldre* was previously the leading variant in Southern dialects and Sinner's (2008: 534) use of *tenir que* as an example of elements that are "quite common" in informal varieties but considered "alien to the language" in prescriptive works (cf. i.a. Hualde 1992: 325; Cabanes Fitor 1996; Martínez Díaz 2002: 87).

Due to its history and the prolonged language contact between Catalan and Spanish, currently all Catalan speakers are, at least, bilingual in these Romance languages, with the only exception being Catalan speakers in Northern Catalonia (southern France) and Alghero (Sardinia, Italy) or in diaspora (e.g. Martines 2020: 315–316, Galindo, De Rosselló and Bernat 2021). This is vital to better understand the variation between the deontic verbal constructions, since both languages have similar systems. Table 1 summarises the main possible deontic constructions in each variety.[1] Prescriptive Catalan grammar describes *tenir de*

[1] Catalan and Spanish have other deontic constructions, such as the complex verbal constructions *fer falta* (Sp. *hacer falta*) or *ser necessari* (Sp. *ser necesario / ser preciso*), which mean 'to be necessary'. Practical constraints precluded the inclusion of *fer falta* in the analysis, despite

and *deure* as only found in old varieties. It claims that the former is nowadays only used in informal settings and the latter only in Valencian varieties or somewhat fixed expressions like *He fet el que devia* 'I did what I must'. In constrast, *haver de* and *caldre* are described as acceptable forms regardless of the setting (IEC 2016: 951). Variants in parentheses in Table 1 are described in prescriptive works as only used in certain varieties. The forms with an asterisk are marked in most modern vernacular varieties and settings (although *haber de* is found in formal written Spanish), as they are seen as less acceptable than their counterparts, generally perceived as genuine or not originating from the contact between Catalan and Spanish. Contrary to folk belief, however, some did not result from this sociolinguistic situation but either from the languages' own diachronic evolution or from contact with other languages (cf. Conde Noguerol 2016 on *caler* in Spanish or Sentí 2015 on *deure* in Catalan).

Table 1. Catalan and Spanish systems of deontic verbal constructions

	Standard Catalan	**Vernacular Catalan**	**Standard Spanish**	**Vernacular Spanish**
Used in all grammatical persons	*haver de* + infinitive (*tenir de* + infinitive)	*haver de* + infinitive *tenir de* + infinitive **tenir que* + infinitive **deure* + infinitive	*tener que* + infinitive *deber* + infinitive (*haber de* + infinitive)	*tener que* + infinitive *deber* + infinitive **haber de* + infinitive
Used in third person	*caldre* + infinitive	*caldre* + infinitive **haver-hi que* + infinitive	*haber que* + infinitive (*caler* + infinitive)	*haber que* + infinitive **caler que* + infinitive

Source: adapted from Stokes (2015: 447).

The use of these constructions is influenced by linguistic constraints and possibly also by sociolinguistic factors, as suggested by previous publications. The elements considered in our study of this phenomenon of variation are discussed in the remainder of this section.

its presence in the sample (54 tokens). However, we intend to incorporate it in future studies of this phenomenon, since it has become widely used in many Catalan varieties and its syntactic behaviour somewhat resembles that of *caldre* (cf. Rigau 1998: 68).

3.1 Catalan deontic constructions and linguistic factors

To understand how speakers use these deontic constructions, we must consider linguistic constraints that affect their acceptability. These and their operationalisation in our study are outlined below.

Syntactic and semantic differences exist between these modal constructions. Firstly, regarding grammatical person, *caldre* and *haver-hi que* are (almost)[2] exclusively used in third person, whereas *haver de, tenir de* and *tenir que* are used in any grammatical person. *Haver-hi que* is a (relatively recent) calque of the Spanish *haber que*, described by prescriptive grammar as a third-person verb (RAE 2010: 2148). As for *caldre*, Catalan prescriptive grammar describes it as a defective verb (IEC 2016: 272). Previous studies have considered grammatical person as a linguistic variable that may interact with their use.

According to Mier (1986: 47), there would be a syntactic restriction in the use of *tenir que* and *tenir de* in that *haver de* is used more frequently in a reflexive construction than the others ("[t]his is true whether the reflexive is the impersonal *s'ha de* or a personal reflexive such as *t'has de*", Mier 1982: 31). In Stokes (2015: 461–462), a more recent study on *haver de* and *tenir que* on Twitter, the reflexive third person pronoun *se* also favours the use of *haver de*, whereas *tenir que* in this context appears "severely limited". Stokes (2015) also finds *tenir que* mostly in first person singular and plural (60.5% of cases), whereas *haver de* appears most frequently in third person (50.8% of instances). Thus, grammatical person is one of the linguistic variables in our study.

Additionally, Rigau (1999: 194), following the traditional Catalan grammarian Anfós Par (1923), argues that *caldre* is a "relativized impersonality" construction:

> Therefore the verb *caldre* behaves as an impersonal verb. But its impersonal character may be relativized by the presence of an argument indicating the person implied or interested in the situation, that is, the argument that shows dative case.

Rigau (1999) states that this dative argument functions as the subject of the sentence, a claim we do not share. Interestingly, however, note that *caldre* may take an experiencer before the inflected form regardless of the infinitive it accompanies when it is not followed by an inflective subjunctive clause. This is not possible with the other constructions because of their syntactic structure, in which the subject is personal and specified for gender and number in the inflected verb. For example, speakers of Central Catalan may utter *{Cal / Hi ha que / S'ha de /*

[2] See Rigau (2005a: 246) on the change in progress by which some speakers would also accept other uses of *caldre* + infinitive. However, no uses of *caldre* + infinitive other than third person were found in our data.

Es té de / Es té que} tenir molta paciència 'One needs to have a lot of patience' as well as *Ens*$_{\text{[exp]}}$ *cal tenir molta paciència* 'We need to have a lot of patience' but not **Ens*$_{\textit{[exp]}}$ *{hi ha*$_{\textit{[3rd pers. sg.]}}$ *que / hem*$_{\textit{[3rd pers. pl.]}}$ *de / tenim*$_{\textit{[3rd pers. pl.]}}$ *de / tenim*$_{\textit{[3rd pers. pl.]}}$ *que} tenir molta paciència* (cf. Rigau 1999: 344).

Thus, while all the constructions in this study express deontic modality in the same linguistic contexts, the sentences in which they appear may need to undergo a few adjustments to meet the requirements of the verb form regarding grammatical person and argument structure. For instance, *haber que* (and *haver-hi que*) generally selects third person clitic pronouns[3] in sentences with verbs with reflexive pronouns, such as *Hay que ducharse* (Cat. *{Cal / Hi ha que} dutxar-se* 'One needs to shower'). In contrast, the deontic periphrases with *haver* and *tenir* cannot appear with such *se*-forms, since a sentence cannot have two third-person *es* pronouns nor the pronoun *es* performing two functions (e.g. *{Hem de / Tenim de / Tenim que} dutxar-nos* vs *{*S'ha de / *Es té de / *Es té que} dutxar-se*; IEC 2016: 895–896). Yet, similarly to the impersonal constructions with *caldre* and *haver-hi que*, these three periphrases can be used with no definite subject with transitive verbs (e.g. *{Cal / Hi ha que / S'ha de / Es té de / Es té que} comprar pa* 'Someone has to buy bread').

Secondly, as seen thus far, all the constructions can be followed by an infinitive clause. However, because of its semantic load (Rigau 2005b: 782), *caldre* can also be followed by a noun phrase (*Cal aigua* 'Water is necessary'), a determiner phrase (*Cal una gàbia* 'A cage is necessary') or an inflected clause, in which case the subjunctive is selected: *Cal que tingui*$_{\textit{[subjunctive]}}$ *aigua* 'It needs to have water'. These are not possible with the other variants due to the conjunction (*que*) or preposition (*de*), which select an infinitive clause (e.g. **Té de {aigua / una gàbia / que tingui* $_{\text{[subjunctive]}}$ *aigua}* '*(S)he has to {water / a cage / have$_{\text{[subjunctive]}}$ water' vs *Té de tenir*$_{\textit{[infinitive]}}$ *una gàbia* '(S)he has to have$_{\text{[infinitive]}}$ a cage'). Thus, we consider only instances of *caldre* followed by an infinitive.

Also because of its semantics, *caldre* allows for the infinitive form to be omitted (e.g. *No cal* '[That] is not necessary'), especially when the context facilitates retrieving the information conveyed by the infinitive. In these cases, the infinitive can be elided in some Spanish deontic constructions (*¿Puede salir ya? Sí, puede* 'May (s)he come out yet? Yes, (s)he may'; Garachana Camarero 2017: 44). With *haber de* and *tener que*, this omission is, on paper, not possible (compare *Comprarlo debería, pero no quiero* 'I should buy it, but I don't want to' and **Comprarlo tendría que, pero no quiero* 'I'd need to buy it, but I don't want to'). Nevertheless, a

[3] But cf. RAE (2010: 2148–2149) for its use with first person plural pronouns due to semantic transfer from *tener que*.

specific pragmatic use of these constructions facilitates the elision of the infinitive, namely, a metalinguistic meaning, as in *No quiero comprarlo, pero {tengo que / he de}* 'I don't want to buy it, but I have to' (Krivochen 2020: 15).

