e-ISSN 2499-1562

Annali di Ca' Foscari. Serie occidentale
Vol. 58 — Num. 58 — Settembre 2024

Comparing Models on the Optionality of Complementizer Omission
A Quantitative Computational Study on German and Italo-Romance

Giuseppe Samo
Beijing Language and Culture University, China

Elena Isolani
University of Cambridge, UK

Abstract Different studies in generative grammar have tried to explain the optionality with respect to the complementizer omission across languages and across structures. Specifically, the complementizer omission in declarative embedded contexts introduced by the so-called bridge verbs. In this paper, we test two models postulating different derivations and marking different predictions to the nature of the complementizer omission: one model stipulates the deletion of the complementizer (complementizer deletion), whereas the second model focuses on the verbal elements of the embedded clause (complementizer rise, when present). To reach our goal, we explore large-scale datasets (syntactically annotated treebanks of German, Italian and Old Florentine) and adopt simple computational models to compare and test the models under investigation. Our results suggest that the predictions of the complementizer deletion hypotheses are confirmed by German data, while those of the complementizer rise model are partially corroborated by Italian data. We consider this study as a blueprint for finer-grained research.

Keywords Complementizer deletion. Bridge verbs. Italo-Romance. German. Quantitative computational syntax.

Edizioni Ca'Foscari

Peer review

Submitted 2024-07-03
Accepted 2024-09-10
Published 2024-10-08

Citation Samo, Giuseppe; Isolani, Elena (2024). "Comparing Models on the Optionality of Complementizer Omission". *Annali di Ca' Foscari. Serie occidentale*, 58(58), 287-308.

DOI 10.30687/AnnOc/2499-1562/2024/01/015

1 Introduction

The lack of co-occurrence has always represented a perfect diagnostic for the understanding of which elements compete for the same syntactic position and therefore are in a complementary distribution. Competing for the very same position has also had important implications with respect to the building of fine-grained maps in syntactic cartography – for example, the fact that foci and wh- interrogative elements compete for the same position in the Left Periphery (Rizzi 1997; but see Rizzi, Bocci 2017).

A well-studied case is related to the lack of co-occurrence between the verb and the complementizer in verb second languages (V2; Holmberg 2015). Since verb second (V2) environments can be observed in those embedded contexts lacking an overt complementizer, Den Besten (1983) claimed that the verb moves to C in main clauses in West Germanic. We illustrate the phenomenon in (1), using bold for embedded verb and complementizers and underline for the main verb, whose nature plays a fundamental role in eliciting or not the complementizer:

(1) German (a, b, c from Samo 2019a, 26 ex. 43a, b, c; d from UD-HDT, hdt-s14414)
   a. *Giotto* ***malte*** *dieses Fresko*
      Giotto painted this fresco
      'Giotto painted this fresco'
   b. *Der Stadtführer* <u>*sagt*</u>*,* ***dass*** *Giotto dieses Fresko* ***malte***
      The city.guide says, that Giotto this fresco painted
      'The tourist guide says that Giotto painted this fresco'
   c. *Der Stadtführer* <u>*glaubt*</u> *Giotto* ***malte*** *dieses Fresko*
      the city.guide believes that Giotto painted this fresco
      'the tourist guide believes that Giotto painted this fresco'
   d. *Wir* <u>*glauben*</u>*,* ***dass*** *es einige Punkte in der Vereinbarung* ***gibt****,*
      We believe that there some points in the agreement are
      die wir noch weiter diskutieren müssen
      that we still yet discuss must
      'We believe that there are some points in the agreement that we still need to discuss'

The data in (1) show that the inflected verb cannot be located in the second position of the embedded clause introduced by verbs like *sagen* 'say': a complementizer, in this case *dass* 'that', is present (1b). However, when the embedded clause is introduced by a so-called bridge verb (see Poletto 2014) such as German *glauben* 'to think', the inflected verb can optionally reach the 'second slot' of the sentence (1c, 1d).

