Silvia Ballarè* and Caterina Mauri

Subjunctive/indicative alternation with *verba putandi* in spoken Italian

https://doi.org/10.1515/flin-2024-0053
Received November 19, 2024; accepted May 14, 2026; published online June 22, 2026

Abstract: This article investigates the alternation between indicative and subjunctive moods in complement clauses introduced by *verba putandi* (e.g. 'think' or 'believe') in contemporary spoken Italian. Drawing on spontaneous conversations from the KIParla corpus, the study tests whether mood selection is governed by degrees of epistemic certainty, by extralinguistic factors, or by linguistic properties internal to the construction. Multivariate statistical models evaluate the contribution of several predictors, including the grammatical person and tense of the subordinate verb, its adjacency to the complementizer, the lemma involved, register, and speakers' educational background. The results show that mood alternation is not primarily driven by epistemic stance or education level, but by usage-based and structural parameters: third-person, present-tense, and adjacent clauses strongly favour the subjunctive, while first- and second-person contexts, past tense, and non-adjacency favour the indicative. Register exerts a measurable influence, whereas educational background does not. Overall, the findings challenge traditional evidential accounts of mood in Italian and suggest that the indicative/subjunctive alternation in spoken language is best understood as a gradient, probabilistic phenomenon shaped by frequency patterns, morphosyntactic dependency, and register variation.

Keywords: mood alternation; subjunctive; spoken Italian; register variation; language variation

1 Introduction: the subjunctive/indicative alternation

1.1 Aims and scope

The alternation between subjunctive and indicative mood, in contexts where prescriptive grammars allow for mood alternation, has received attention from scholars

***Corresponding author: Silvia Ballarè**, University of Bologna, Bologna, Italy,
E-mail: silvia.ballare@unibo.it
Caterina Mauri, University of Bologna, Bologna, Italy, E-mail: caterina.mauri@unibo.it

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

interested in understanding the limits of speakers' freedom in mood selection, be they connected to epistemic stance, processing efficiency, linguistic competence or extralinguistic aspects. Despite the great number of studies addressing this issue, the explanation is still controversial.[1]

In this paper, our aim is to investigate the alternation between indicative and subjunctive moods in spoken Italian, based on corpus data, and to identify the factors determining the speakers' choice for one mood or the other. The analysis focuses on finite completive clauses introduced by *verba putandi* like *pensare* 'to think' or *credere* 'to believe' at present tense: in such syntactic contexts, standard Italian indeed allows the verb in the completive clause to be inflected in the subjunctive (1) or in the indicative mood (2).

(1) Italian

a. ***penso*** ***che*** *ci* ***sia*** *una* *sola*
think:IND.1SG COMP LOC be:SBJV.3SG INDEF:F.SG only:F.SG
Francesca *nella* *mia* *rubrica*
Francesca in:DEF.F.SG my:F.SG contacts
'I think that there is just one Francesca in my address book' (KIParla, BOA3020)

b. ***credo*** ***che*** *i* *miei* *dati* ***rappresentino***
believe:IND.1SG COMP DEF.M.PL POSS:M.1PL data represent:SBJV.3PL
sufficientemente *la* *realtà*
adequately DEF:F.SG reality
'I believe that my data adequately represent reality' (KIParla, BOD1003)

(2) Italian

a. ***penso*** ***che*** ***può*** *essere* *utile*
think:IND.1SG COMP can:IND.3SG be:INF useful
'I think that it can be useful' (KIParla, PTB022)

b. ***credo*** ***che*** *forse* ***ci*** ***sono*** *più* *persone*
believe:IND.1SG COMP maybe LOC be:IND.3PL more people
che *si* *muovono*
REL REFL move:IND.3PL
'I believe that maybe there are more people travelling' (KIParla, TOD2001)

1 This paper is the result of a continuous collaboration between the two authors. For the purposes of Italian academia, Silvia Ballarè is responsible for Sections 1.3 and 2, Caterina Mauri is responsible for Sections 1.2 and 3. The two authors are jointly responsible for Sections 1.1 and 4. We would like to thank two anonymous reviewers for their insightful comments, which helped us make the argumentation and the methodology sounder.

Two different types of explanations have been put forward in the literature. According to traditional grammars (Serianni 2010 [1989] § XIV, 48), mood selection depends on the degree of certainty that the speaker wants to express over the propositional content contained in the completive clause. The indicative is to be preferred in context in which the speaker is sure about the conveyed content, while the subjunctive shows a higher degree of uncertainty. On the other hand, it has been argued that the subjunctive/indicative alternation correlates with extralinguistic factors, both connected to the communicative situation and to the social profile of speakers. According to Berruto (2012), indicative (vs. subjunctive) is indeed more frequent in interactions with a lower degree of formality and/or in which speakers with lower educational achievements are involved. As is well known, linguistic and extralinguistic factors may coexist, and the presence of the latter does not exclude that of the first (see the discussion below).

Yet, evidence for these hypotheses was mostly based on qualitative analyses, which made little or no use of corpus data. In this paper, we aim to look for such evidence and observe the subjunctive/indicative alternation in completive clauses in actual language use, based on data collected between 2017 and 2024, accessible through the KIParla corpus of spoken Italian (Mauri et al. 2019). By examining spontaneous interactions between speakers with different social characteristics, we will identify the linguistic and extralinguistic factors that may explain the alternation between indicative and subjunctive in these constructions, and we will measure their respective importance. More specifically, through the analysis, we aim to answer the following research questions:

a) What is the role played by the speaker's confidence or certainty regarding the content of the completive clause?
b) To what extent do extralinguistic factors play a role in mood selection?
c) Are there other linguistic factors that should be considered to explain the alternation between indicative and subjunctive?

After discussing the main approaches and explanations to the subjunctive/indicative alternation as attested across different languages (Section 1.2), we will focus on the Italian situation (Section 1.3). Section 2 is devoted to the methodological choices and the empirical ground of the research: the corpus of spoken data employed for this study is described and the parameters and the tools of analysis are explained. Section 3 discusses the results and shows that the expectations that we had at the beginning of this study, based on the existing literature, are to be debunked. In Section 4 we will discuss the relevance of these results for an overall account of subjunctive/indicative alternation, and a general theory of mood selection.

1.2 The subjunctive/indicative alternation

The relation between the use of subjunctive and the speaker's epistemic stance has been argued to be significant across languages. Palmer (2001) and Givón (2001) propose that the distinction between indicative and subjunctive moods can be interpreted with reference to the realis-irrealis distinction (cf. Mauri and Sansò 2016), positioning the subjunctive within the epistemic-evidential section of the broader "non-actualization" domain. When a language allows for an alternation between subjunctive and indicative mood in the same syntactic context, Givón (2001: 313) argues that the choice of one mood over the other results in some difference in meaning. Considering the case of Spanish and Italian, he claims that the use of subjunctive instead of indicative mood encodes a different epistemic stance by the speaker, as in (3): the choice of the indicative mood in (3a) reveals less surprise than in (3b), the context in which the subjunctive is instead selected to convey a mirative stance (2001: 322).

(3) Spanish (adapted from Givón 2001: 322)

a. *Lo* *increíble* *era* *que* *Pedro* *no* *lo* ***sabía***
DEF:M.SG incredible was COMP Pedro NEG it know:IND.IPFV:3SG
'The incredible thing was that Pedro didn't know it.'

b. *Lo* *increíble* *era* *que* *Pedro* *no* *lo* ***supiera***
DEF:M.SG incredible was COMP Pedro NEG it know:SBJV.IPFV:3SG
'The incredible thing was that Pedro *should* not know it.'

In this paper, we are concerned with the subjunctive-indicative alternation in complementation. Let us consider Spanish again: as shown in (4), the indicative is used with verbs like *saber* 'to know' (4a), which imply that the speaker is certain about the situation, but the subjunctive is employed with verbs like *dudar* 'to doubt' (4b), which encode uncertainty. A similar opposition can be observed in example (5) from Russian, where the complement clause following the verb *govorit'* 'to say' shows the indicative (5a), while the subjunctive has be to selected after *somnevat'sya* 'to doubt' (5b).

(4) Spanish (adapted from Palmer 2001: 118)

a. *Yo* *sabía* *que* *el* ***estaba*** *ahí*
I know:PST.1SG COMP he be:IND.PST:3SG here
'I knew that he was here.'

b. *Dudo que* ***aprenda***
doubt:1SG COMP learn:SBJV.PRS:3SG
'I doubt that he's learning.'

(5) Russian (adapted from Noonan 2007: 107)
a. *Ja govorju, čto Boris* ***pridët***
I say:1SG COMP Boris come:IND.FUT:3SG
'I say that Boris will come.'
b. *Ja somnevajus', čtoby Boris* ***prišël***
I doubt:1SG COMP Boris come:SBJV:3SG
'I doubt that Boris will come / came.'