According to the Catalan prescriptive grammar, this is possible with all the Catalan deontic periphrases except for *haver de*, again, because of the preposition *de* (IEC 2016: 52); yet a preliminary exploration of the sample reveals that *haver de* appears without an infinitive where it is retrievable from the context and is given the metalinguistic meaning presented above. For instance, in our data, the interviewer may ask whether the prospective owner of a cat needs to take it out for walks: *{Ha de / Cal} treure'l a passejar?* 'Does it need to be walked?' In this context, the informant may not use the infinitive (e.g. *Si vol, sí, però no {ha de / cal}* 'If (s)he wants to, sure, but (s)he doesn't need to'). Thus, since different allostructions are possible, instances where the infinitive is dropped are considered in the study.

Furthermore, *haver-hi que* is semantically different from the other constructions because it requires a participant (a noun phrase) that refers to an entity capable of displaying intentionality (RAE 2010: 2148). Therefore, for instance, it is possible to express the need for more rain in a region with the other allostructions (*Aquí cal que plogui més* or *Aquí {ha de / té de / té que} ploure més* 'It needs to rain more here') but not with the impersonal form *haver-hi que* (**Aquí hi ha que ploure més*, Sp. **Aquí hay que llover más*).

Another possible linguistic constrain is sentence polarity. Rigau (2005a: 256) claims that *caldre* is mainly used in negative sentences. A preliminary analysis of the sample seems to support this observation. Table 2 shows that 76.32% of tokens of *caldre* appear in negative sentences. Simultaneously, it seems worth considering how polarity interacts with the other constructions, since *caldre* merely accounts for 29 of the 118 instances (24.58%) in which the variants are used in negative sentences.

Lastly, there seems to be a restriction regarding the verb tenses in which these constructions are used. In a study on *haber de, tener de* and *tener que*, the variable "verb tense and mood" is statistically significant (Blas Arroyo 2015). Furthermore, these deontic constructions have an element of existentiality to their meaning. They express that *there is* a need or that the obligation *exists* for someone to do something, rather than strictly that someone needs or has the obligation to perform an action (cf. Rigau 1999: 326 on *caldre*). Therefore, the preferred verb tense for these constructions might be the present indicative (*tinc de, has de, cal*, etc.). However, this variable was not analysed due to time constraints.

Linguistic factors like, at least, the ones just outlined should be considered in studies of this phenomenon of variation. However, few steps have been taken in this direction until now. It has even been suggested that *tenir que* and *haver de* are in free variation (Stokes 2011: 129). Our study includes the linguistic vari-

Table 2. Variants in the study by sentence polarity

Variant	Sentence polarity	Total
caldre + infinitive	Affirmative	9/38 (23.68%)
	Negative	29/38 (76.32%)
haver de + infinitive	Affirmative	625/689 (90.71%)
	Negative	59/689 (8.56%)
	Neutral/question	5/689 (0.73%)
tenir de + infinitive	Affirmative	35/39 (89.74%)
	Negative	2/39 (5.13%)
	Neutral/question	2/39 (5.13%)
tenir que + infinitive	Affirmative	262/293 (89.42%)
	Negative	28/293 (9.56%)
	Neutral/question	3/293 (1.02%)
haver-hi que + infinitive	Affirmative	1/1 (100%)

ables grammatical person, sentence polarity, markedness (regarding the historical development of the constructions) and priming, i.e. uses of *haver de* by the informants may be influenced (primed) by the interviewer's preceding conversational turns.

3.2 Sociolinguistic factors and variation in Catalan

This study also considers social factors, which, as outlined, may equally interact with the phenomenon of variation we discuss to probe the suitability of variationist linguistics to attest or dismiss free variation. Variationist linguistics posits that languages interact with social variables that are relevant in their speech communities. Sociolinguistic factors that have been shown to play a role in explaining language choices in contemporary Catalan include speakers' first language(s) (Gonzàlez et al. 2014: 64; Flors 2015: 36), gender (Pujolar 2001), class (DGPL 2015), linguistic attitudes (Ubalde 2013; Bretxa 2019; Martínez Díaz 2019) or individual preferences (Flors Mas and Vila i Moreno 2014).

These and other extralinguistic factors also correlate with speakers' use of linguistic variables in other languages (e.g. Ball 2010; Chambers and Schilling 2013). Additionally, sociolinguistic factors may interact with one another, as Comajoan (1998: 87) argues regarding the effect of the social environment surrounding speakers and their attitudes towards varieties:

> In sum, intergenerational transmission depends greatly on the demo-linguistic characteristics of the languages in contact and on the representations (as evidenced by their attitudes) that the speakers ascribe to the languages. Both factors are intertwined, given that major exposure to specific social networks may affect the individual's attitudes and behavior regarding what language to use.

Scholars have thus identified extralinguistic variables that influence the choices of Catalan speakers. However, studies considering the relationship between Catalan deontic verbal constructions available to speakers and social factors (such as age, linguistic attitudes or first language(s)) are still scarce.

To the best of our knowledge, only a few studies have dealt with this issue in some detail. Mier (1982, 1986)[4] interviews 43 speakers in Barcelona and examines five phonetic and morphologic and five social variables: age, sex, occupational level, use of Spanish in childhood and declared written skills in Catalan. One of the morphologic variables is the distribution of *haver de, tenir que and tenir de*, which are found to correlate with all the social variables.

More specifically, regarding speakers' age, *tenir que* is "increasingly used over time" (Mier 1986: 47). This construction is frequently used by most younger speakers in the sample (born between 1956 and 1960) and usually not perceived as a Spanish interference. The author concludes that its use "does not seem to be stigmatized" but "could become so in the future" due to the social changes and an increased interest in Catalan that began after the reinstatement of Spanish democracy, a few years prior to the study (Mier 1986: 56–57). *Tenir de* is only used by six speakers, all but one of whom were 50 or older at the time. Results lead Mier to suggest that this construction "is an old form that is rapidly becoming obsolete" (1986: 47). Interestingly, none of the speakers who use *tenir de* also produce *tenir que* (Mier 1986: 47).

Alongside grammatical person (Section 3.1), Stokes (2015) identifies a statistically significant correlation between *haver de* and *tenir que* and diatopic factors. This author gathered a collection of 9,558 tokens of these variants over two weeks in 2013, covering eight cities across Catalonia, Valencia and the Balearic Islands.

[4] This study predates the publication of the normative Catalan grammar currently in force, so Mier relied upon a historical grammar published in 1952 by a renowned philologist, Francesc de Borja Moll, and the prescriptive grammar by Pompeu Fabra (1974). The former described *haver de* + INF as "the most normal and the only [deontic periphrasis] in modern literary language", *tenir de* + INF as "normal in Valencian but not allowed in written language for it is an interference from Castilian" (Moll 1952: 336–337; Mier 1986: 46). The latter states that using *tenir de* instead of *haver de* is not advisable, whereas using *tenir que* in its place is "absolutely inadmissible" (Fabra 1974: 88).

Tenir que amounts to 11% of the sample, and *haver de* is clearly preferred in all cities except those in Valencia.

Some of the limitations of his study, as described by Stokes (2015: 463–464), are that, first, only two constructions are analysed, which contrasts with the higher number of possibilities available to speakers; second, the methods might not suit the data; third, only present indicative forms are considered; and, lastly, more linguistic variables may influence this phenomenon of variation. These limitations were all considered in the design of our study (see Section 4).

Given these previous publications, age and sex are included as sociolinguistic variables in our study. Additionally, since our sample consists of students in secondary education, we do not consider occupational level or written skills. However, the data come from Manlleu and Mataró, two cities with different demolinguistic landscapes (see Section 6), so we include this variable in the study. Furthermore, language(s) of identification and an index measuring speakers' exposure to Catalan in the media and cultural products[5] are analysed as indicators of speakers' attitudes towards Catalan. Likewise, speakers' social environment is reflected by an index measuring their use of Catalan within their social network.[6]

[5] This index was determined from answers to four questions in a sociolingusitic questionnaire. Informants listed the three television programs, musical artists, books and webpages they had been recently most exposed to and identified the languages used in each ('only or mostly Catalan', 'only or mostly Spanish', 'Spanish and Catalan equally' or 'a different language'). To obtain a value representing their exposure to Catalan in the media and cultural products, each item (television program, music, book or webpage) was scored between 0 (no use of Catalan) and 1 (exclusive use of Catalan). Then, the percentages of Catalan, Spanish and other languages in each participant's answers were calculated. For example, informants who only consumed culture in Catalan scored 1 for their exposure to Catalan and 0 for their exposure to Spanish and other languages, while those equally exposed to Catalan and Spanish scored 0.5 for their exposure to Catalan and to Spanish but 0 for other languages. Thus, a score was obtained for each participant's exposure to different languages in cultural activities so that the sum of the three percentages (exposure to Catalan, Spanish and other languages) corresponds to an informant's overall use of languages in this sphere and, therefore, always equals 100%.

[6] In the questionnaire, participants were also asked to list the twenty individuals with whom they interacted most in a week (i.e. their social network, in dyads), the relationship with them, the frequency with which they interacted and the language they used ('Catalan', 'Catalan = Spanish', 'Spanish' or 'Other (specify)'). These answers were transformed into three language use indexes (LUI), one for Catalan, one for Spanish and a third for other languages. Our study uses Catalan LUI. To obtain it, each informant was scored with 100 for dyads exclusively in Catalan, 50 for dyads using Catalan alongside another language, 33 for Catalan in combination with two other languages and 0 for answers not including Catalan. The sum of an informant's scores was divided by the number of peers in the list, which produced an index of use of Catalan between 0 (no use) and 100 (exclusive use of Catalan with all peers). For more details, see Vila, Ubalde, Bretxa & Comajoan (2020).