Once established that two elements compete for the same position, it is important to understand, from a formal point of view,

hierarchies of priorities of syntactic elements. Does the verb move there because there is no complementizer? Does the complementizer appear in (1b) and (1d) because the verb cannot move there? Are these two phenomena interconnected or totally independent? Is there a third way? Is there a pure optionality across phenomena and across languages? To answer these questions, we run a study in the spirit of Quantitative Computational Syntax (Merlo 2016; Samo, Merlo 2019; 2021s), by comparing models on the basis of linguistic data retrieved from large-scale datasets via a quantitative analysis and simple computational models.

Let us briefly introduce the models under investigation, which will be discussed in detail in section 2. A first model, which we label Complementizer Deletion (henceforth, CD) follows Den Besten's intuition (1983), which can be summarized as follows: the complementizer is absent (Ø), the verb moves to C, as in (2a). On the other hand, a second model, which we label Complementizer Rise (CR), would assume that the complementizer emerges since the verb cannot reach the activated functional projection (indicated by 0 ← X ← in 2b).[1] The two models are summarized in (2).

(2) a. CD
if $[_C [_T [_V [_C \text{Ø} [_T \text{verb} [_{VP}\text{<verb>}]]]]]$
then $[_C [_T [_V [_C \text{verb} [_T\text{<verb>}[_{VP}\text{<verb>}]]]]]$
b. CR
if $[_C [_T [_V [_C\text{< 0>} [_T \text{verb} [_{VP}\text{<verb>}]]]]]$
$\leftarrow X \leftarrow$
then $[_C [_T [_V [_C \text{comp} [_T \text{verb} [_{VP}\text{<verb>}]]]]]$

These models can be compared to a third model in which the selection is fully random, representing a form of null hypothesis.

To reach our goals we proceed as follows. Section 2 presents the core properties of the linguistic phenomenon and core ingredients of the formal models under investigation. We then run our study: the hypotheses are introduced in section 3, materials and methods are presented in section 4, while section 5 discusses the results. Finally, section 6 concludes.

1 Theoretically, other models can be tested, such as those stipulating the presence/absence of the complementizer simply regulated by the relevant functional projection, as it has been proposed for, among others, Chinese (see Xu 1993 *inter alia*). We leave this and other comparisons to future works.

2 Core Properties of the Linguistic Phenomenon and Models

The question on whether the declarative complementizer is omittable has not only been at the center of the debate in Germanic languages studies (Holmberg 2015), but it has also extensively affected the research on (Italo-)Romance languages which largely bridges the theoretical assumptions developed for Germanic to the status of Romance complementizer drop (see Poletto 1995). This primarily concerns the nature and the relation between the selecting verb available in the root clause and the subsequent embedded verb. As a matter of fact, Vikner (1994), analyzing embedded verb movement in Germanic, elaborated a system of verbal classification in order to detect the main verbs that do not compulsorily need a complementizer and, consequently, are eligible to select V2 in the embedded clause.

Hence, they allow for verb movement from the inflectional domain to the complementizer phrase. Along these lines, *bridge* verbs stand for the root predicate triggering V2 (3a), whereas *non-bridge* verbs are the ones which do not allow for V2 (3b):

(3) German (Vikner 1994, 132 ex. 39b, 40b)
   a. *Watson behauptete, dieses Geld hatte Moriarty gestohlen*
      Watson claimed this money had Moriarty stolen
      'Watson said (that) this money had Moriarty stolen'
   b. **Holmes bewies, dieses Geld hatte Moriarty gestohlen*
      Holmes proved this money had Moriarty stolen
      'Holmes proved (that) this money had Moriarty stolen'

In this regard, Poletto (1995) identified a parallel between Germanic languages and standard Italian observing that the verbs which select V2 in Germanic embedded clauses, likewise license complementizer drop in Italian, hence reaching the conclusion that complementizer deletion must involve verb movement to the complementizer phrase on a par with V2 in Germanic.