Noonan (2007: 102) proposes that the subjunctive can only be fully characterised when viewed in contrast with the indicative. If we take the indicative to represent the mood most closely associated with prototypical declarative clauses, then the term *subjunctive* functions as a general designation for any mood distinction in complement clauses that does not fall within the indicative domain, and that typically involves dedicated inflectional forms or stem alternations.

Within this framework, Noonan suggests that the defining property of the subjunctive lies in signalling that the interpretation of the subordinate clause is not autonomous: some component of its meaning is anchored to, or shaped by, information supplied by the main clause. This interpretive dependency can manifest itself in different domains. It may concern temporal anchoring between the two clauses, an epistemic relation expressing how the speaker positions the embedded content with respect to knowledge or belief (as illustrated in (4) and (5)), or a pragmatic relationship in which the embedded proposition is treated as presupposed or backgrounded. The latter is exemplified in (6), where the use of the subjunctive conveys that the content of the complement is taken for granted, and therefore presented as certain within the discourse.

(6) Spanish (Noonan 2007: 109)
Lamento que Juan salga esta noche
regret:1SG COMP John leave:SBJV:3SG this night
'I regret that John will leave tonight'

In many languages, subordinate complement clauses with subjunctive verbs alternate with infinitival constructions. As noted by Nikolaeva (2007, cf. also Timberlake 2007: 326), when used in complement position, the two forms often convey a comparable meaning, typically associated with an irrealis or projected future event.

Languages differ considerably, however, in the extent to which particular predicates admit this alternation and in the specific properties that condition it. Italian provides a clear illustration of this cross-linguistic diversity: verbs of saying, for example, may take either a complement clause introduced by *che* with a subjunctive verb or an infinitival clause, as shown in (7).

(7) Italian

a. *Digli* *di* *andare* *a* *casa*
tell:IMP.2SG:3SG.DAT to go:INF to home
'tell him to go home.'

b. *Digli* *che* *vada* *a* *casa*
tell.IMP.2SG:3SG.DAT COMP go:SBJV:3SG to home
'tell him that he should go home.'

1.3 Subjunctive/indicative alternation in Italian

In Italian, the subjunctive is mainly employed in subordinate clauses, where it alternates with the indicative. Its selection has been argued to be linked with an array of linguistic factors, and especially with semantics. More specifically, subjunctive can be selected in order to modulate the epistemic commitment of the speaker or to underline their role as a non-source of the information conveyed in the subordinate clause (Renzi 2019; Squartini 2008, 2010). For example, the verb *pensare* 'to think' (especially if at 1st person) is said to be followed by an indicative verb when the propositional content of the subordinate clause is considered to be (more) certain by the speaker (see examples (1) and (2)). On the other hand, a verb like *sostenere* 'to claim' (especially if at 3rd person) can be followed by the subjunctive if the speaker wants to distance themselves from the content of the subordinate clause, as in (8a) (Renzi 2019: 21) versus (8b).

(8) a. CORIS,[2] journalistic prose
L' *obiettivo?* *Qualcuno* *sostiene* *che* *sia* *ambiziosissimo,*
DEF:M.SG goal someone assert:IND.3SG COMP be:SBJV:3SG ambitious:SUP:M.SG
ma *chiaro:* *la* *conquista* *del* *mondo*
but clear DEF:F.SG conquest of:DEF.M.SG world
'The goal? Someone claims that it is highly ambitious but very clear: the conquest of the world.'

2 Tamburini 2021.

b. UniverS-ITA corpus,[3] 0564
Alcuni sostengono che, attraverso la didattica a distanza,
some assert:IND.3PL COMP through DEF:F.SG distance learning
gli studenti hanno la possibilità di svolgere
DEF:M.PL students have:IND.3PL DEF:F.SG chance of conduct:INF
più lavoro in autonomia
more work in autonomy
'Some claim that, through distance learning, students have opportunities to carry out more work autonomously.'

The use of subjunctive in completive clauses has also been described as expressing the speaker's stance or attitude towards the propositional content. Examples like (9a) versus (9b) show the use of subjunctive after evaluative or emotive predicates like *regret*, *be surprised*, *be glad* (cf. also (6) above).

(9) a. UniverS-ITA corpus, 0992
Sono contento che oggi ci sia stato questo
be:IND.1SG happy COMP today LOC AUX.SBJV:3SG be:PST.PTCP this
progetto
project
'I am happy that today there was this project'

b. ItTenTen20, 414457
Per Gaspare: sono contento che è tutto ok
for G. be:IND.1SG happy COMP be.IND.3SG everything ok
'For Gaspare: I am happy that everything is okay"

In these instances, while the situation described by the subjunctive verb form is factual, it can be contended that the speaker's intention is not to report the fact but to provide an evaluation of it (Palmer 1986: 119). Nordström (2010: 42) suggests that these scenarios align with an analysis of the indicative-subjunctive distinction based on the speaker's propositional attitude towards the proposition's factuality. The subjunctive form may also indicate that the proposition's factuality does not align with the speaker's expectations, serving a mirative function (DeLancey 1997, cf. also (3b)).

Recent studies have shed light on the role of other linguistic factors at play in the mood selection. Subjunctive is undergoing a grammaticalization process, involving "lexical routinization […] and structural conventionalization" (Poplack et al. 2018: 217).[4] More precisely, in relation to Italian, it has been observed that the selection of the subjunctive is not linked to semantic factors (such as the presence of irrealis

3 Grandi et al. 2023.

4 See also Poplack et al. (2013) on French and Torres Cacoullos et al. (2017) on Spanish.

matrices, of utterances featuring explicit indicators of non-factual modality, and the semantic class of the governor); what is important, instead, is the lexical identity of the governor (Poplack et al. 2018). Digesto (2019, 2021), thanks to the adoption of an empirical methodology, showed that, in current uses, subjunctive can be triggered by lexical and morphosyntactic factors: the mood is in fact preferred when some specific lexical governors are employed in the main clause or when *essere* 'to be' is the inflected verb of the subordinate one.

Furthermore, in subordinate clauses introduced by a factual governor, the information structure of the sentence becomes relevant: as also argued by Noonan (2007), subjunctive is preferred when the subordinate clause conveys a thematic content (Ballarè 2025; Ballarè and Cerruti 2023), as in (10) (cf. (6) above from Spanish). In the example, the subordinate clause introduced by *il fatto che* 'the fact that' conveys thematic content ("Turin is no longer an industrial city") and displays a verb in the subjunctive mood. The predicative content, which is more relevant from an informational point of view, is located in the main clause ("(it) has changed everything") and shows a verb in the indicative mood.

(10) PTD004, KIParla, (see Ballarè and Cerruti 2023: 89)

sì	e	poi	sicuramente	il	fatto	che	torino	non	
yes	and	then	for sure	DEF.M.SG	fact	COMP	Turin	NEG	
sia	più	una	città	industriale	ha				
be:SBJV:3SG	anymore	INDEF.F.SG	city	industrial	AUX:IND:3SG				
cambiato	tutto								
change:PST.PTCP	everything								

'Yes and then for sure the fact that Turin is no longer an industrial city has changed everything'

The scrutinized mood alternation has been largely investigated from a sociolinguistic perspective, and indicative mood, when in alternation with subjunctive, is considered to be (more) typical of sub-standard varieties. It is furthermore highly stigmatized and is more prone to be employed in informal styles and in the productions of speakers with lower educational achievements (see Schneider 1999; Lombardi Vallauri 2003; Berruto 2012; Prandi 2012 *inter al.*). Conversely, subjunctive is more typical of formal styles and written texts. However, it is important to note that most of these hypotheses have been formulated at a time when there was little or no access to (spoken) data. Recent and more empirically based studies (see Ballarè and Cerruti 2023; Digesto 2019, 2021), unexpectedly, reduced the weight of extralinguistic factors in the mood. In fact, as discussed above, linguistic factors have proved to be more useful in explaining mood alternation.

Starting from the hypotheses put forward in the literature, in this paper we aim to examine spontaneous Italian spoken data to compare the respective weight of extralinguistic and linguistic factors at play in the subjunctive/indicative alternation, focusing on completive clauses after verbs of cognition. It is interesting to note that, in these cases (i.e. with *pensare* 'to think' and *credere* 'to believe', see below), previous studies have shown considerable variability: the subjunctive, for instance in the data analyzed by Digesto (2019), is used in approximately 70 % of cases.

2 Data and methods

In this section, we provide an overview of the data used in the study and we describe the methods adopted for data extraction, annotation and analysis. Our study is based on the KIParla corpus (Mauri et al. 2019), which is a corpus of spoken Italian, and is composed of 4 modules: KIP, ParlaTO, ParlaBO and KIPasti. One of the most important characteristics of the KIParla corpus is that it gives access to a wide set of metadata and, thus, allows for sociolinguistic investigations.