4. Methodology

This study explores the variation among the five deontic constructions discussed so far by analysing a sample of 1,060 tokens produced by 64 informants of the RESOL project, one of the largest longitudinal studies of Catalan to date (for details, see Bretxa and Vila 2012; Bretxa and Vila 2014; Bretxa et al. 2016). The informants in this project were recorded in the cities of Manlleu and Mataró twice in four years – aged approximately 12 years old at T1 and 16 years old at T2 – during a role-play task in which they played a pet store clerk with the aim of convincing their customer (the interviewer) to buy an animal in their store. For the purpose of this study, tokens obtained at different times are treated as belonging to different individuals, since our aim is not to compare the performance of each speaker across time but to explain variation between equivalent constructions. Table 3 summarises the distribution of the allostructions in the data. Tokens were searched for manually and included in the sample regardless of verb tense.

Table 3. Distribution of deontic verbal constructions in teenage Central Catalan speech

Variant	**N (percentage)**
caldre	38 (3.6%)
haver de	689 (65%)
haver que	1 (0.1%)
tenir de	39 (3.7%)
tenir que	293 (27.6%)
Total	**1,060 (100%)**

The independent linguistic and sociolinguistic variables considered are shown in Tables 4 and 5. The data for the last three sociolinguistic variables in Table 5 were obtained from the informants' responses to the comprehensive sociolinguistic questionnaire used in the RESOL research project (Bretxa 2014: 130–133). Regarding language of identification, most informants identified with Catalan (coded as '1'), Spanish ('2') or both ('3'), with one exception: an informant who answered 'Portuguese' at T1 (coded as '77') but, incidentally, answered 'Catalan and Spanish' at T2.

Table 4. Linguistic variables in the study

Dependent variable			**Linguistic variables**			
A	**B**	**C**	**D**	**E**	**F**	**G**
Token	**Informant code**	**Deontic construction**	**Grammatical person**	**Priming**	**Sentence polarity**	**Markedness**
no cal tenir-lo fora\ ‘you don’t need to keep him outdoors\’	M0054SCI	1	3_sg	0	0	1

C: 1 = ‘*caldre*, 2 = ‘*haver de*’, 3 = ‘*tenir que*’, 4 = ‘*tenir que*’, 5 = ‘*haver-hi que*’
D: 1 = ‘1st.pers.sing.’, 2 = ‘2nd.pers.sing.’, 3 = ‘3rd.pers.sing.’, 4 = ‘1st.pers.pl.’, 5 = ‘2nd.pers.pl.’, 6 = ‘3rd.pers.pl.’
E: 0 = ‘no priming’, 1 = ‘priming’
F: 0 = ‘negative’, 1 = ‘affirmative’, 2 = ‘neutral/question’
G: 1 = ‘Forms perceived as genuine in Catalan (*haver de, tenir de, caldre*)’, 2 = ‘Forms perceived as contact-originated in Catalan (*tenir que, haver-hi que*)’

Table 5. Sociolinguistic variables in the study

Dependent variable			**Sociolinguistic variables**					
A	**B**	**C**	**H**	**I**	**J**	**K**	**L**	**M**
Token	**Informant code**	**Deontic construction**	**Age**	**Sex**	**City**	**Language of identification**	**Index of Catalan in the social network**	**Index of Catalan in culture and media**
no cal tenir-lo fora\ ‘you don’t need to keep him outdoors\’	M0054SCI	1	1	2	1	1	1	0.45

H: 1 = ‘12 y.o. (T1)’, 2 = ‘16 y.o. (T2)’
I: 1 = ‘male’, 2 = ‘female’
J: 1 = ‘Mataró’, 2 = ‘Manlleu’
K: 1 = ‘Catalan’, 2 = ‘Catalan and Spanish’, 3 = ‘Spanish’, 77 = ‘Portuguese’
L: Numeric value from 0 (no use of Catalan with the social network) to 1 (exclusive use of Catalan)
M: Numeric value from 0 (no exposure to media or cultural products in Catalan) to 1 (exposure to media and cultural products in Catalan only)

Univariate and multivariate approaches have been used to investigate linguistic variation (cf. Jacewicz et al. 2009: 245–246, Pichler 2010: 592). However, following Hinrichs and Szmrecsanyi (2007: 470), only multivariate techniques are utilised in this exploratory study, since several variables (e.g. language of identification, the index of Catalan in the social network and the index of Catalan in culture and media) reflect interrelated phenomena. These authors, following observations on the limitations of univariate analysis by Gries (2003: 185) and others, state that:

> Whenever the set of independent variables exceeds a couple of (possibly not entirely independent) factors, corpus-based research into variation in time and space should adopt multivariate methodologies, which have long been state-of-the-art in variationist linguistics and in the social sciences in general.
>
> (Hinrichs and Szmrecsanyi 2007: 470)

We employed decision trees, an exploratory multivariate statistical technique which combines descriptive and predictive analysis and projects results into a graphic that makes interpreting the results an intuitive task. Thus, a decision-tree induction algorithm was applied to the data using IBM's SPSS software (version 25) to test the hypothesis that the independent variables could, in some combination, predict most deontic constructions used by the informants.

Decision-tree induction is a predictive *ad hoc* classification technique, like discriminant analysis or neural networks (Berlanga Silvente, Rubio Hurtado and Vilà Baños 2013: 66). The technique is used in various research areas, including natural language processing in medicine (Gordon et al. 2022), research on aphasic speech (Fromm et al. 2021) and child language acquisition and development (Kim et al. 2019). Compared to other techniques, it presents several advantages.

Decision trees can be used with large sets of discrete and continuous variables (Schmid 2010: 195; Song and Yu 2015: 130). Results can be displayed in an easy-to-interpret graph (Schmid 2010: 195; Pérez 2011). They are fast, robust, accurate and unambiguous (Schmid 2010: 195; Berlanga Silvente, Rubio Hurtado and Vilà Baños 2013: 68; Song and Yu 2015: 130). Lastly, they are easy to create and need little parameter adjustment to work effectively (Schmid 2010: 195; Song and Yu 2015).

Nevertheless, they also present limitations. For example, not all relations identified between the variables are causal, even if they are selected because they improve the underlying statistical model (Song and Yu 2015: 134). Therefore, the nature of the correlations discovered in this study is not specified, i.e. Section 5 merely describes the relationships identified by the decision trees. Similarly, depending on the training data, results may not be generalisable due to over- or underfitting. This, however, is not an issue in this study, only intended as a first exploration of the sample. Lastly, changes in the training data may affect algorithm performance (Schmid 2010: 188). This was evidenced here. The first dataset included instances of *caldre* followed by inflective subjunctive clauses. With these data, the algorithm produced slightly worse classification results (up to 91.3% of correct classifications) than it did after they were removed (up to 92.6% of correct classifications; see Section 5).

There are several growing methods for decision trees (Rokach and Maimon 2015: 81). We used chi-square automatic interaction detector (CHAID), created specifically to work with nominal variables (Rokach and Maimon 2015: 79), like most of the ones in our study. This algorithm chooses the independent variables presenting the strongest interaction with the dependent variable (Berlanga Silvente, Rubio Hurtado and Vilà Baños 2013: 68). It operates with a merge threshold, against which it compares pairs of values of the independent variables (Rokach and Maimon 2015: 79–80). Because the dependent variable in our study is nominal, the statistical test used to compare pairs of values is a Pearson chi-squared test. If two values generate a p value greater than the threshold ($p = 0.05$), CHAID merges them and searches for a new pair. This happens as many times as necessary until one of the following situations occurs: all pairs of values that are significantly different from the values in the dependent variable are found, a node contains all the cases it can or the tree reaches its maximum depth. In the first scenario, the best variable to split a node is selected "such that each child node is made of a group of homogenous values" of that variable. Splits are only performed if the p value adjusted with the Bonferroni correction method of the best variable is lower than a certain split threshold. The result of this process is a multidirectional decision tree obtained rapidly and effectively (Berlanga Silvente, Rubio Hurtado and Vilà Baños 2013: 68).

5. Results

Three models were generated using decision trees and different combinations of the ten independent variables. This section describes the results of this exploratory study.

The first model analysed the data in connection with the sociolinguistic variables – speakers' age, sex, city, language of identification, index of Catalan in culture and media and index of Catalan in the social network. Three of these were included in the resulting tree (Table 6). No validation measure was used for this exploratory analysis, and the maximum tree depth was set at ten so that the algorithm could use as many variables as relevant for the statistical model (the maximum was set at a greater value than the total of independent variables fed into the model). The default settings were used for the minimum of cases per child and parent nodes. The resulting tree has a depth of two nodes below the root and a total of 18 nodes, 13 of which are terminal.

Table 6. Model summary for the decision tree generated from the sociolinguistic variables

Specifications	Growing Method	CHAID
	Dependent Variable	Deontic_constructions
	Independent Variables	Age, Sex, City, Language_of_identification, Catalan_culture_index, Social_network_index
	Validation	None
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Catalan_culture_index, Age, Sex
	Number of Nodes	18
	Number of Terminal Nodes	13
	Depth	2

The model classifies 71.7% of tokens overall correctly (Table 7). Specifically, the percentage of correct classifications is higher for *haver de* (89.7%) and lower for *tenir que* (48.5%). *Caldre, tenir de* and *haver que* are never predicted correctly. The estimated risk is 0.283, with a standard error of 0.014.