When the complementizer is absent, the following properties are at work: (a) the main verb needs to belong to the bridge verb class, (b) the embedded verb must be inflected for irrealis morphology, (c) the subordinate clause cannot be left-dislocated (see Poletto 1995):

(4) *Credo (che) sia già partito*
   believe (that) be.pst.sbjv.3sg already left
   'I believe (that) he has already left'

Following Vikner's (1994) assumptions, Poletto (1995) predicted that, in the absence of the declarative complementizer *che* 'that', the embedded verb undergoes a head-movement towards the left-periphery

functioning as an alternative checker of the omitted complementizer. In featural terms, the raised verb is able to check the same bundle of functional features that the complementizer would do, which, in turn, can be felicitously dropped. Structurally speaking, Poletto (1995) hypothesized that the embedded verb in complementizer deletion structures targets a low C-projection, specifically Fin°.[2]

A further pioneering study within the field of complementizer deletion in standard Italian dates back to Giorgi and Pianesi (1997) who agreed with Poletto (1995) in terms of verb movement to CP in this context, but ruled out any associations with the Germanic languages' system of V-to-C movement.[3]

Therefore, most literature tradition on complementizer deletion in standard Italian is composed of remarkable contributions aimed to define empirical patterns that can predict the distribution of the declarative complementizer and factors that may affect its omission.

It is also noteworthy the case of the lack of the declarative complementizer in Florentine, which reveals a series of similarities with standard Italian, but, as extensively reported by Cocchi and Poletto (2002; 2007), a more flexible portrait can be depicted. Cocchi and Poletto (2002) observed that not only does Florentine license complementizer deletion under the same structural conditions of Italian, but it also features complementizer deletion when a different combination of main and embedded verbs are available. To put it differently, the lack of complementizer in Florentine occurs regardless of the bridge or non-bridge status of the selecting verb and of the irrealis or realis nature of the embedded verb, provided that a clitic-like element intervenes between the main and the subordinate predicate. Therefore, the strict condition affecting the complementizer omission is not the nature of the verbs under analysis, but the availability of a clitic-like item, such as a pronominal clitic, a preverbal negator or an auxiliary, in an intermediate position (see Cocchi, Poletto 2002; 2007):

(5) Florentine (Cocchi, Poletto 2002, 3 ex. 9)
Gli dispiace la un venga a casa
he is sorry she-subj.cl not comes at home
'He is sorry she doesn't come home'

2 Poletto's (1995) assumption on the position of the declarative complementizer in standard Italian is in line with a series of studies (Ledgeway 2005; Paoli 2007; Colasanti 2018 *inter alia*) which do not provide a fixed location for this item, but promote a more flexible view according to which the complementizer, akin to other functional exponents, can navigate the syntactic structure and can fill various projections on the basis of a featural-checking criterion.

3 Giorgi and Pianesi firmly stated that verb movement to the CP in complementizer deletion configuration is the result of the irrealis (or subjunctive) mood property of the verb itself which realizes a "syncretic category […] projecting the agreement and the mood features" (1997, 239).

Along the lines of the account provided by Poletto (1995) for CD1 in standard Italian, Cocchi and Poletto (2002; 2007) attempted to unify complementizer deletion under the 'alternative checking hypothesis', according to which two elements are alternative checkers if they check the same bundle of formal features while competing for the analogous structural projection (Zanuttini 1997; Obenauer 2001; Cocchi, Poletto 2002; 2007). In this regard, in Italian the alternative checking configuration affects the declarative complementizer and the embedded verb, while in Florentine, it involves the declarative complementizer and the intervening element between the main and the embedded predicate. Therefore, Cocchi and Poletto (2002) hypothesized that in Florentine complementizer drop configurations, the subordinate verb is stranded in the inflectional domain, whereas the clitic-like item interposing between the two verbs is displaced towards the left-periphery. Moreover, as opposed to complementizer deletion in Italian, where the embedded verb raises to the low left-periphery, namely to FinP, to check the irrealis feature, in Florentine, the feature involved is associated with a higher functional projection, that is ForceP, where the clitic-like element moves to. One of the leading arguments in favor of a distinct structural projection hosting respectively the embedded verb in Italian complementizer drop and the clitic-like exponent in Florentine complementizer drop is that whereas the former licenses a preverbal lexical subject intervening between the main and the embedded verb, the latter rules it out:

(6) Cocchi, Poletto 2002, 9 ex. 21a, 21b
   a. *Credo **Gianni** abbia telefonato*
      believe Gianni has.sbjv called
      'I believe Gianni has called'
   b. **Maria mi ha detto **Gianni** un ha portato il libro*
      Maria to-me has said Gianni not has brought the book
      'Maria told me Gianni has not brought the book'

According to Cocchi and Poletto (2002), the ungrammaticality of (6b) is due to the highest functional projection filled by the clitic-like element *un* 'not'. By assuming that the negator undergoes a displacement to ForceP, the fine-structure of the left-periphery does not permit any other functional exponents to precede it. On the other hand, the movement towards FinP in (6a) does not inhibit other left-peripheral dislocations to the left of the embedded verb.