The first one (KIP: 121 interactions, 69:23:08 hours of recording, 581.784 tokens) is a collection of interactions recorded in the academic setting in Bologna and Turin between 2016 and 2019, involving professors and students; it has been built in order to maximize register variation, in that it includes 5 diverse types of situation (fixed vs. free topic, symmetric vs. asymmetric relations between speakers): free conversations, semi-structured interviews, office hours, exams and lessons. The ParlaTO (67 interactions, 48:51:14 hours of recording, 561.388 tokens) and ParlaBO (85 interactions, 65:43:25 hours of recording, 701.354 tokens) modules consist of semi-structured interviews recorded in Turin between 2018 and 2020 and in Bologna between 2021 and 2024, respectively. In order to build the modules and maximize social differentiation, diverse kinds of speakers were involved: they have different ages (from 16 years old to over85), have different educational achievements (from elementary school diploma to PhD), and different occupations (intellectual, laborer, retailers, …).

Finally, the KIPasti module is a collection of kitchen-table conversations recorded in 13 Italian regions between 2020 and 2024 (63 interactions, 42:36:11 hours of recording, 482.887 tokens). The module was balanced according to the population size of Northern, Central and Southern Italy, and diverse kinds of speakers were recorded, maximizing age and education differentiation, always keeping the absolute ecology of the situation: no ad hoc meeting has been organized, but every recorded meal would have taken place even if it had not been recorded (cf. *naturally-occurring* data, Troiani et al. 2024).

Table 1: Dataset.

Lexical governor	Occurrences
Pensare ('to think')	455
Credere ('to believe')	222
	Tot. 677

We extracted all the inflected forms of the two verbs that are more frequently used in the *putandi constructions*, i.e. *pensare* 'to think' and *credere* 'to believe', in the present indicative and followed by *che* 'that': [*pensare/credere*$_{\_\text{PRS.IND}}$ che [*clause*]]. Then, we manually removed all the cases in which the subordinate clause did not allow for mood alternation (i.e. cases in which the verb of the subordinate clause was in the future tense or the conditional mood) and cases in which it was not possible to decide whether the verb was inflected at the indicative or at the subjunctive due to morphological syncretism (i.e. first person plural). The total number of occurrences is shown in Table 1.

The decision to limit the analysis to these two verbs only is related to two further considerations. The first concerns their high frequency of use, which makes them sociolinguistically (more) neutral, unlike other variants – such as, for instance, *ritenere* 'to consider (something) to be true'- which are considerably rarer, as they typically occur in formal speech produced exclusively by educated speakers (e.g., *ritenere* occurs only once in KIPasti and not in a complement clause, and is entirely absent from the spontaneous conversations in KIP). Furthermore, including other governors (which tend to be rarer and associated with more formal speech by highly educated speakers) would introduce collinearity between governor, register, and educational level. Focusing on only two governors allows for an analysis that avoids overlap among these factors. Finally, the last reason pertains to verb semantics: *pensare* and *credere* are semantically largely overlapping, whereas other verbs that can still be classified as *verba putandi* (e.g., *immaginare* 'to imagine', *stimare* 'to estimate') may convey distinct shades of meaning. Further studies that broaden the scope of the investigation are nonetheless desirable, as they may help capture a potential lexical routinization of the subjunctive.

All the extracted occurrences were manually annotated according to the mood of the verb of the subordinate clause (subjunctive vs. indicative) and nine parameters (i–ix): i and ii are extralinguistic in nature, while iii–ix refer to morphosyntactic, lexical, and pragmatic features relevant for the *putandi* constructions. Let us start with the extralinguistic aspects.

I. Interactional context (formality): lessons, exams and office hours (h), semi-structured interviews (m); free conversations and kitchen table conversations (l).

This parameter deserves a detailed discussion, since we introduce a classification that was not proposed in the literature before. The classification between low (l), medium (m), and high (h) formality has been made following the well-known distinction between *transactional* and *personal* interactions (Gumperz 1964).[5] The first ones refer to cases in which there is a socially defined goal, and the main aim is to transfer informational content (see also Brown and Yule 1983); for this reason, participants "in a sense suspend their individuality in order to act out the rights and obligations of relevant statuses" (Gumperz 1964: 149). This is clearly the case in lessons, office hours and exams, where professors and students communicate playing their social role, which, in addition, places them in an asymmetric relation, in terms of status and relational power. The topic is rather fixed, with perhaps the partial exception of office hours, where a small range of subjects can be found (for example, discussions about students' dissertations, requests for information about university curricula, …). Given the social characteristics of these interactional contexts, it is here that speakers select higher registers (classified as (h), high formality).

At the opposite end, there are personal interactions, where "participants act as individuals, rather than for the sake of specific social tasks" (Gumperz 1964: 149), as typically occurs within peers. In these exchanges, what is of greater relevance is the interpersonal relationship between individuals, which allows the communication of personal thoughts and experiences. The register selected, in this case, falls at the lower end of the scale (classified as (l), low formality). For our analysis, we decided to consider free conversations and kitchen-table conversations as personal interactions. In these cases, speakers involved are always close members of a certain social network: for example, there are recordings of family meals or conversations between flatmates or friends. The relationship between the speakers is more symmetrical, and the interactional roles are less rigidly codified by the context.

Between these two poles, we decided to insert semi-structured interviews. These interviews were collected following the traditional Labovian methodology (Labov 1984; Tagliamonte 2006: 37–49), in which the interviewer had structured sets of questions about different topics (community and neighborhood, if and how the city had changed over time, local traditions, hobbies, and interests, …) and their main goal was to elicit narratives of personal experiences, following the informant interest

5 See also *referential* versus *emotive* (Jakobson 1960) and *ideational* versus *impersonal* (Halliday 1970).

or will to talk about certain topics. In other words, the selection of the topic is, at least to some extent, guided by the interviewer. In these interviews, there is a transfer of informational content, but also personal views and opinions are shared. Moreover, to elicit less controlled data, multiple informants (friends, couples etc.) were often involved simultaneously. The data collection was designed to record informal exchanges, but it is important to stress that these interactions maintain a rather rigid structure in which the speakers' roles are codified (one asks questions, the other(s) respond), one of them acts according to their role of interviewer while the other is at least expected to act as individual, sharing their thoughts and opinions. It is also important to consider that in the interviews present in the KIParla corpus, interviewers and interviewees have diverse relationships: sometimes they are people who know each other well (for example, students interviewing relatives and friends), while in other cases, that was not the case (for example, students/researchers interviewing members of local associations). In this latter scenario, an intermediary was often used to introduce the two people, and they also took part in the interaction. For the purposes of this study, we have chosen to place these interactions at an intermediate level between the *transactional* and *personal* ones (classified as (m), medium formality).

The classification and the relevant parameters are summarized in Table 2; note that the values in the table must obviously be considered as gradable and not as discrete categories.

II. Speaker educational achievement (education): low (l); medium (m); high (h).

The second extralinguistic parameter concerns speakers' education. To account for the expansion of schooling observed in Italy over the last few decades, we consider both speakers' age range and their highest educational degree, correlating the two factors when assigning the level of educational achievement. Over time, the substantial increase in university enrolments and graduates has meant that secondary-

Table 2: Interactional contexts.

	Kind of interaction	**Topic**	**Speakers' relation**	**Roles' codification**	**Values**
Office hours, exams, lessons	Transactional	Fixed	Asymmetric	Strict	H
Semi-structured interviews	Personal/transactional	(More) fixed	Symmetric	(More) strict	M
Free conversations, kitchen table conversations	Personal	Free	Symmetric	Loose	L

Table 3: Education/age range.

	Elementary school diploma	Middle school diploma	Technical or professional school diploma	High school diploma	BA/MA degree	Post-graduate degree
Young (under 30 years)	L	L	L	M	H	H
Adult (31–60 years)	L	L	M	M	H	H
Senior (over 61 years)	L	L	M	H	H	H

school diplomas, especially technical and vocational ones, have gradually lost part of their selective value, including in the labour market. For this reason, the same diploma may correspond to different levels of educational achievement depending on whether the speaker is young, adult, or senior.

The values low (L), medium (M), and high (H) were therefore assigned on the basis of two types of metadata, as shown in Table 3, namely age range and educational degree, following the sequential hierarchy of educational qualifications within the Italian school system. The extreme cases receive a uniform classification: elementary and middle-school diplomas are classified as L across all age ranges, whereas BA/MA degrees and postgraduate qualifications are classified as H across all age ranges. Secondary-school qualifications, by contrast, are assigned different values depending on speakers' age. A technical or vocational diploma is classified as L for young speakers (under 30) and as M for all others; a general high-school diploma is classified as H for senior speakers (over 61) and as M for all others. In this way, the classification system aims to reflect changes in access to education across the Italian population over the past century.

Moving now to the linguistic parameters, we considered several features that could have an effect on modal selection.