Table 7. Classification results for the decision tree generated from the sociolinguistic variables

Observed	**Predicted**					
	caldre	**haver_de**	**tenir_que**	**tenir_de**	**haver_que**	**Percent correct**
caldre	0	36	2	0	0	0.0%
haver_de	0	618	71	0	0	89.7%
tenir_que	0	151	142	0	0	48.5%
tenir_de	0	25	14	0	0	0.0%
haver_que	0	0	1	0	0	0.0%
Overall percentage	0.0%	78.3%	21.7%	0.0%	0.0%	71.7%

Growing Method: CHAID
Dependent Variable: Deontic_constructions

Figure 1 shows the resulting decision tree with the sociolinguistic variables selected and the predicted category for each node highlighted in grey (note that this is never *caldre, tenir de* or *haver que*). Node 0 displays the percentages of each deontic construction. The main predictive variable in this case is the index of Catalan in culture and media. Age and sex are also selected.

The most frequent variant in most nodes is *haver de*. The factors that seem to favour *tenir que* are, firstly, a low index of Catalan in the media and culture, with values between 0 and 0.08 (node 2) or between 0.17 and 0.22 (node 5); secondly, being 12 years of age for speakers with an index between 0 and 0.08 (node 12); and, thirdly, being female for speakers with an index between 0.13 and 0.17 (node 14).

However, an overall correct prediction percentage of 71.7% means that 28.3% of instances (i.e. nearly 300) of the deontic constructions in the sample are misclassified. Thus, the information provided by the sociolinguistic variables cannot account for all the variability in the sample: restricting the analysis to these independent variables might lead to a possibly erroneous conclusion that over 25% of the sample is due to free variation. Next, therefore, the linguistic variables are used in order to compare the new results to this model.

The second model analyses the data with the linguistic variables (grammatical person, priming, sentence polarity and markedness). Table 8 shows that the specifications regarding validation, maximum tree depth and minimum of cases per child and parent nodes remained the same as in the previous model for comparability. The model generated a decision tree with a depth of four nodes and a total of nine, five of which are terminal.

The classification and risk results for this model are a considerable improvement on the previous ones, as shown in Table 9. Correct classifications amount to 92.6% overall, reaching 100% for *haver de* and *tenir que*, even though *caldre, tenir de* and *haver que* are still consistently misclassified. The estimated risk of misclassifications is, thus, set at 0.074 (with a standard error of 0.008).

Figure 2 shows the decision tree with the linguistic variables. The main predictive independent variable in this model is markedness, which differentiates between *haver de, caldre* and *tenir de* (Node 1) and *tenir que* and *haver-hi que* (Node 2). This split therefore separates the two most frequent allostructions. In Node 1, encompassing most of the sample (72.3%), most instances correspond to *haver de* (89.9%), whereas in Node 2, nearly all tokens are *tenir que* (99.7%).

Node 1 diverges into two nodes according to sentence polarity. Node 3 includes all instances of the allostructions perceived as genuinely Catalan in negative form (8.5% of the sample), and Node 4 contains the rest of tokens (63.8% of the data). The most frequent variant in both is *haver de*. In Node 3, this construction represents 65.6% of the total, whereas in Node 4, this percentage rises to

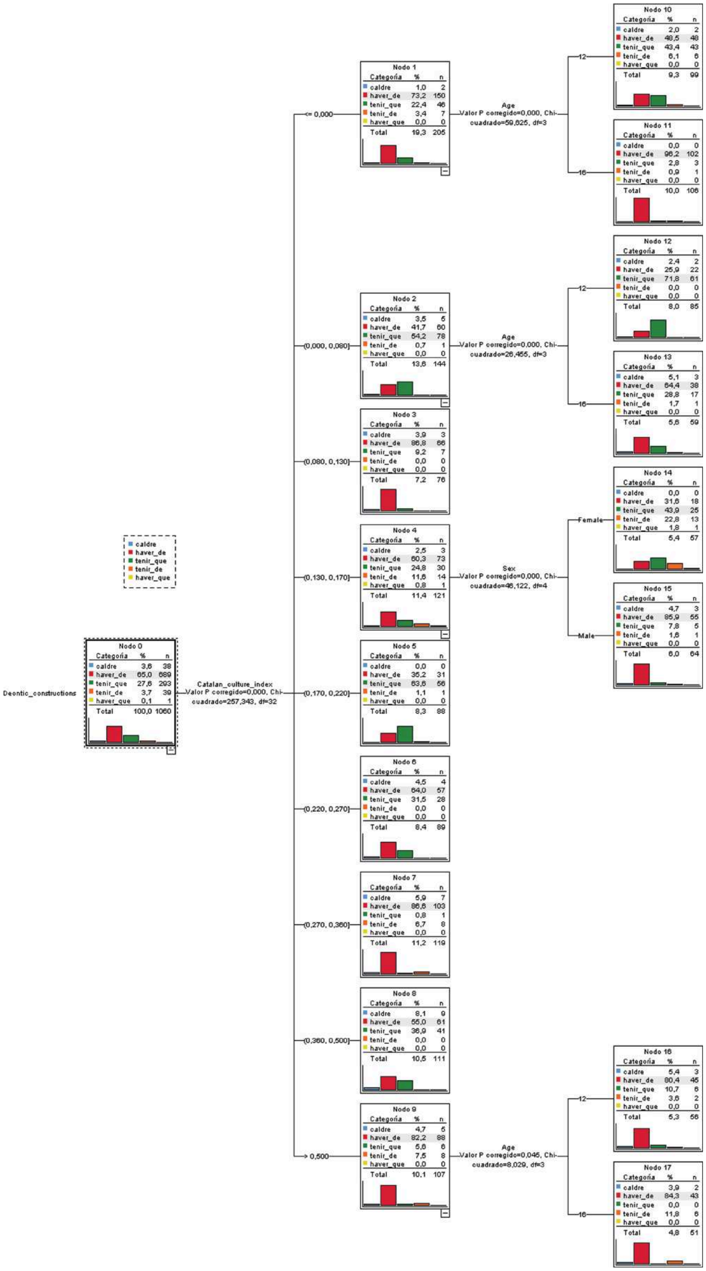


Figure 1. Decision tree of the sample of deontic constructions with sociolinguistic variables

Table 8. Model summary for the decision tree generated from the linguistic variables

Model summary		
Specifications	Growing Method	CHAID
	Dependent Variable	Deontic_constructions
	Independent Variables	Grammatical_person, Priming, Sentence_polarity, Markedness
	Validation	None
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Markedness, Sentence_polarity, Priming, Grammatical_person
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	4

Table 9. Classification results for the decision tree generated from the linguistic variables

Observed	**Predicted**					
	caldre	**haver_de**	**tenir_que**	**tenir_de**	**haver_que**	**Percent correct**
caldre	0	38	0	0	0	0.0%
haver_de	0	689	0	0	0	100.0%
tenir_que	0	0	293	0	0	100.0%
tenir_de	0	39	0	0	0	0.0%
haver_que	0	0	1	0	0	0.0%
Overall percentage	0.0%	72.3%	27.7%	0.0%	0.0%	92.6%

Growing Method: CHAID
Dependent Variable: Deontic_constructions

93.2%. Furthermore, while the second most frequent variant in Node 3 is *caldre* (32.2%), in Node 4, *haver de* is followed by *tenir de* (with only 5.5% of tokens).

Node 4 splits into Nodes 5 and 6 through the variable priming. On the one hand, this classifies all instances in Node 6 as primings of *haver de* (18% of the sample); on the other, Node 5 includes 45.5% of the data, in which *haver de* is also the most frequent variant (90.5%), followed by *tenir de* (7.6%) and *caldre* (1.9%).

Lastly, Node 5 divides into two more nodes by the variable grammatical person. Node 7 includes instances of third person (singular and plural) and first person plural (in total, 22.8% of the sample). Node 8 comprises second person (singular and plural) and first person singular (22.9% of the sample). The main difference between these nodes lies, again, in the second most frequent construction, since *haver de* is the most common in both, with 93% of occurrences in Node 7 and 88.1% in Node 8. The distribution of the variants is more even in Node 7 (where *caldre* accounts for 3.7% of cases and *tenir de* for 3.3%) than in Node 8, where the only other variant is *tenir de* (11.9% of tokens).

In short, this model shows that all linguistic variables are relevant for predicting the occurrences of the constructions in the sample. More specifically, markedness influences *tenir que* in that only 0.3% of the constructions perceived as contact-originated are of *haver-hi que* instead of *tenir que*. As for *caldre*, it is most likely to occur in negative sentences. Focusing on *tenir de*, we can see that no variable in this model specifically predicts its occurrence. Most instances of this construction appear in non-negative sentences, never as primings and mostly in second- or first-person singular forms. Finally, the variant *haver de* is favoured throughout the nodes in the decision tree and can, therefore, be described as the default or unmarked construction. It is one of the forms perceived as genuine and used more in affirmative or neutral sentences (although also, to a lesser degree, in negative sentences), in contexts where the speakers are primed by the interviewer (but also otherwise) and more so in third person and first person plural than in second-person and first-person singular.

Thus, since this model does not seem to distinguish particularly well between the constructions perceived as genuine, specifically between *tenir de* and *haver de*, and since the classification and risk results show that 7.4% of cases (i.e. more than 78 instances) are still misclassified, the next step in our analysis consists in using all the independent variables as input for a new model. Ideally, this new attempt would correctly classify all the deontic constructions in the sample.

The ten independent variables were fed into the last model, which selected four (markedness, sentence polarity, index of Catalan in the social network and city) as relevant to predict instances. Table 10 displays a summary of this third decision tree with a depth of four and 12 nodes in total (eight of which are terminal).

Table 11 shows the results of the classification of tokens, which match those of the previous model. The three variants with fewer instances are again classified incorrectly. The risk estimate and the standard error values also remain stable (0.074 and 0.008, respectively).