The account provided by Cocchi and Poletto (2002) in order to explain the ungrammaticality of (6b) has a double-edged consequence: whereas it adequately justifies its ill-formedness, it concomitantly predicts that the opposite order between the clitic and the preverbal lexical subject is expected. Structurally speaking, if the negator fills a high structural projection like Force, it naturally precedes other

left-peripheral or inflectional items like a preverbal lexical subject. However, a sentence like (7) results ungrammatical as well:

(7) Florentine (Cocchi, Poletto 2002, 11 ex. 23)
**Mi dispiace un Gianni viene stasera*
I am sorrynot Gianni comes tonight

The solution proposed by Cocchi and Poletto (2002) to account for the ungrammaticality of (7) is that there exist some phonological form (PF) constraints that force the negator, as a clitic element, to form a unique unit with the verb at the phonological level, hence provoking a strict adjacency between the clitic and embedded verb at the surface form.

Cocchi and Poletto (2007) proposed a different analysis to account for complementizer drop in Florentine; still embracing the 'alternative checking hypothesis', they ruled out the movement of the intervening clitic, whereas they relied on the Agree operation. In other words, the clitic does not raise to any left-peripheral projection, but it remains within the IP along with the embedded verb being probed by Force°. The postulation of the Agree operation can solve the issue related to (7) without referring to PF constraints. Indeed, if the intervening clitic remains adjacent to the embedded verb, it comes straightforward that a preverbal subject cannot be sandwiched between them.

In sum, Cocchi and Poletto (2002; 2007) treated complementizer deletion in Italian and Florentine in a similar vein, positing that they both instantiate alternative checking between the declarative complementizer and an additional exponent: while the former licenses embedded verb movement to Fin alternating with the complementizer, the latter resorts either to the displacement towards Force of the clitic-like element available in an intermediate position between the main and the embedded verb (Cocchi, Poletto 2002) or to the Agree operation taking place between Force° and the clitic in its merge position.

A more recent approach to complementizer deletion in Florentine highlights the role of the embedded verb (see Isolani 2023). More specifically, given the account provided by Cocchi and Poletto (2002) for the ungrammaticality of (7) relying on some PF constraints and not on purely syntactic basis and given the ambiguous status of the intervening clitic-like element (Cocchi, Poletto 2007), the proposal advanced by Isolani (2023) focuses on the role of the embedded verb rather than the clitic itself.[4] Along these lines, a stricter link

4 Notwithstanding the syntactic base account provided for the ungrammaticality of (7), the proposal advanced by Cocchi and Poletto (2007) results controversial in terms of the vague definition of intervening clitic assumed. Cocchi and Poletto (2007) established the head nature of the clitic, but they distinguished between subject/object clitic, encoding argumental feature and negators/auxiliaries bearing declarative feature. However, this distribution does not make explicit the selection of one or the other set of features

between complementizer omission in Italian and Florentine is established whereby they both involve verb movement towards the left-periphery: the former to FinP and the latter to ForceP.

The exclusion of the clitic-like displacement in Florentine structures rests on several pieces of evidence showing its vague and optional nature. In particular, Isolani (2023) observed that this item is not obligatory as Cocchi and Poletto (2002; 2007) predicted. In order to verify that, Florentine was excluded as, given its subject clitic nature, it is very unlikely that the selecting verb is adjacent to the embedded verb without any intervening pronominal clitic. Conversely, Pisano was taken into account; on a par with Florentine, Pisano shows a more flexible approach to complementizer deletion by accepting more combinations of main and embedded verbs than standard Italian, hence abstracting away from pure (1).[5] Additionally, Pisano does not present subject clitic, thus the main and embedded verb can potentially be adjacent:

(8) Pisano
   a. *Ha* *detto* *viene* *da solo*
      has said comes.ind alone
      'He has said that he comes alone'
   b. *Penso* *venga* *da solo*
      think come.sbjv alone
      'I think he comes alone'
   c. *Mi dispiace* *venga* *da solo*
      I am sorry come.sbjv alone
      'I am sorry he comes alone'

The well-formedness of (8a,b,c), exhibiting different combinations of main and embedded verbs, reveals the optionality of the intervening item. Therefore, if the clitic is not obligatory, Cocchi and Poletto's proposal (2002) on clitic movement in Florentine complementizer

in different configurations. Also, it remains obscure why pronominal clitics like reflexives, locatives and partitives are not included in the set of interveners (cf. Isolani 2023).

5 As mentioned by Isolani (2023), the status of complementizer drop in Pisano still needs a proper investigation. It seems, indeed, not only to depart from the case of Italian, but also from Florentine, as not all the combinations of main and embedded verbs available in the latter are admissible in Pisano. For instance if the selecting verb is non-bridge and the embedded verb is realis, the structure is degraded:

(2) Pisano
   **Mi dispiace rompono sempre tutto*
   I am sorry break.ind always everything
   'I am sorry they always break everything'

Replying to the intuition of one of the reviewers of this article, the introduction of a clitic exponent between the main and the embedded verb does not significantly improve the grammaticality judgment of (2), retaining a strong preference for an embedded verb inflected for irrealis morphology in the subordinate clause.

drop structures is called into question; the clitic cannot function as the main character in the derivation if it can be omitted.[6]

The solution advanced by Isolani (2023) is that in Florentine, the embedded verb raises towards the left-periphery, in the same vein as in Italian, whereas the clitic, if present, can move along with the verb.[7] If this is truly the case, the sentence in (7) can be discarded by assuming that, in Florentine, the verb is unable to follow a preverbal lexical subject owing to its prominent position, from which it can only precede it. Thus, this approach permits the elimination of (7) relying on syntactic analysis rather than on PF constraints. As also mentioned in footnote 4, Cocchi and Poletto's account (2007) is ruled out even though it syntactically accounts for the ill-formedness of (7) because of the vague definition and adoption of clitics as interveners. The complete rejection of a clitic-base account given its optionality permits to exclude any ambiguity related to the status and classification of pronominal clitics.

In short, Isolani (2023) proposed a parallel analysis of complementizer deletion in Florentine, based entirely on embedded verb movement along with the proposal advanced for Italian by Poletto (1995). In this regard, the occurrence of the intervening clitic element is fundamentally irrelevant. Complementizer deletion in Italian and Florentine can, hence, be unified under the verb movement to the CP hypothesis, whereby the predicate reaches distinct structural projections according to the feature in need to be checked.

Finally, some words on the generation site of the complementizer. The complementizer is usually considered as generated in dedicated functional projections in the LP (cf. Rizzi 1997). However, building on Leu (2015), Samo (2019a, ch. 4) proposes an IP internal nature of the complementizer for Germanic, drawing on locality constraints. A different base-generation site of the complementizer would not affect, at this stage, with our tools, the results of our study.

6 It is worthwhile reminding that the reason why complementizer deletion structures in Florentine are well-formed only if an intervening clitic-like element occurs is due to the subject clitic nature of the language itself. Since structures without an overt subject clitic are ungrammatical in Florentine, it results that in subordinate configuration lacking the declarative complementizer, the embedded verb must be preceded by (at least) a subject clitic. From the opposite perspective, the ungrammaticality of a construction without the declarative complementizer and without an intervening clitic-like element, presumably of the subject nature, is not due to the lack of former, but actually to the absence of the latter, which is mandatory in a subject clitic variety like Florentine.

7 Other evidence provided by Isolani (2023) to support this hypothesis concerns the order between the embedded verb and other left-peripheral items, revealing that the verb can only precede focalized or topicalized constituents, whereas it naturally precedes hanging topics. This piece of evidence further upholds the view that the embedded predicate in this configuration is displaced towards a significantly high position in the structure, presumably in Force.