III. Lemma of the main clause verb (v1_lemma): *pensare* ('to think') or *credere* ('to believe').

We considered the verb of the main clause (v1), since it could possibly be linked with lexical routinization, that is to be intended as the process by which specific lexical items (in this case, the governors) become strongly associated with particular grammatical forms (in this case, the mood of the subordinate clause verb) regardless of their semantics (Poplack et a. 2018). However, it should be noted that, as already discussed, only two governors are considered in this analysis and therefore further studies would be desirable to verify the impact of lexical items from a larger sample.

IV. Tense of the subordinate clause verb (v2_tense_morp): present (pres), past (pas).

v. Subordinate clause verb person (v2_person): 1st; 2nd; 3rd.
vi. Lemma of the subordinate clause verb (v2_essere): *essere* 'to be' (yes), other (no).

We also took into account some characteristics of the subordinate clause verb, such as tense and person. More specifically, in order to monitor morphological features, all cases in which the inflectional component of the verb was conjugated in the present tense were considered *present*, thus including cases of compound pasts in which the auxiliary is conjugated in the present tense; all cases in which the inflectional component of the verb was conjugated in the past tense were annotated as *past*, thus including simple pasts and compound pasts in which the auxiliary is conjugated in the past tense. The potential relevance of determined time reference (see Section 1.2) was not considered in the analysis, as only contexts in which the main clause verb appears in the present tense were extracted. As a result, it would not have been possible to assess the impact of this parameter in a rigorous manner, but only partially. This aspect may be addressed in future research.

The person of the subordinate clause verb was included to assess whether mood selection is affected by the referential anchoring of V2. In particular, 1st- and 2nd-person forms identify one of the speech-act participants, while 3rd-person forms refer to an external participant or situation; this distinction may interact with the morphological opposition between indicative and subjunctive.

Finally, we coded whether v2 coincides (or not) with *essere* 'to be'; the lemma *essere* indeed shows a significantly higher frequency than all the other verbs in the dataset, which may likely lead to routinized constructions, and it has been argued to play a role in favoring the use of subjunctive (cf. Digesto 2019, 2021).

vii. Phrasal adjacency between *che* ('that') and the subordinate clause verb (adjacency): adjacent (adj); non-adjacent (non-adj).

Furthermore, we considered the adjacency between the complementizer and the subordinate clause verb. In coding adjacency, we considered to be *adjacent* also all the cases in which clitics and/or negative markers are inserted between *che* ('that') and the subordinate clause verb, as shown in (11).

(11) KIParla, BOA3003

penso	*che*	*lo*	*conosca*
think:ind.1sg	comp	3sg.obj:m	know:sbjv:3sg

'I think he/she knows him'

viii. Same subject (same_subj): same subject (ss), different subject (ds).

According to grammars, when the subject of the main clause and the complement clause is the same, the choice of non-finite verbs in the dependent clause is more common, as in (12), but in colloquial Italian also the indicative may be selected (Serianni 2010 [1989] § XIV, 6), as in (13). Real corpus data actually show that speakers/writers may employ the subjunctive even in these cases, as shown in (14) even if they are quite rare. For this reason, we decided to consider this feature and tag whether the main clause and the subordinate clause had the same subject or not.

(12) KIParla, PTB018
non credo di essere stata forse in tema
NEG believe:IND.1SG of be:INF be:PST.PTCP maybe in topic
'I do not think I was maybe on topic'

(13) KIParla, TOD2007
penso che forse anch'io a volte faccio
think:IND.1SG COMP maybe also I sometimes do:IND.1SG
qualcosa che dà fastidio agli altri
something REL give:IND.3SG annoyance to:DEF.M.PL other:PL
'I think that maybe sometimes I do things that annoy others, too.'

(14) itTenTen20, 1438413
Per il resto penso che io sia più annoiata
for the rest think:IND.1SG COMP I be:SBJV:1SG more bored
'For the rest, I think that I am more bored'

IX. ASSERTIVE ILLOCUTION (assertive_ill): yes, no.

To monitor the speaker's stance more directly, we distinguished between cases in which the main clause has an assertive illocution and others, such as questions and/or dubitative clauses, characterized by hedging elements that could weaken the utterance assertivity and/or discourse markers conveying uncertainty, as in (15) – and negative clauses (see Digesto 2021; Serianni 2006, 2010 [1989] § 14, 52). In some studies on languages similar to Italian (see, e.g., Torres Cacoullos et al. 2017 on Spanish), sentence polarity has been shown to be a relevant parameter for the selection of mood in subordinate clauses. However, preliminary investigations on this dataset, as well as comparable studies on Italian (Ballarè and Cerruti 2023), have yielded opposite results. For this reason, the present analysis classifies sentences not according to their polarity but rather according to their degree of assertiveness, distinguishing between genuinely assertive sentences and those exhibiting a different illocutionary force, expressed through the presence of negation markers and/or other adverbial or discourse elements conveying uncertainty.

(15) KIParla, PBA005
sinceramente non lo so forse *penso che fosse*
honestly NEG 3SG.OBJ:M know:IND.1SG perhaps think:IND.1SG COMP be:SBJV:3SG
una sorta di gara tra le famiglie di bologna
a sort of competition between DEF:F.PL family:PL of Bologna
per dimostrare il potere.
to show DEF power
'Honestly, I don't know, maybe I think it was some sort of competition between the families of Bologna to show their power.'

The data were then analyzed adopting a statistical tool that is useful in cases in which there is a small sample but a rather large number of parameters, i.e. conditional inference tree and random forest (Levshina 2015; Tagliamonte and Baayen 2012).

A conditional inference tree is a flowchart-like model built by using statistical hypothesis testing to select splits. It evaluates the association between each feature and the target using *p*-values, only selecting significant splits. The tree grows by recursively partitioning the data, ensuring each split is statistically significant. In other words, each split occurs under a specific set of linguistic and/or extralinguistic circumstances, revealing what factor plays a role over mood selection.

Random forests consist of several inference trees, and the output is a ranking of the importance of each parameter, and it allows us to discuss if and how the selected parameters play a role in the scrutinized mood alternation.

A single inference tree builds one path from the root to the leaves, focusing on partitions that maximize homogeneity for each branch split. This means that even parameters with moderate relevance can appear significant if they contribute to reducing impurity at crucial splits in the tree. In contrast, random forests average the importance of parameters over many trees, which reduces the influence of individual splits and emphasizes more consistent parameters across multiple trees. Random forests can capture interactions between variables better than a single tree because they aggregate results from multiple trees built on bootstrapped samples and subsets of features. A parameter that may seem less important in individual trees might contribute meaningfully in combination with others across the forest. The inference tree, however, doesn't always capture these interactions fully because it only has a single decision path for each combination.

Finally, random forests reduce variance through averaging, making them less sensitive to noisy or less important variables. As a result, parameters with only localized significance might appear important in an inference tree but less so in a random forest, where the importance score is averaged over many trees and focuses on features consistently important across samples. An inference tree may thus highlight parameters with localized or conditional importance that random forests filter out due to their

averaging approach and reduced variance. This explains why some parameters might appear significant in an inference tree but not in a random forest model.

The two statistical techniques outlined above – random forest and inference tree- are primarily of exploratory nature. Consequently, a logistic regression will also be performed (Gries 2019), with the aim of providing a clearer quantification of the predictors' effect on mood selection after *verba putandi*. A mixed effects model was not used because, unfortunately, it was not possible to consider the speaker as a random effect. In the future, it will certainly be necessary to take into account the possible weight of the individual speaker in mood selection.

All statistical analyses were conducted in R. Data were imported from Excel spreadsheets, and the following packages were used: *party* for conditional inference trees (*ctree*) and conditional random forests (*cforest*); Hmisc to assess the model fit using Somers' D statistics; *glm* for logistic regression.

3 Results: going beyond expectations

In this section, we discuss the results of our analysis, which partly confirm the expectations we had based on the literature, but mainly reveal new factors at work, leading us to explain mood alternation within a larger, and more complex, picture. In 3.1, we will identify and explain the specific factors at play, following the visualization offered by the conditional inference tree; in 3.2, we discuss the ranking within the random forest and in 3.3 the results of the logistic regression. Finally, in 3.4 we will present the parameters that resulted to be relevant in explaining variation.

3.1 Inference tree

Figure 1 shows the conditional inference tree, that was generated by the statistical analysis conducted on our data. The c-index of the model is 0.8 and thus, according to Hosmer and Lemeshow (2000: 162), has excellent discrimination.

The first feature that proves to be significant in order to split the data into internally homogeneous subsets, thus having relevance with respect to the choice of subjunctive versus indicative, is the PERSON OF THE SUBORDINATE CLAUSE VERB (V2_PERSON). Indeed, if the verb of the complement clause is conjugated in the first or second person, thus communicating an evaluation regarding the speech act participants, in most cases the speaker selects the indicative mood. Note that this is the only situation where the indicative is more frequent than the subjunctive (first and second column on the right, Node 10 and 11). However, this refers only to a small part of our dataset, given that V2 was *not* inflected at the third person only in 58 cases out of 677.