Figure 3 shows the resulting decision tree. The first split is performed, as before, through the variable markedness. Node 1 is then divided by sentence

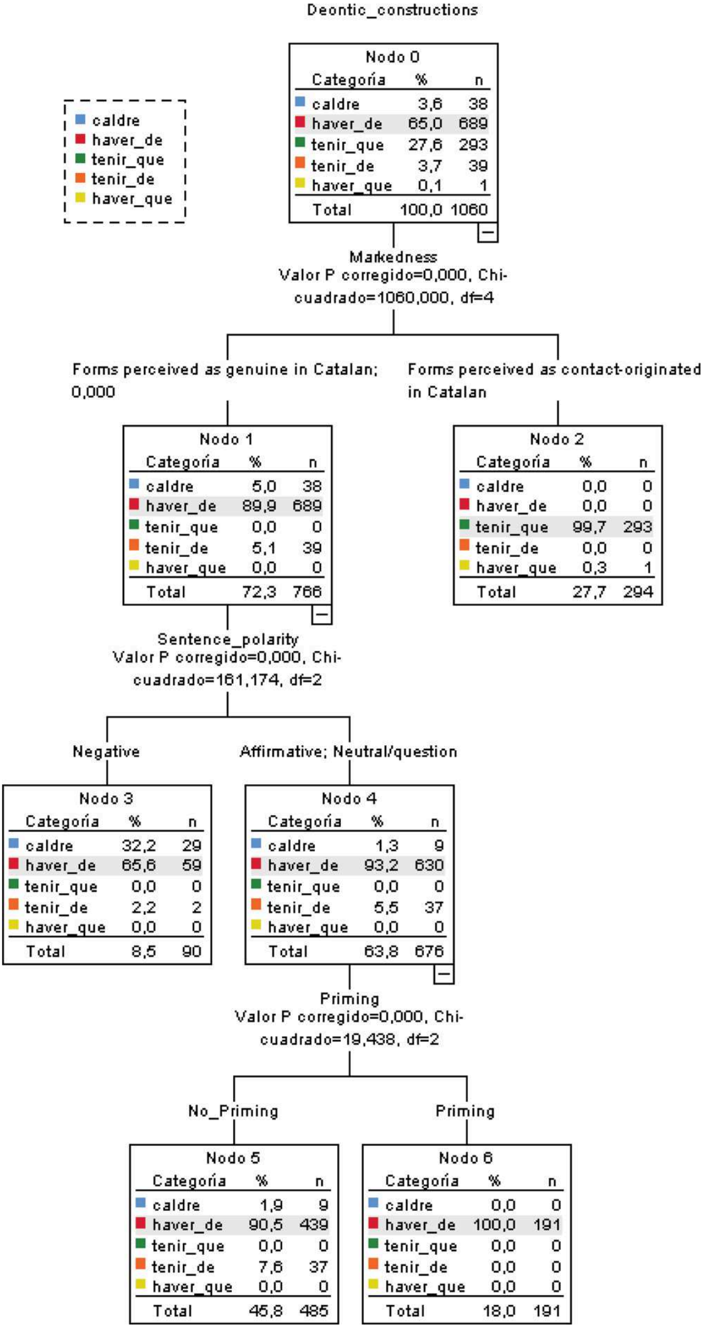


Figure 2. Decision tree of the sample of deontic constructions with linguistic variables

Table 10. Model summary for the decision tree generated from all the independent variables

Model summary		
Specifications	Growing Method	CHAID
	Dependent Variable	Deontic_constructions
	Independent Variables	Age, Sex, City, Language_of_identification, Catalan_culture_index, Social_network_index, Grammatical_person, Priming, Sentence_polarity, Markedness
	Validation	None
	Maximum Tree Depth	10
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	Markedness, Sentence_polarity, Social_network_index, City
	Number of Nodes	12
	Number of Terminal Nodes	8
	Depth	4

polarity into Nodes 3 and 4 (equivalent to Nodes 3 and 4 in the previous tree). Next, Node 4 is split into five more nodes through the index of use of Catalan in the social network. Nodes 5 and 6 each group nearly 6% of the sample, Node 7 accounts for 24.2%, and Nodes 8 and 9 encompass around 14% each. *Haver de* is the most frequent variant in all these nodes, albeit with different percentages than in the previous tree. In Node 5, corresponding to constructions by speakers with an index of Catalan use in their social network of 0, *haver de* is found in 75.8% of instances and *tenir de* in the other 24.2%. Node 6 groups speakers with an index between 0 and 0.11. In this node, *haver de* accounts for 91.8% and *tenir de* for the rest of the tokens (8.2%). In Node 7, *haver de* is used in 98.4%, *caldre* in 1.2% and *tenir de* in 0.4% of cases. In Node 8, the percentage of *haver de* descends to 86.6%, followed by *tenir de* (10.7%) and *caldre* (2.7%). Finally, Node 9 groups only instances of *haver de* (98.6%) and *caldre* (1.4%).

Table 11. Classification results for the decision tree generated from all the independent variables

Observed	Predicted					
	caldre	haver_de	tenir_que	tenir_de	haver_que	Percent correct
caldre	0	38	0	0	0	0.0%
haver_de	0	689	0	0	0	100.0%
tenir_que	0	0	293	0	0	100.0%
tenir_de	0	39	0	0	0	0.0%
haver_que	0	0	1	0	0	0.0%
Overall Percentage	0.0%	72.3%	27.7%	0.0%	0.0%	92.6%

Growing Method: CHAID
Dependent Variable: Deontic_constructions

On the last level, Node 8 is split into Nodes 10 and 11 through the variable of city. Node 10 includes constructions by informants from Mataró and represents 6.9% of the sample. This node shows instances of *haver de* (93.2%), *caldre* (4.1%) and *tenir de* (2.7%). In contrast, Node 11 contains 7.2% of the sample and corresponds to speakers in Manlleu. Most (80.3%) of the tokens here are *haver de*, 18.4% *tenir de* and 1.3% *caldre*.

In short, the variables used in this combined model to classify the instances in the sample are markedness, sentence polarity, index of Catalan in the social network and city. It seems, from the order of the selected variables in the decision tree and from the results of the former models, that the linguistic variables are more relevant in explaining the variability of the data than the sociolinguistic factors. As with the previous tree, *haver de* is the predicted variant in almost all nodes except for Node 2, which gathers the forms perceived as contact originated. Regarding *tenir de*, most instances are produced in non-negative sentences by speakers with a very low (up to 0.11) or very high (between 0.75 and 0.98) index of Catalan in the social network, especially those in Manlleu. Thus, the differences between the variants perceived as genuine (*haver de, tenir de* and *caldre*) are somewhat clearer in this tree than in the previous model.

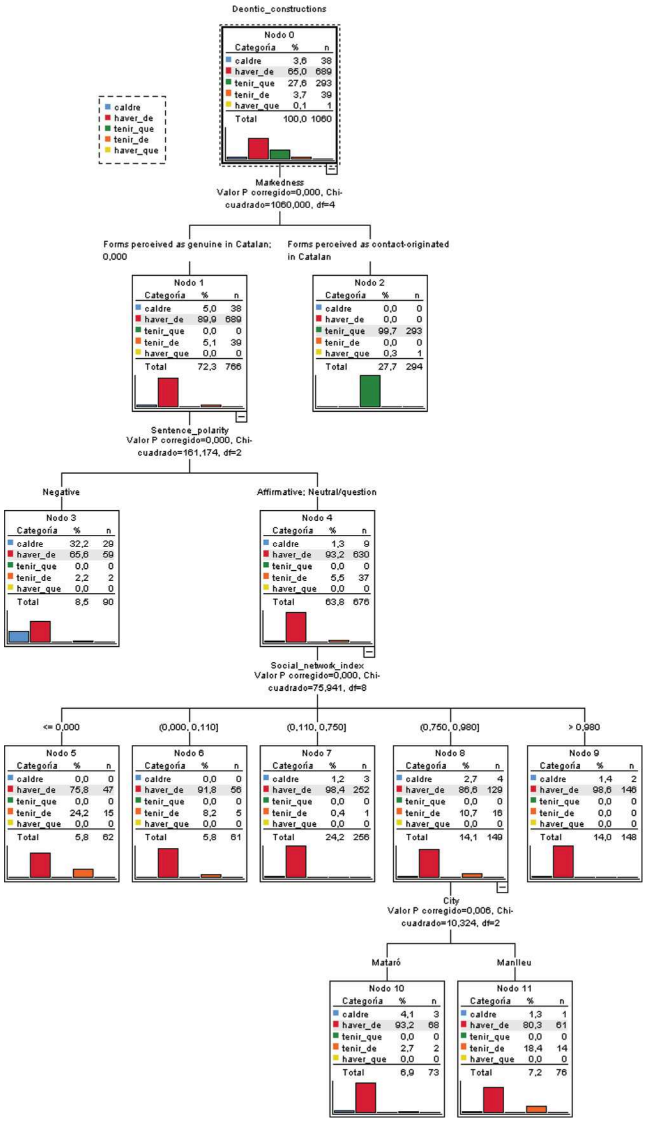


Figure 3. Decision tree of the sample of deontic constructions with all the independent variables

6. Discussion of results and possible future lines of research

The hypothesis of this study was that some combination of the variables analysed would predict most tokens in the sample. This was demonstrated for all the trees generated. However, some results, summarised below, merit more attention in future studies.