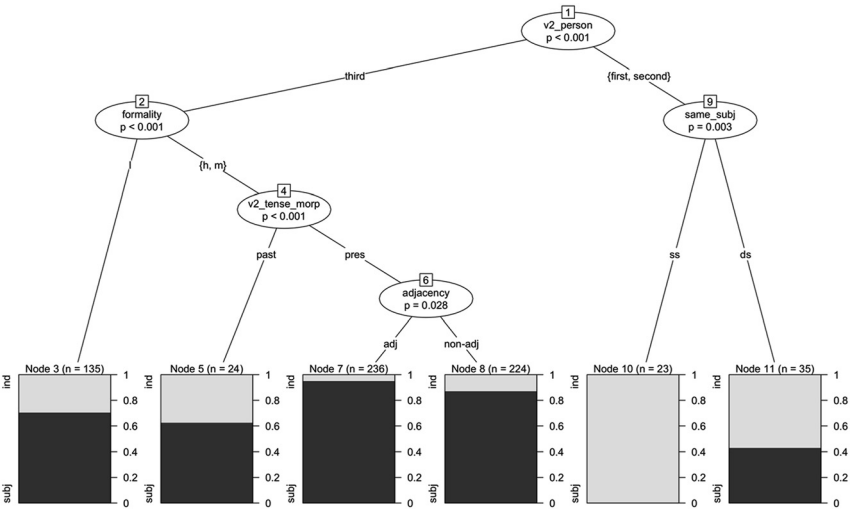

Figure 1: Inference tree of *putandi* constructions.

To account for the 58 occurrences in which V2 is conjugated in the 1st or 2nd person, the distinction between SAME/DIFFERENT SUBJECT between the main and subordinate clause becomes relevant. Almost half of these cases (23/58) indeed show the same subject for the main clause and the subordinate one: in these occurrences, the speakers' behavior aligns with the expectations outlined by standard grammatical descriptions, selecting the indicative mood (Node 10). On the other hand, if the main

Table 4: Same versus different subject – distribution.

	Ind	Subj	Tot.
Same subject	23 (100 %)	0 (0 %)	23
Different subject	110 (67 %)	544 (33 %)	654

The Fisher exact test statistic value is <0.00001. The result is significant at $p < 0.01$.

clause and the subordinate one have different subjects, there is more variability in mood selection (Node 11). Crucially, there are no instances of same-subject contexts when V2 is conjugated in the 3rd person. The distribution of the parameter in the dataset is shown in Table 4.

In general, we observe that same subject occurrences are extremely less frequent than different subject ones (23 vs. 654), which confirms the preference for non-finite verbs in same subject completive clauses. Furthermore, the absence of subjunctive in the few same subject occurrences shows that in a *putandi* construction, if the subject of V1 is the same as the subject of V2, V2 likely requires the indicative mood. Since variation is attested only for one of the two possible values, i.e., different subject, the predictor will be excluded in the next steps of the analysis.

When the subordinate clause verb is inflected at the third person, subjunctive is more frequent and other factors, both extralinguistic and linguistic, come into play. Overall, we observe that 3rd person for V2 is attested in 619 cases, which means 91 % of the occurrences. In 469 cases (69 %), it is associated to 1st person in the main clause (structure: $V1_{\_1}$ – $V2_{\_3}$), probably due to the evaluative nature of statements introduced by verbs of thinking, which typically convey some speaker's belief concerning a third party. The higher frequency of 3rd persons in V2 within *putandi* constructions may constitute in itself a factor that supports the use of subjunctive for these cases.

Italian is indeed presented as a case point in discussing the markedness of the subjunctive as against the indicative (Greenberg 2005: 46), because, in the subjunctive, forms for 1st and 2nd person singular are for many verbs identical to the form for 3rd person singular (e.g. *sia* '1st/2nd/3rd be.SUBJ'), while they differ in the indicative. Given the higher frequency of 3rd persons in V2 (619 out of 677 cases), speakers are likely to expect to find 3rd person in this type of constructions and, consequently, to interpret the use of a subjunctive form in the complement clause as referring to a third party. We may thus explain the wider use of indicative for the less frequent cases (i.e. 1st and 2nd person in V2) as a disambiguating strategy, chosen to avoid potential misunderstanding when expressing beliefs on the speech act participants. This hypothesis is further supported by the fact that, in most of the cases[6] in which V2 is inflected at 1st or 2nd person subjunctive, speakers indeed employ a personal pronoun to disambiguate verbal morphology, as in (16):

(16) KIParla, TOD2005

penso	*che*	***tu***	*lo*	*sappia*	*meglio*	*di*	*me*
think:IND.1SG	COMP	you	3SG.OBJ:M	know:SBJV.2SG	better	than	me

'I think you know it better than me'

6 Considering the possible cases of ambiguity (that is, sentences in which the subordinate verb is inflected in the subjunctive and in the singular, the subject of the main and subordinate clauses differ, and no impersonal constructions are involved), the pronoun is inserted in 7 out of 10 cases. In the remaining three occurrences, in two instances ambiguity does not arise because of a difference in gender or number between the subject of the main and the subordinate clause.

If we now focus on the occurrences where V2 is conjugated in 3rd person, the node that turns out to be relevant in Figure 1 is connected to the interactional context, that is, an extralinguistic factor. The parameter of FORMALITY has three values that were grouped in two by the algorithm: data associated with h (high, i.e. lessons, office hours, exams) and m (medium, i.e. semi-structured interviews) behave more uniformly compared to the ones annotated with l (low, i.e. free conversations and kitchen table conversations). Indeed, in more informal productions, the indicative is overall more frequent than in the other two types (cf. first column on the left, node 3). The low degree of formality alone leads to the higher frequency of indicative, as we can see from the fact that in the tree, given the value l (low), Node 2 (formality) is directly connected to Node 3 below. This means that, at least according to the results of the inference tree, in informal interactions there is less variability and no other linguistic parameters are useful to explain the mood alternation, showing that the existing variability does not seem to be related to linguistic factors. Instead, the algorithm has selected two additional parameters that are useful in explaining mood alternation in more formal productions (m, h), and these parameters are linguistic in nature.

In contexts of medium and high formality, V2 TENSE (node 4) acquires importance. Note that, in this case, priority was given to morphological features: we classified all the occurrences based on the tense in which the inflected component of the verb was conjugated (see Section 2). As can be seen in Figure 1, in the rare cases where the verb is inflected in the past tense (24 occurrences), the indicative is employed in almost half of the occurrences (node 5). On the other hand, when the verb is in the present tense, we observe a general neat preference for the subjunctive, and another parameter assumes relevance, namely the ADJACENCY between *che* 'that' and the verb of the subordinate. What the tree shows is that the choice of subjunctive mood is extremely likely in the combination of (i) 3rd person in V2, (ii) medium or high formality, (iii) present tense in V2 and (iii) adjacency between the main clause and the dependent verb lead to the use of subjunctive. On the contrary, if some phrasal element (like a lexically realized subject, or some adverbial element) separates the complementizer from the V2, we have more chances to find the indicative, although in a minority of cases.

The higher availability of indicative in non-adjacent cases may be due to the fact that the syntactic distance of V2 from the main clause weakens the dependency relation. One could argue that indicative may be selected, in case of phrasal distance, because the clause is not construed as dependent: the main clause behaves as a sort of an evidential discourse marker, equivalent to *secondo me* 'to me' (see Lo Baido 2024) and, as a consequence, the complement clause is not felt as such, but rather as a main clause characterized by inferential evidentiality. This is exemplified in (17),

where *penso che* is used to signal the inferential source of what the speaker is going to say, as shown by the pause (represented by the dot in brackets) and the reformulation discourse marker *cioè* 'I mean' following it.

(17) KIParla, TOD2009

anche	*il*	*fatto*	*di*	*aver*	*preso*	*la*	*malaria*
also	DEF.M.SG	fact	of	have:INF	take:PTCP.PST	DEF.SG:F	malaria
penso	*che*	*(.)*	*cioè*	*mi*	*ha*	*un*	*sacco*
think:IND.1SG	COMP		I.mean	1SG.DAT	have:IND.3SG	INDEF.SG.M	bag
cresciuto.							
grow:PTCP.PST							

'Even the fact that I got malaria, I think that (.) I mean, it made me grow a lot.'

Example (17) is particularly revealing of the pauses and possible disfluencies that characterize spoken interaction, which may favor the reorganization of the complement clause as a main utterance. The more V2 is distant from V1, the more it will be perceived as an independent verb, leading to the choice of the indicative mood.

3.2 Random forest

In this section, we aim to discuss the statistical distribution of the parameters, together with their overall importance, as revealed by the random forest technique (see Figure 2, the corresponding importance scores are reported in Table 5). The

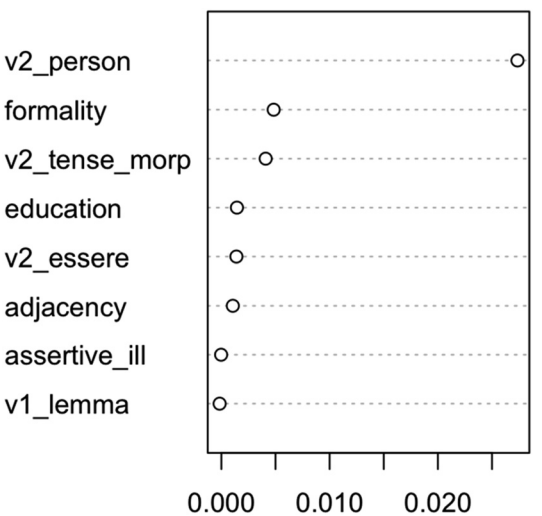

Figure 2: Random forest – *putandi* constructions.

c-index of the model is 0.9 and thus, according to Hosmer and Lemeshow (2000: 162), has outstanding discrimination. It is important to remark that inference trees and random forests can yield slightly different insights into parameter significance due to differences in their structure and mechanisms.

Figure 2 displays the parameters that we considered and their importance in explaining the mood alternation in the whole dataset. If we compare it to Figure 1, we observe a good level of correspondence, with the predictors that emerged as relevant in the inference tree ranking high also in the random forest. In the rest of this section, we will begin with the most important parameter according to Figure 2 and proceed toward the less important ones, presenting the statistical distributions associated with each parameter individually. Our aim is to discuss the direction in which each parameter may influence mood selection (indicative vs. subjunctive), either jointly or independently. For each parameter we provide a table showing absolute values, together with percentage values (per row) in parentheses. Even if the main goal of this analysis is to assess the impact of all the parameters together, we also provide the results of a chi-square or Fisher exact test to verify the statistical significance of the distribution of each parameter. Results are reported only when relevant.

The two most important factors in the forest are linguistic in nature. Let us start with V2_PERSON: the person of the subordinate clause verb is the most important parameter in the random forest. The distribution in Table 6, which is highly significant from a statistical point of view, confirms what was observed in the previous Section. With the first and second person, there is a neat preference for the indicative mood (79 % and 67 %, respectively), and only minor differences are observed between the two persons. On the other hand, if the verb is in the third person, the situation is reversed: the subjunctive mood is selected in the vast majority of cases (85 %).

Table 5: Random forest – numeric values.

Parameter	Value
v2_person	0.027
formality	0.005
v2_tense_morph	0.004
education	0.001
v2_essere	0.001
adjacency	0.001
assertive_ill	0.000
v1_lemma	0.000

Table 6: V2_person – distribution.

	Ind	Subj	Tot.
First	27 (79 %)	7 (21 %)	34
Second	16 (67 %)	8 (33 %)	24
Third	90 (15 %)	529 (85 %)	619

The chi-square statistic is 120.771. The *p*-value is <0.00001. The result is significant at $p < 0.01$.

Table 7: Formality – distribution.

	Ind	Subj	Tot.
High	6 (10 %)	56 (90 %)	62
Medium	69 (15 %)	390 (80 %)	459
Low	58 (37 %)	98 (60 %)	156

The chi-square statistic is 45.4666. The *p*-value is <0.00001. The result is significant at $p < 0.01$.

The second parameter that has importance in mood selection is an extralinguistic one, namely the degree of FORMALITY, which had already been identified as relevant in the inference tree. Table 7 shows the distribution of indicative and subjunctive across high, medium and low registers.

Whereas the distribution of V2 PERSON is rather unbalanced, with highly infrequent values (i.e. first and second person) correlating with the indicative mood, the three degrees of formality do not show extreme distributional differences. Interestingly, the interactions characterized by high (lectures, exams, office hours) and medium (semi-structured interviews) formality are more similar to each other than the highly informal contexts (free conversation and kitchen-table conversation). In the latter, more variability is indeed observed, while in higher registers the subjunctive is selected in 90 % and 80 % of the cases; as revealed by the inference tree, the small percentage of indicatives in higher registers is influenced by further linguistic factors, such as V2 tense morphology and adjacency, which, in these cases, acquires relevance (cf. Section 3.1).

V2 TENSE MORPHOLOGY occurs as fourth in the ranking of the random forest. The two values of this parameter show a very unbalanced distribution, as shown in Table 8, with morphological past occurring only in 39 cases out of 677 (5 %).

This ratio may be influenced by the fact that in our dataset the *putandi* verbs in V1 are inflected at present tense (see Section 2), and this tends to trigger completive clauses either inflected for present or for recent past (i.e. through forms where the auxiliary is a morphological present, e.g. *abbia fatto*). More remote past forms are not excluded but are rarer. Once again, as we observed for V2_person, the indicative mood is more likely for the infrequent value (past, in this case) than for the frequent one.

Table 8: Morphological tense – distribution.

	Ind	Subj	Tot.
Present	114 (18 %)	524 (82 %)	638
Past	19 (49 %)	20 (51 %)	39

The Fisher exact test statistic value is 0. The result is significant at $p < 0.01$.

3.3 Logistic regression

The logistic regression model reveals several statistically significant predictors, confirming the analysis discussed in Sections 3.1 and 3.2, while others do not appear to exert a meaningful effect. In order to interpret the meaning of the values, it should be specified that the indicative was selected as the baseline level. This means that parameters with negative coefficients are associated with the selection of the indicative, whereas parameters with positive coefficients are associated with the selection of the subjunctive.

Three predictors have a significant negative influence and thus trigger the selection of the indicative mood: a low degree of formality, the non-adjacency between *che* ('that') and the subordinate clause verb, and the assertivity of the illocution.

Conversely, several linguistic features display strong positive associations. The presence of present tense morphology significantly increases the odds of the

Table 9: Logistic regression.

Predictor	Beta	SE	*z*_value	*p*_value	Odds_Ratio	Significance
educationl	−0.134	0.356	−0.377	0.7059	0.874	
educationm	−0.194	0.328	−0.59	0.5554	0.824	
formalityl	−1.605	0.517	−3.103	0.0019	0.201	**
formalitym	−0.181	0.518	−0.35	0.7267	0.834	
v1_lemmapensare	−0.41	0.264	−1.555	0.1199	0.664	
v2_tense_morppres	1.59	0.406	3.916	0.0001	4.902	***
v2_personsecond	−0.258	0.772	−0.334	0.7382	0.772	
v2_personthird	1.759	0.604	2.911	0.0036	5.809	**
v2_essereyes	0.125	0.247	0.506	0.6131	1.133	
adjacencynon-adj	−0.795	0.242	−3.287	0.001	0.451	**
same_subjss	−17.44	764.773	−0.023	0.9818	0	
assertive_illyes	−0.865	0.273	−3.172	0.0015	0.421	**

outcome, while the use of third-person verb forms has the largest effect size, indicating a very strong predictive role.

Other variables, such as education level, the V1 lemma *pensare*, the V2 lemma *essere*, and second-person verb forms, do not reach statistical significance ($p > 0.1$), suggesting that they do not contribute meaningfully to the prediction within this model, once again confirming the results of the random forest.

Table 9 summarizes the results of the logistic regression model, reporting coefficients, standard errors, significance levels (** for p value <0.01; *** for p value <0.001), and odds ratios for each predictor.

Overall, the reduction in deviance from the null model (670.40) to the residual deviance (504.39) and the resulting AIC of 528.39 indicate that the model achieves a substantial improvement in explanatory power relative to the baseline. In other words, this decrease in deviance reflects the extent to which the included predictors succeed in accounting for the variability in mood selection, demonstrating that the final model provides a markedly better fit to the data than would be obtained by assuming no effect of any predictor. The significant predictors point to the critical role of formality, adjacency, assertive illocution, and specific morphosyntactic features (tense and person marking) in shaping the outcome under investigation.

3.4 Explaining variation

After discussing the results of the inference tree, random forest, and logistic regression models separately, we can now draw these strands together and provide an integrated account of the factors that shape mood alternation in the dataset.

Overall, the analysis highlights the weight of specific linguistic factors and register in explaining mood alternation. Both the logistic regression and the random forest models converge in identifying V2_person and formality as the most influential predictors in the dataset. In the random forest model, these variables exhibit the highest importance values, while in the logistic regression they emerge as statistically significant or strong contributors – specifically, *low formality* predicts the use of indicative mood and *third-person verb forms* predicts the use of subjunctive mood.

In particular, V2_person proves especially relevant. As discussed in Section 3.1, the preference for the indicative in first- and second-person contexts may reflect their overall lower frequency within *putandi* constructions, which typically express opinions about third parties. In Italian, the subjunctive form is identical across all three singular grammatical persons. Consequently, a singular subjunctive verb in V2 is more likely interpreted as referring to a third person. The use of the indicative for

first or second person may therefore serve a disambiguation function, clarifying the referential interpretation of the verb.