In the first decision tree, results regarding the index of Catalan in culture and media in relation to the Spanish-influenced variant *tenir que* are unexpected. Previous publications show that exposure to television, music and the Internet in a target language may positively affect speakers' proficiency levels (e.g. Peters 2018; De Wilde, Brysbaert and Eyckmans 2020; Muñoz and Cadierno 2021). Nevertheless, instead of a lower use of variants proscribed in prescriptive works, an increase in the indices of Catalan in culture and media between many nodes on the first decision tree correspond to more use of *tenir que* (cf. *tenir que* tokens between Nodes 1 and 2, 3 and 4; 4 and 5 and 7 and 8 in Figure 1). Yet Node 2 shows the highest percentage of *tenir que* and corresponds to speakers with low indexes of Catalan in the media, a finding which can be easily explained from the literature. Thus, further exploration of the role of exposure to Catalan in culture and the media in the present phenomenon of variation may shed light on these complex results.

Another interesting finding of the first tree is that male informants mainly use the standard *haver de*, whereas female speakers show a preference for *tenir que*. This agrees with changes from below identified in previous studies of Catalan varieties (see Turell 1995). However, when considering its effect on sociolinguistic variation, "[w]e clearly cannot talk about gender independently of other aspects of social identity, as no variable correlates simply with gender or social category" (Eckert 1997: 73). Therefore, future endeavours should use tools capable of exploring the relations between independent variables more deeply than decision trees.

In the second and third models, which show the best classification performances, as mentioned above, *haver de* proves to be the unmarked variant. This is also indicated by the interviewers predominantly using this construction (in front of the other two variants described as acceptable in the prescriptive grammar), as reflected by the priming results. Language use in the sample thus reflects, to a certain extent, the extended belief that 'the *correct* deontic construction in Catalan is mainly *haver de*, although *caldre* can also be used, and other forms are a product of language contact with Spanish and should therefore be avoided'. This has been repeatedly reproduced as a recommendation in numerous style manuals, language handbooks and online fora (e.g. SLUOC 2016: 48; CPNL n.d.; CNLO 2012). In this sense, Mier's (1986: 57) prediction that the genuine Catalan form *tenir de* could become stigmatised seems supported by the findings.

The results of the second model also suggest that the linguistic variables considered here cannot explain the use of the constructions originating from language contact (see Terminal Node 2 in Figure 2). Thus, the sociolinguistic situation and related factors might override grammatical constraints in this case. However, as noted above, because of the underlying assumption in variationist linguistics, our results do not allow us to dismiss the possibility that other linguistic variables not included in our study may play a role in the use of these constructions.

Results in relation to sentence polarity (Figures 2 and 3) align with Rigau's (2005a: 256) observation that *caldre* is most frequently used in negative form. Furthermore, it seems that the other constructions may also be constrained by sentence polarity. This has yet to be explored in the literature. Therefore, it will be interesting to observe if the tendencies identified here are related to the communicative task carried out by the informants or to the grammar of these constructions.

Still on the second model, the variable grammatical person seems relevant to identifying the instances of the defective form *caldre*, as expected. However, it is also worth noting that *tenir de* is most frequently used in first-person singular and second person. In light of these findings, and since there seems to be no previous literature on the use of this variant in contemporary Catalan varieties, it would be interesting to explore its uses in other samples and check whether this preference is also found or whether it may be related to the characteristics of the sample and the informants' communicative purpose in our study.

Additionally, with regard to this understudied variant, Figure 3 shows that it is mostly used by speakers with either a very low index of Catalan in their social network (of up to 0.11) or quite a high one (between 0.75 and 0.98). This finding suggests that speakers with values away from the two extremes might be avoiding its use, perhaps, in line with Mier's (1986: 57) prediction, because it is a stigmatised construction.

Figure 3 also points to possible geographical differences between the two cities in the sample. As observed, Nodes 10 and 11 comprise similar portions of the data in numerical terms (6.9% and 7.2%, respectively) but show different distributions of the tokens. Informants in Manlleu use *tenir de* seven times more frequently than those in Mataró. Diatopic differences in the use of deontic periphrases were also found in Stokes (2015), where Valencian dialects make a more frequent use of the Spanish-influenced *tenir que* than the other varieties. This is consistent with our data, since 78.84% (231) of the total occurrences of *tenir que* (293) were produced by informants in Mataró, a city in the Barcelona Metropolitan area where Spanish was the predominant first language of the population in 2008 and 2013, at 63.1% and 64.3%, respectively (SPL 2008: 174, DGPL 2015: 43). Only 21.16% (62) of tokens were produced by adolescents in Manlleu, located in central Catalonia, where 50.4% and 49.1% of the population claimed Catalan as their first language in 2008 and 2013 (SPL 2008: 174, DGPL 2015: 43).

Lastly, although this study operationalised the speakers' perceptions of the diachronic origins of the constructions as the linguistic variable 'markedness', future endeavours might construe it as a sociolinguistic factor due to its relation with linguistic attitudes and discourses on normativity. It would be interesting, for example, to survey speakers' opinions on the 'genuineness' of each variant to empirically test the observations made in our study and previous works (e.g. Mier 1986).

This study is, thus, a successful first exploration of the phenomenon of variation at hand in our sample of contemporary Central Catalan by adolescent speakers, since it serves to identify different paths to be explored further. Additionally, future studies of this phenomenon should consider issues which were left out of the present study exclusively due to practical constraints (e.g. verb tense as a linguistic variable or the allostruction *fer falta*).

7. Can variationist linguistics prove the (non)existence of free variation?

This study takes a variationist approach to linguistics in that it attempts to correlate speakers' morphosyntactic behaviour and (socio)linguistic data. The constellation of variants analysed was designed to capture more options available to speakers than previous publications, which focused on *haver de* and *tenir que*. These constructions account indeed for most of our sample. However, not to consider other constructions with equivalent meanings may lead to partial and inaccurate descriptions of variation phenomena. Therefore, more detailed examinations of this sample will also consider the verb phrase *fer falta*. Despite this limitation, by considering five variants, this exploratory study provides an overall view of the alternatives at the speakers' disposal in vernacular Catalan varieties to express deontic meanings.

Furthermore, the statistical method used is well suited to large datasets and many independent variables. As shown, the intuitive nature of the statistical products generated is compatible with robust and statistically significant results. Since only statistically significant correlations are identified and included in the final decision tree, this method is appropriate for exploratory studies like this one and studies greatly conditioned by practical constraints. The correlations between the variables identified contribute to a better understanding of the factors which are likely to interfere in the use of the constructions. However, a relation of causality between the variables cannot be assumed. Nevertheless, two of the models reach very high correct classification percentages (92.6%), a strong indicator that the selected variables should be considered to explain the use of these constructions.

Remarkably, the second decision tree performs as well as the third one without resorting to sociolinguistic variables, which might raise the question of whether social factors are at all necessary to explain variation phenomena. This provocative question, however, deserves to be considered in much more depth than can be achieved in this paper.

The few tokens of *caldre, tenir de* and *haver que* in the sample were not correctly classified by any of the models, which highlights the shortcomings of this method with limited data. Future studies will need datasets with more instances of these variants or tools which are better equipped to deal with few realisations. Yet, within our approach, the fact that the models consistently grouped the only instance of *haver-hi que* with *tenir que* might indicate that (at least in relation to the independent variables in this study), in the sample, the occurrences of these variants – not by chance, the two originated by language contact – behave in similar ways.

As explained in Section 3.1, only instances of *caldre* followed by infinitive clauses were included in the sample to preserve equivalence between the variants as much as possible. However, future approaches might make it possible to consider other uses of this construction, although the semantic and syntactic differences between the variants should not be ignored (see Section 3).

As happens frequently in linguistic research, the independent variables may interact not only with the dependent variable but also with each other to some extent. This is why many authors advocate for multivariate methods and univariate tools were not used (see Section 4). However, the results do not provide detailed information on the potential relationships between the independent variables. Instead, they inform us of the combinations of variables relevant for classifying most of the speakers' productions.

Lastly, our approach means that the percentage of cases misclassified by the models cannot be taken to represent the amount of free variation in the sample. Rather, results indicate that the independent variables used failed to provide the information needed to correctly classify all the tokens. Thus, since the independent variables in the second and third models succeed in classifying all the instances of *haver de* and *tenir que*, it seems that free variation does not play a role in accounting for the use of these constructions in the sample (i.e. free variation can be ruled out from the results obtained for these variants). Nevertheless, the portion of the sample that does not correlate with the independent variables selected by the models (7.4% of instances) cannot be attributed to free variation solely from these results. According to the assumptions underlying variationist linguistics, a complementary set of independent variables might improve the performance of the models and ultimately rule out free variation for all the variants. Simultaneously, however, note that:

> Even if large numbers of independent variables are factored into the analysis, we can never be certain that other important but unknown variables are not responsible for the observed variation. Instead, causative links must be considered in light of the results of the study and the possible existence of confounding factors that have gone uncontrolled. This is the primary limitation of adopting an observational approach to linguistics. (Grieve 2021: 6–7)

Thus, through the lens of variationist linguistics, free variation may prove to be an unverifiable notion, but it still plays an important role in linguistics, like dark matter in cosmology. According to Robson (2018), for years, cosmologists observed rotation curves in spiral galaxies that contradicted their expectations, based on their knowledge that galaxies were formed by stars and gas. Several studies pointed to a discrepancy in the mass of spiral galaxies. To account for the flat rotation curves observed, Ostriker, Peebles and Yahil (1974) hypothesised that there was a spherical halo of unknown matter around spiral galaxies. This unknown matter is commonly known as dark matter, and it is believed to "provid[e] a large contribution to the gravitational field at large distances from the center of the galaxy", even though this hypothesis "has yet to be verified", mostly because of the unknown nature of this matter. The dark matter hypothesis boosted numerous advancements in physics, including the postulation of new hypothetical particles (Robson 2018). However, the dark matter hypothesis is not uncontroversial, and different modified gravity theories have emerged to explain the flat rotation curves of spiral galaxies independently of dark matter (e.g. Milgrom 1983).