With respect to register, it is noteworthy that data from semi-structured interviews – situated between the two poles of formality – align more closely with the patterns observed in formal productions, whereas kitchen-table and free conversations display a distinct distribution. This suggests that the most relevant parameter for differentiating data along this variable is not the interpersonal relation between speakers (symmetric vs. asymmetric), but rather the degree of role codification (strict vs. loose) and the fixity of the topic (fixed vs. free). When the interactional setting requires speakers to adhere to codified roles and to discuss a predetermined topic, the subjunctive tends to occur more frequently. Conversely, in less structured, more spontaneous exchanges, speakers exhibit a greater tendency to use the indicative.

Furthermore, V2_tense_morphology emerges as another relevant factor, ranking third in the random forest and displaying a strong positive and significant coefficient for the present tense in the logistic regression. This pattern can be explained by considering the markedness hierarchy between the past and present tenses (Greenberg 2005: 48; Vellupilai 2016) as well as between the subjunctive and indicative moods (Dahl 1985; Greenberg 2005: 46). The data indicate that in contexts where speakers select a less frequent and more marked tense (i.e., the past), they tend to employ the mood that is generally more frequent and less marked (i.e., the indicative), which is thus more cognitively accessible.

According to inference tree and logistic regression, non-adjacency between V1 and V2 may also trigger the use of the indicative mood, due to a weaker dependency relation. This tendency is consistent with the nature of spoken discourse, which is characterized by pauses, interruptions, and disfluencies (cf. example (17)). The greater the distance between V1 and V2, the more V2 is perceived as an independent clause (cf. Bybee et al. 1994), leading to the selection of the mood typically associated with main clauses, namely the indicative. A similar effect may arise from a strong assertive illocutionary force, which favors the indicative as the preferred mood for expressing epistemic certainty.

Finally, it is worth noting that the analysis found no significant effect for speakers' educational background or for the lemma of the verbs in either the main or subordinate clause. These factors, therefore, do not appear to play a meaningful role in explaining the observed patterns of variation. However, the lack of a lexical effect for the main clause verb should be interpreted with caution, since the present analysis only includes two highly frequent and semantically similar governors (*pensare* 'to think' and *credere* 'to believe'). Future studies including a broader range of lexical governors may help to better assess the role of lexical routinization in mood selection.

Table 10: Frequency of V2_lemma in the dataset and frequency of subjunctive forms for each lemma.

Lemma_V2	Absolute frequency	Relative frequency	Absolute frequency of subj	Frequency of subj/tot. lemma
Essere 'to be'	**426**	63 %	366	86 %
Avere 'to have'	**70**	10 %	43	61 %
Potere 'can'	30	4 %	26	87 %
Fare 'to do'	23	3 %	15	65 %
Stare 'to stay'	9	1 %	5	56 %
Andare 'to go'	7	1 %	7	100 %
Dovere 'must'	6	1 %	1	17 %
Sapere 'to know'	6	1 %	5	83 %
Mancare 'to miss'	5	1 %	4	80 %
Venire 'to come'	4	1 %	2	50 %

It may be interesting to note that two factors found to be significant in the statistical analysis – namely, the person and the adjacency between *che* and the subordinate clause verb – seem to point to the development of a construction that favors the use of subjunctive, namely [*putandi*_VERB [*che* $V_{3SG.SUBJ}$]]. From this perspective, it may be relevant to consider the lemma of the subordinate clause verb, which was coded as *essere* 'to be' versus other. Despite its low importance, it is worth digging a little deeper into this distribution, reaching a finer-grained level of analysis. Table 10 shows the ten more frequent lemmas attested for V2, with the respective relative frequency in the dataset: *essere* is six times more frequent than *avere* 'to have', the other auxiliary of Italian, which is the second most frequent verb.

Even more interestingly, the difference in frequency that we observe between *essere* and *avere* for V2 in our dataset is significantly higher than in the whole KIParla corpus, in which *essere* is just twice as frequent than *avere,* as shown in Table 11.

The especially high frequency of *essere* in *putandi* constructions is probably connected to its use as copula: in evidential contexts the expression of judgments and opinions is frequently conveyed through evaluations with the structure [copula *essere* + nominal/adjectival predicate]. The strong correlation between *essere* and the subjunctive (86 % of the cases vs. 61 % of the occurrences of *avere* 'to have', see Table 10) is therefore to be explained as a manifestation of the Conserving Effect of token frequency (Bybee 2010): given that highly frequent items often resist change, we expect that particularly high-frequency verbs will be the ones to resist change and preserve irregular forms and rare moods, like the subjunctive. As a consequence, the high frequency of the lemma within *putandi* constructions favors the rise of a routinized construction involving the use of subjunctive, namely [*putandi*_VERB

Table 11: Frequency list of verbs in the KIParla corpurs: first ten.

Lemma	Absolute frequecy	Relative frequency
Essere 'to be'	100,869	41,513.0883
Avere 'to have'	43,876	18,057.3641
Fare 'to do'	24,446	10,060.8607
Dire 'to say'	16,205	6,669.24025
Andare 'to go'	14,650	6,029.27305
Sapere 'to know'	9,473	3,898.65553
Stare 'to stay'	8,210	3,378.86223
Potere 'can'	7,045	2,899.40127
Dovere 'must'	6,656	2,739.30658
Volere 'want'	5,650	2,325.28278

[*che sia*]], exemplified in (18). This construction occurs in 29 % of the occurrences considered in our dataset (195/676).

(18) KIParla, PBC010
infatti e poi a parte quello ***penso*** ***che sia***
indeed and then apart_from that think:1SG that be:SBJ.3SG
anche molto utile
also very useful
'Indeed, and besides that, I also think it is very useful.'

4 Conclusions: going beyond expectations

This study set out to test long-standing assumptions about mood alternation after *verba putandi* in Italian, assessing both semantic–epistemic explanations and sociolinguistic accounts against spontaneous spoken evidence. The results show that neither the speaker's epistemic stance nor their educational background plays a decisive role in determining mood choice. Instead, the alternation emerges from the interaction of morphosyntactic, interactional, and usage-based factors whose effects become visible only when examined through corpus-based and quantitative methods.

A first crucial finding concerns the person of the subordinate verb. The models consistently reveal that the subjunctive is strongly favoured in third-person contexts, while first- and second-person forms overwhelmingly select the indicative. This asymmetry aligns with the frequency profile of *putandi* constructions, which

predominantly express beliefs about third parties. The formal syncretism of the singular subjunctive paradigm further reinforces this tendency: because the form *sia* is equally compatible with 1st, 2nd and 3rd person singular, speakers appear to rely on the indicative to avoid ambiguity when the reference is to the speech-act participants. Person effects therefore reflect both distributional frequencies and referential disambiguation strategies, rather than differences in certainty or commitment.

The second major factor is interactional context, operationalised as degree of formality. The analysis demonstrates a clear divide between loosely structured, spontaneous exchanges (free and kitchen-table conversations), where the indicative is proportionally more frequent, and contexts with codified roles and partly fixed topics (semi-structured interviews, office hours, exams, lessons), which instead favour the subjunctive. Importantly, the decisive property is not the symmetry of speaker roles, but the degree to which the interactional frame constrains topic management and role behaviour. This confirms that register is best understood not as a proxy for prestige or education, but as an index of interactional organisation, which shapes how speakers construe clause linkage.

Two further parameters, tense morphology and adjacency, shed additional light on the constructional dynamics at play. Past-tense forms in the subordinate clause, though infrequent, tend to favour the indicative, an effect interpretable through markedness: when one dimension of the verbal paradigm is marked (past), speakers gravitate toward the less marked mood (indicative). In turn, non-adjacency between *che* 'that' and V2 decreases the likelihood of subjunctive. In spoken discourse, pauses, parenthetical elements, and reformulations often weaken the syntactic and prosodic integration between the two clauses. As distance increases, the complement clause is more readily construed as a main clause (sometimes even approaching the behaviour of evidential comment clauses) thus favouring the indicative. These findings illustrate how prosodic fragmentation and dependency weakening interact with grammar in shaping mood selection.

Finally, the analysis provides evidence for emerging constructional patterns. The recurrent clustering of third-person contexts, present-tense morphology, and adjacency, together with the extremely high frequency of *essere* 'to be' as V2, supports the hypothesis that a routinised construction [*putandi*_VERB [*che sia*]] is developing, sustained by the Conserving Effect of token frequency. This perspective reconciles statistical observations with usage-based theory: entrenched high-frequency strings preserve less common morphological forms, such as the subjunctive, even as the overall system undergoes levelling.