Similarly, linguists have observed the use of different but equivalent structures for decades (e.g. deontic constructions in vernacular Catalan), apparently with no connection to (extra)linguistic factors. While some argue that empirical studies will eventually account for most variation in language, others claim that a percentage of the variability in language use corresponds to an 'invisible' (unverifiable) object, free variation, which can only be studied indirectly. Both stances contribute to the advancement of models of grammar (and linguistics more generally). Our study exemplifies that the presence of free variation can be ruled out but not demonstrated from a traditional variationist approach. Therefore, compelling proof of its eventual existence can only be obtained from other perspectives on the study of language.

8. Conclusion

This paper has dealt with the (dis)advantages of variationist linguistics in relation to free variation through the example of its application to the little-researched variation between deontic constructions in spoken Catalan. It has questioned its capacity to demonstrate that free variation does (not) exist, since a fundamental premise in variationist linguistics is that the right set of variables can account for the data.

However, as exemplified by the last two models in our exploratory study, different sets of independent variables may yield similar results, which poses some important questions: which of all the factors that yield good results should be used to explain the data? Precisely how do they influence speakers' choices? And, above all, would other independent variables produce similar (or even better) results? If so, (how) could one ever finish analysing a particular dataset, provided that the percentage of correct classifications never reached 100%?

Despite this conundrum regarding the (im)possibility of proving free variation within the variationist framework, variationist linguistics has generated a wealth of valuable knowledge over time. Indeed, the need to combine usage- and system-based approaches has been repeatedly stated in the literature. Lastly, this paper has claimed that the notion of free variation benefits scholars in different areas of linguistics regardless of whether its existence can be demonstrated, just like the dark matter hypothesis has boosted our knowledge of physics by sparking interest in the implications of its existence on the one hand and potential alternative explanations of empirical observations that initially seemed to contradict established knowledge on the other.

Acknowledgements

We thank the two anonymous reviewers for their insightful feedback, which helped improve an earlier version of this paper.

References

Adger, David & Trousdale, Graeme. 2007. Variation in English syntax: Theoretical Implications. *English Language and Linguistics* 11 (2): 261–278.

Adger, David. 2016. Language Variability in Syntactic Theory. In *Rethinking Parameters*, Luis Eguren, Olga Fernandez-Soriano & Amaya Mendikoetxea (eds), 49–63. New York: Oxford University Press.

Adli, Aria, García García, Marco & Kaufmann, Göz. 2015. System and Usage: (Never) Mind the Gap. In *Variation in Language: System- and Usage-based Approaches*, Aria Adli, Marco García García & Göz Kaufmann (eds), 1–25. Berlin, Boston: De Gruyter.

Bader, Markus. 2020. Analyzing Free Variation with Harmony – A Case Study of Verb-Cluster Serialization. *Zeitschrift für Sprachwissenschaft* 39 (3): 407–437.

Baechler, Raffaela & Pröll, Simon. 2019. Analyzing Language Change through a Formalist Framework. In *Morphological Variation. Theoretical and Empirical Perspectives*, Antje Dammel & Oliver Schallert (eds), 63–94. Amsterdam, Philadelphia: John Benjamins.

Ball, Martin J. (ed). 2010. *The Routledge Handbook of Sociolinguistics Around the World*. London, New York: Routledge.

Berlanga Silvente, Vanesa, Rubio Hurtado, María J. & Vilà Baños, Ruth. 2013. Cómo aplicar *árboles de decisión* en SPSS [How to apply *decision trees* in SPSS]. *Revista d'Innovació i Recerca en Educació* 6 (1): 65–79.

Blas Arroyo, José L. 2015. The Scope of Language Contact as a Constraint Factor in Language Change: The Periphrasis *haber de* plus Infinitive in a Corpus of Language Immediacy in Modern Spanish. *International Journal of Bilingualism* 19(5): 499–524.

Bretxa i Riera, Vanessa. 2014. El salt a secundària. Canvis en les tries lingüístiques i culturals dels preadolescents mataronins en la transició educativa. PhD dissertation. Barcelona: Universitat de Barcelona.

Bretxa i Riera, Vanessa. 2019. Capítol 7. Actituds i representacions lingüístiques. In *Els usos lingüístics als territoris de llengua catalana*, coord. by Direcció General de Política Lingüística i Xarxa CRUSCAT-IEC., 94–109. Barcelona: Generalitat de Catalunya, Departament de Cultura.

Bretxa, Vanessa, Comajoan, Llorenç, Ubalde, Josep & Vila, F. Xavier. 2016. Changes in the Linguistic Confidence of Primary and Secondary Students in Catalonia: A Longitudinal Study. *Language, Culture and Curriculum* 29 (1): 56–72.

Bretxa, Vanessa & Vila, F. Xavier. 2012. Els canvis sociolingüístics en el pas de primària a secundària: el projecte RESOL a la ciutat de Mataró [Sociolinguistic changes in the move from primary to secondary education: The RESOL project in the city of Mataró]. *Treballs de Sociolingüística Catalana* 22: 93–118.

Bretxa, Vanessa, & Vila, F. Xavier. 2014. L'evolució dels usos lingüístics dins l'aula des de sisè de primària fins a quart d'ESO [The development of patterns of language use in the classroom from sixth grade to the fourth year of the mandatory secondary education program]. *Revista de Llengua i Dret* 62: 106–123.

Cabanes Fitor, Vicent. 1996. Les perífrasis modals de necessitat-obligació i probabilitat en català. Seguiment diacrònic: segles XIII al XIX. [The modal periphrases of necessity-obligation and probability in Catalan. Diachronic monitoring: 13th to 19th centuries]. *Caplletra. Revista Internacional de Filologia* 20: 129–164.

Cappelle, Bert. 2009. Can We Factor Out Free Choice? In *Describing and Modeling Variation in Grammar*, Guido Seiler, Andreas Dufter & Jürg Fleischer (eds), 183–202. Berlin, New York: De Gruyter.

Chambers, J.K. & Schilling, Natalie. 2013. *The Handbook of Language Variation and Change*. Chichester: Wiley-Blackwell.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Comajoan, Llorenç. 1998. The Sociolinguistic Situation of Catalan at the Turn of the 21st Century: Immigration and Intergenerational Transmission. *Catalan Review*, XVIII (1–2): 67–95.

Conde Noguerol, María E. 2016. Hacia una caracterización semántica del verbo *caler*. [Towards a semantic characterization of the verb *caler*]. *Sintagma* 28: 93–106.

Consorci per la Normalització Lingüística d'Osona. 2012. Haver de / *tenir que. *Ets i uts…* https://blogs.cpnl.cat/etsiuts/2012/01/13/haver-de-tenir-que/ (19 February 2023).

Consorci per la Normalització Lingüística. N. d. 25. Les perífrasis verbals. *Gramàtica*. https://www.cpnl.cat/gramatica/52/25-les-perifrasis-verbals (19 February 2023).

Cornips, Leonie. 2015. The No Man's Land between Syntax and Variationist Sociolinguistics: The Case of Idiolectal Variability. In *Variation in Language: System- and Usage-based Approaches*, Aria Adli, Marco García García & Göz Kaufmann (eds), 147–171. Berlin, Boston: De Gruyter.

De Wilde, Vanessa, Brysbaert, Marc & Eyckmans, June. 2020. Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition* 23(1): 171–185.

DGPL = Direcció General de Política Lingüística (ed). 2015. *Enquesta d'usos lingüístics de la població 2013*. Barcelona: Generalitat de Catalunya. Departament de Cultura. Direcció General de Política Lingüística. https://www.idescat.cat/cat/idescat/publicacions/cataleg/pdfdocs/eulp2013.pdf (19 February 2023).

Eckert, Penelope. 1997. Gender and sociolinguistic variation. In *Readings in Language and Gender*, Jennifer Coates (ed), 64–75. Oxford: Blackwell.

Ellis, Rod. 1999. Item Versus System Learning: Explaining Free Variation. *Applied Linguistics* 20(4): 460–480. (19 February 2023).

Espuny, Janina. 1998. Aspectes de la interferència lèxica castellana en el discurs oral català. In *Oralment: Estudis de variació funcional*, Lluís Payrató (ed), 275–290. Barcelona: Publicacions de l'Abadia de Montserrat.

Fabra, Pompeu. 1974. *Gramàtica catalana*. Barcelona: Editorial Teide.

Flors Mas, Avel·lí. 2015. Young People and Languages in Catalonia: The State of the Question. In *Linguapax Review 2015. The Role of Youth in Language Revitalisation / El paper dels joves en la revitalització lingüística*, Josep Cru (ed), 30–42. Barcelona: Linguapax International.

Flors Mas, Avel·lí & Vila i Moreno, F. Xavier. 2014. Justifying Preferences. How Catalan Adolescents Explain Their Linguistic Choices. *Treballs de Sociolingüística Catalana* 24: 173–199.

Fromm, Davida, Greenhouse, Joel, Pudil, Mitchell, Shi, Yichun & MacWhinney, Brian. 2021. Enhancing the Classification of Aphasia: A Statistical Analysis Using Connected Speech. *Aphasiology*, 1–28.

Galindo, Mireia, De Rosselló, Carles & Bernat, Francesc. 2021. *El castellà a la Catalunya contemporània: Història d'una bilingüització*. Benicarló: Onada.

Garachana Camarero, Mar. 2017. Los límites de una categoría híbrida. Las perífrasis verbales. In *La gramática en la diacronía: la evolución de las perífrasis verbales modales en español* [Lingüística Iberoamericana 69], coord. by Mar Garachana Camarero, 35–80. Madrid: Iberoamericana, Frankfurt a. M.: Vervuert.

Gonzàlez Balletbò, Isaac, Pujolar Cos, Joan, Font Tanyà, Anna & Martínez Sanmartí, Roger. 2014. *Llengua i joves. Usos i percepcions lingüístics de la joventut catalana*. Barcelona: Generalitat de Catalunya.

Gordon, Alexandra J., Banerjee, Imon, Block, Jason, Winstead-Derlega, Christopher, Wilson, Jennifer G., Mitarai, Tsuyoshi, Jarrett, Michael, Sanyal, Josh, Rubin, Daniel L., Wintermark, Max & Kohn, Michael A. 2022. Natural Language Processing of Head CT Reports to Identify Intracranial Mass Effect: CTIME Algorithm. *The American Journal of Emergency Medicine* 51: 388–392.

Gries, Stefan T. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York, London: Continuum.

Grieve, Jack. 2021. Observation, experimentation, and replication in linguistics. *Linguistics* 59(5): 1343–1356.

Hinrichs, Lars & Szmrecsanyi, Benedikt. 2007. Recent Changes in the Function and Frequency of Standard English Genitive Constructions: A Multivariate Analysis of Tagged Corpora. *English Language and Linguistics* 11(3): 437–474.

Hualde, José I. 1992. *Catalan: Descriptive Grammar*. London: Routledge.

Institut d'Estudis Catalans (IEC). 2016. *Gramàtica de la llengua catalana*. Barcelona: Institut d'Estudis Catalans.

Jacewicz, Ewa, Fox, Robert A. & O'Neill, Caitlin. 2009. Articulation Rate across Dialect, Age, and Gender. *Language Variation and Change* 21: 233256.

Joos, Martin. 2012. The Isolation of Styles. In *Readings in the Sociology of Language*, Joshua A. Fishman (ed), 185–191. Berlin, Boston: De Gruyter.

Kager, René. 2004. *Optimality Theory*. Cambridge: Cambridge University Press.

Kiesling, Scott F. 2011. *Linguistic Variation and Change*. Edinburgh: Edinburgh University Press.

Kim, Hyun M., Schluter, Philip, McNeill, Brigid, Everatt, John, Sisk, Rose, Iusitini, Leon, Taleni, Leali'ie'e, Tautolo, El-Shadan & Gillon, Gail. 2019. Integrating Health, Education and Culture in Predicting Pacific Children's English Receptive Vocabulary at 6 Years: A Classification Tree Approach. *Journal of Paediatrics and Child Health* 55: 1251–1260.

Krivochen, Diego G. 2020. *Algunos problemas de la sintaxis de los auxiliares modales*. Unpublished manuscript. Faculty of Linguistics, Philology and Phonetics. University of Oxford.

Labov, William. 1966. The Linguistic Variable as a Structural Unit. *Washington Linguistics Review* 3: 4–22.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Martines, Josep. 2020. General Lexicon. In *Manual of Catalan Linguistics*, Joan A. Argenter & Jens Lüdtke (eds), 311–350. Berlin, Boston: De Gruyter.

Martínez Díaz, Eva. 2002. Las perífrasis modales de obligación "tener que + infinitivo" y "haber de + infinitivo": Variación e interferencia en el español de Barcelona. PhD dissertation. Barcelona: Universitat de Barcelona.

Martínez Díaz, Eva. 2019. Las actitudes lingüísticas del castellanohablante en la sociedad catalana. [The linguistic attitudes of the Spanish speaker in Catalan society]. *Tonos digital. Revista de estudios filológicos* 37: 1–30.

Mier, Jeanne Z. 1982. A Sociolinguistic Study of Selected Aspects of Catalan. PhD dissertation. Michigan: University of Michigan.

Mier, Jeanne Z. 1986. Estudi sociolingüístic de certs aspectes de la llengua catalana. [Sociolinguistic study of certain aspects of the Catalan language]. *Treballs de Sociolingüística Catalana* 6: 33–112.

Milgrom, Mordehai. 1983. Modification of the Newtonian Dynamics as a Possible Alternative to the Hidden Mass Hypothesis. *The Astrophysical Journal* 270: 365–370.

Moll, Francesc de Borja. 1952. *Gramática histórica catalana*. Madrid: Editorial Gredos.

Muñoz, Carmen & Cadierno, Teresa. 2021. How do differences in exposure affect English language learning? A comparison of teenagers in two learning environments. *Studies in Second Language Learning and Teaching* 11(2): 185–212.

Ostriker, Jeremiah P., Peebles, Philip J. E. & Yahil, Amos. 1974. The Size and Masses of Galaxies and the Mass of the Universe. *The Astrophysical Journal* 193: L1–L4.

Par, Anfós. 1923. *Sintaxi catalana segons los escrits en prosa de Bernat Metge (1398)*. Halle: Max Niemeyer.

Pérez, César. 2011. *Técnicas de segmentación. Conceptos, herramientas y aplicaciones*. Madrid: Gaceta Grupo Editorial.

Peters, Elke. 2018. The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *International Journal of Applied Linguistics* 169(1): 142–168.

Pichler, Heike. 2010. Methods in Discourse Variation Analysis: Reflections on the Way Forward. *Journal of Sociolinguistics* 14(5): 581–608.

Pujolar, Joan. 2001. *Gender, Heteroglossia and Power: A Sociolinguistic Study of Youth Culture*. Berlin: Mouton De Gruyter.

Real Academia Española. 2010. *Nueva gramática de la lengua española*. Madrid: Espasa Libros.

Rigau, Gemma. 1998. La variació sintàctica. Uniformitat en la diversitat. [Syntactic variation. Uniformity in diversity]. *Caplletra. Revista Internacional de Filologia* 25: 63–82.

Rigau, Gemma. 1999. Relativized Impersonality: Deontic Sentences in Catalan. In *Semantic Issues in Romance Syntax*, Esthela Treviño & José Lema (eds), 193–230. Amsterdam, Philadelphia: John Benjamins.

Rigau, Gemma. 2005a. Estudi morfosintàctic del verb *caldre* en el català antic i en l'actual [Morphosyntactic study of the verb *caldre* in Old and modern Catalan]. *Caplletra. Revista Internacional de Filologia* 38: 241–258.

Rigau, Gemma. 2005b. Number Agreement Variation in Catalan Dialects. In *The Oxford Handbook of Comparative Syntax*, Guglielmo Cinque & Richard S. Kayne (eds), 775–805. Oxford, New York: Oxford University Press.

Robson, Brian A. 2018. The Story of Dark Matter. In *Essentials on Dark Matter*, Abraão J. Capistrano de Souza (ed). London: IntechOpen.

Rokach, Lior & Maimon, Oded. 2015. *Data Mining with Decision Trees. Theory and Applications*. Singapore: World Scientific.

Schmid, Helmut. 2010. Decision Trees. In *The Handbook of Computational Linguistics and Natural Language Processing*, Alexander Clark, Chris Fox & Shalom Lappin (eds), 180–196. New York: Blackwell Publishing.

SPL = Secretaria de Política Lingüística (ed). 2009. *Enquesta d'usos lingüístics de la població 2008*. [Survey of linguistic uses of the population 2008]. Barcelona: Generalitat de Catalunya. Departament de Vicepresidència. Secretaria de Política Lingüística. https://llengua.gencat.cat/web/.content/documents/dadesestudis/altres/arxius/eulp2008.pdf (19 February 2023).

Seiler, Guido. 2015. Syntactization, Analogy and the Distinction between Proximate and Evolutionary Causations. In *Variation in Language: System- and Usage-based Approaches*, Aria Adli, Marco García García & Göz Kaufmann (eds), 239–263. Berlin, Boston: De Gruyter.

Sentí, Andreu. 2015. Modal Verbs, Future and Grammaticalization in Old Catalan. A Cognitive Approach. *Catalan Journal of Linguistics* 14: 179–198.

Servei Lingüístic de la Universitat Oberta de Catalunya. 2016. *Guia pràctica de català*. Barcelona: Fundació per a la Universitat Oberta de Catalunya.

Sinner, Carsten. 2008. Spanish and Catalan in Contact: Orality and Informal Contexts. *Oihenart* 23: 521–543.

Song, Yan-Yan & Lu, Ying. 2015. Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry* 27(2): 130–135.

Stokes, Craig R. 2011. Castilian Transcodic Markers in Internet Catalan: Analysis of Generic, Regional and Linguistic Factors. PhD dissertation. University at Albany, State University of New York.

Stokes, Craig R. 2015. The Use of Catalan Verbal Periphrases *haver de* and *tenir que* on Twitter. *Sociolinguistic Studies* 9(4): 445–466.

Turell Julià, Maria T. 1995. The "variationist" view-point of variation: evidence from Catalan-speaking communities. *Catalan Review* 9(2): 275–290.

Ubalde, Josep. 2013. Language Attitude Adoption: A Cross-sectional Study on Attitudes Towards Catalan among Adolescents from Catalonia and La Franja. *Treballs de Sociolingüística Catalana* 23: 287–302.

Vila, F. Xavier, Ubalde, Josep, Bretxa, Vanessa & Comajoan-Colomé, Llorenç. 2020. Changes in language use with peers during adolescence: a longitudinal study in Catalonia. *International Journal of Bilingual Education and Bilingualism* 23(9): 1158–1173.