Taken together, the results invite a reframing of mood alternation in Italian. Rather than a domain governed by epistemic evaluation or by speakers' educational characterization, it emerges as a probabilistic phenomenon shaped by distributional biases, morphosyntactic markedness, and the register. This usage-based account

helps explain why alternation persists despite strong prescriptive norms, and why it remains largely resistant to social evaluation in actual spoken practice. Beyond the Italian case, the findings have broader implications for the cross-linguistic study of mood systems, suggesting that future research should systematically integrate frequency patterns, dependency relations, and discourse organisation into theories of clause-linking and modality.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: Silvia Ballarè and Caterina Mauri. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: Silvia Ballarè: The research here presented was partly developed within the project M4C2 funded by the Italian Ministry of University and Research and the European Union – NextGenerationEU, titled "CHANGES – Cultural Heritage Active Innovation for Sustainable Society" – CUP F53C22000700006. Caterina Mauri: The research here presented was partly developed within the PRIN 2022 PNRR project "DiverSIta: diversity in spoken Italian", coordinated by Caterina Mauri (University of Bologna; MUR project code P2022RFR8T – CUP J53D23017320001). The project is funded by the European Union – NextGenerationEU with the The National Recovery and Resilience Plan (NRRP) – Mission 4 'Education and Research' – Component 2 'From Research to Business' – Investment 1.1 m Public Call PRIN 2022 NRRP published with the DDN. 1409 of 14.09/2022.
Data availability: The raw data can be obtained on request from the corresponding author. The corpus from which they were extracted is freely accessible (www.kiparla.it).

Abbreviations

1	1st person
2	2nd person
3	3rd person
AUX	auxiliary
COMP	complementizer
DAT	dative
DEF	definite
F	feminine
FUT	future
IND	indicative

INDEF	indefinite
INF	infinitive
IPFV	imperfective
LOC	locative
M	masculine
NEG	negation
OBJ	object
PL	plural
POSS	possessive
PRS	present
PST	past
PTCP	participle
REFL	reflexive
REL	relative
SBJV	subjunctive
SG	singular
SUP	superlative

References

Ballarè, Silvia. 2025. Tra norma e variazione: l'alternanza congiuntivo/indicativo nello scritto formale di studenti universitari. In Nicola Grandi (ed.), *L'italiano scritto degli studenti universitari. Quadro sociolinguistico, tendenze tipologiche, implicazioni didattiche*, 151–163. Milano: Franco Angeli.

Ballarè, Silvia & Massimo Cerruti. 2023. Sociolinguistic variation, or lack thereof, in the use of the Italian subjunctive. *Sociolinguistica* 37/1(2023). 75–93.

Berruto, Gaetano. 2012. *Sociolinguistica dell'italiano contemporaneo*, 2nd edn. Rome: Carocci.

Brown, Gillian & George Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

Dahl, Östen. 1985. *Tense and aspect systems*. London: Blackwell.

DeLancey, Scott. 1997. Mirativity: The grammatical marking of unexpected information. *Linguistic Typology* 1(1). 33–52.

Digesto, Salvio. 2019. *A variationist analysis of subjunctive variability across space and time: From contemporary Italian back to Latin*. Ottawa: University of Ottawa PhD Dissertation.

Digesto, Salvio. 2021. Lexicalization and social meaning of the Italian subjunctive. *Cadernos del Linguistíca* 2(3). https://doi.org/10.25189/2675-4916.2021.v2.n3.id609.

Givón, Thomas. 2001. *Syntax: An introduction*. Amsterdam: John Benjamins.

Grandi, Nicola, Silvia Ballarè, Francesca Chiusaroli, Francesca Gallina, Matteo Pascoli & Elena Pistolesi. 2023. *Corpus Univers-ITA*. Bologna: Alma Mater Studiorum - Università di Bologna. https://doi.org/10.60760/unibo/univers-ita.

Greenberg, Joseph. 2005. *Language universals: With special reference to feature hierarchies*, 2nd edn. Berlin: De Gruyter.

Gries, Stefan. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3). 617–647.

Gumperz, John J. 1964. Linguistic and social interaction in two communities. *American Antrhopologist* 66(6). 137–153.

Halliday, Michael A. K. 1970. Language structure and language function. In John Lyons (ed.), *New horizons in linguistics*, 140–165. Harmondsworth: Penguin.

Hosmer, David W. & Stanley Lemeshow. 2000. *Applied logistic regression*. New York: Wiley.

Jakobson, Roman. 1960. Closing statement: Linguistics and poetics. In Thomas Sebeok (ed.), *Style in language*, 350–377. Cambridge, MA: MIT Press.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh & Joeal Scherzer (eds.), *Language in use: Readings in sociolinguistics*, 28–54. Englewood Cliffs, NJ: Prentice Hall.

Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing Company.

Lo Baido, Maria Cristina. 2024. *Forms and functions of meta-discourse. The case of comment clauses in present-day Italian*. Berlin & Boston: Mouton De Gruyter.

Lombardi Vallauri, Edoardo. 2003. Vitalità del congiuntivo nell'italiano parlato. In Teresa Poggi Salani & Nicoletta Maraschio (eds.), *Italia linguistica anno Mille, Italia linguistica anno Duemila. Atti del XXXIV Congresso Internazionale della Società di linguistica italiana (Firenze 19–21 ottobre 2000)*, 609–635. Roma: Bulzoni.

Mauri, Caterina & Andrea Sansò. 2016. The linguistic marking of (ir)realis and subjunctive. In Jan Nuyts & J. Joan van der Auwera (eds.), *Handbook of modality and mood*, 166–195. Oxford: Oxford University Press.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti & Francesco Suriano. 2019. KIParla corpus: A new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds.), *Proceedings of the 6th Italian conference on computational linguistics CLiC-it*. Torino: Accademia University Press. Available at: http://ceur-ws.org/Vol-2481/paper45.pdf.

Nikolaeva, Irina. 2007. Constructional economy and nonfinite independent clauses. In Irina Nikolaeva (ed.), *Finiteness: Theoretical and empirical foundations*, 138–180. Oxford: Oxford University Press.

Noonan, Michael. 2007. Complementation. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. II, 52–150. Cambridge: Cambridge University Press.

Nordström, Jackie. 2010. *Modality and subordinators*. Amsterdam: John Benjamins Publishing Company.

Palmer, Frank Robert. 1986. *Mood and modality*. Cambridge: Cambridge University Press.

Palmer, Frank Robert. 2001. *Mood and modality*, 2nd edn. Cambridge: Cambridge University Press.

Poplack, Shana, Allison V. Lealess & Nathalie Dion. 2013. The evolving grammar of the French subjunctive. *Probus* 25. 139–193.

Poplack, Shana, Rena Torres Cacoullos, Nathalie Dion, Rosane de Andrade Berlinck, Salvio Digesto, Dora Lacasse & Jonathan Steuck. 2018. Variation and grammaticalization in Romance: A cross-linguistic study of the subjunctive. In Wendy Ayres-Bennett & Janice Carruthers (eds.), *Manuals in linguistics: Romance sociolinguistics*, 217–252. Berlin & New York: Mouton de Gruyter.

Prandi, Michele. 2012. Il congiuntivo e i suoi valori: un bilancio. In Remo Bracchi, Michele Prandi & Leo Schena (eds.), *Passato, presente e futuro del congiuntivo*, 97–128. Bormio: Centro Studi Storici Alta Valtellina.

Renzi, Lorenzo. 2019. Ancora su come cambia la lingua. Qualche nuova indicazione. In Bruno Moretti, Alin Kunz, Silvia Natale & Etna Krakenberger (eds.), *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale della Società di Linguistica Italiana (Berna, 6–8 settembre 2018)*, 13–33. Milano: Officinaventuno.

Schneider, Stefan. 1999. *Il congiuntivo tra modalità e subordinazione: uno studio sull'italiano parlato*. Roma: Carocci.

Serianni, Luca. 2006. *Prima lezione di grammatica*. Rome/Bari: Laterza.
Serianni, Luca. 2010 [1989]. *Grammatica italiana. Italiano comune e lingua letteraria*. Torino: Utet.
Squartini, Mario. 2008. Lexical vs. grammatical evidentiality in French and Italian. *Linguistics* 46. 917–947.
Squartini, Mario. 2010. Mood in Italian. In Björn Rothstein & Rolf Theroff (eds.), *Mood in the languages of Europe*, 237–250. Amsterdam: John Benjamins.
Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
Tamburini, Fabio. 2021. I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC e DiaCORIS. In Emanuela Cresti & Massimo Moneglia (eds.), *Corpora e studi linguistici. Atti del LIV congresso SLI (online 2021)*, 189–197. Milano: Officinaventuno.
Timberlake, Alan. 2007. Aspect, tense, mood. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. II, 280–333. Cambridge: Cambridge University Press.
Torres Cacoullos, Rena, Dora LaCasse, Michael Johns & Johan De La Rosa Yacomelo. 2017. El subjuntivo: hacia la rutinización. *Moenia* 23. 73–94.
Troiani, Giorgia, John Du Bois & Andrey Filchenko. 2024. Corpus as a slice of life: Representing naturally occurring language and its speakers. *Research in Corpus Linguistics* 12/2. 174–202.
Vellupilai, Viveka. 2016. Partitioning the timeline: A cross-linguistic survey of tense. *Studies in Language* 40(1). 93–136.