The complementizer deletion patterns are primarily drawn along the split between the main and embedded verb and their idiosyncratic properties: the selecting verb is likely to reconcile with Vikner's (1994) classification of bridge and non-bridge verb, whereas the embedded verb needs to comply with a specific inflectional morphology. The positive outcome of the combination of main and embedded verb results in the movement of the embedded verb and in the omission of the complementizer. However, as hinted in this section, no clear statement has ever been put forth to disentangle the cause-effect conflict concerning complementizer deletion and verb movement. In other words, although some authors seem to implicitly prefer one or the other option, it remains obscure whether verb movement takes place because of the absence of the complementizer and of the subsequent need to check some relevant feature that would remain unchecked otherwise or whether the lack of the complementizer is a direct consequence of the verb raising towards the left-periphery, inevitably triggering a complementary distribution configuration.

Summing up, we can identify the theoretical grounds for the two models.

1. *Model CD*: The verb of the main clause selects the underlying CP. The relevant features of such a selection should be checked by the complementizer. In its absence, due to featural checking requirements, the verb has to be raised to the C layer.

(9) CD
if [$_C$ [$_T$ [$_V$ [$_C$ Ø [$_T$ verb [$_{VP}$<verb>]]]]
then [$_C$ [$_T$ [$_V$ [$_C$ verb [$_T$<verb>[$_{VP}$<verb>]]]]

2. *Model CR*: The verb of the main clause selects the underlying CP and its verbal root. The verb raises to the C layer to comply with such a selection, but different factors may undermine the required movement. The complementizer thus emerges to check the relevant requirements.

(10) CR
if [$_C$ [$_T$ [$_V$ [$_C$< 0> [$_T$ verb [$_{VP}$<verb>]]]]
← X ←
then [$_C$ [$_T$ [$_V$ [$_C$ comp [$_T$ verb [$_{VP}$<verb>]]]]

In the remainder of this article, we run a quantitative and computational study to compare the two models. Section 3 presents the quantification of the hypotheses.

3 Quantifying the Hypotheses

To compare the two models, we follow an approach inspired by the studies in Quantitative Computational Syntax (Merlo 2016), which explores large-scale datasets and simple computational models. An important contribution of this framework is to adopt frequency as a dependent variable to test linguistic proposals. Frequency acts as a measure of syntactic computation *id est* frequency depends on grammar and may reveal important facts of the underlying structure. Following Samo and Merlo (2021, 29) the quantitative dimension of structures in large datasets allows us "to develop investigations of the correlation between quantitative linguistic properties and theory-driven abstract linguistic representations and operations".

We operate as follows. We decided to collect data in German, a V2 language, Italian and Old Florentine (14th century). These three languages are presented in syntactically annotated treebanks under the guidelines of Universal Dependencies (UD; De Marneffe et al. 2021). UD allow for a fine-grained syntactic search and for the retrieval of lemmas, uninflected verbal form (e.g. the retrieval of the annotated lemma *credere* 'believe' from the Italian treebanks will provide us with all the inflections of the verb), as well as the mood, when annotated, of the inflection (e.g. SUB for subjunctive). Please note that the treebanks for Old Florentine follow the contemporary Italian lemma instructions, which favor our analysis. Details on the size of the treebanks and the query structures are given in section 4 ("Materials & Methods") and in the supplementary materials.

The first independent variable under investigation is represented by the type of verb in the main clause. We have decided to isolate a restricted set of four verbs labeled as bridge verbs – the German and Italian forms of the verbs 'to believe', 'to know', 'to think' and 'to hope'. As a control group, we use non-bridge verbs that usually license the presence of the complementizer ('to regret', 'to notice', 'to confirm' and 'to doubt'): we label this group "License" **[tab. 1]**.

The second independent variable is the presence (Comp) or the absence (V-to-C) of the complementizer in the embedded clause introduced by the verbs in Table 1 in the relevant treebanks. Finally, a third independent variable is the mood of the inflection of the verb in the embedded clause. Relying on the annotation scheme, and simplifying our model in just two values, we have decided to only operate our counts with respect to subjunctive (SUB) and non-subjunctive (non-SUB).

The two models, CD and CR, make different predictions with respect to the computational costs of the structures. As a reminder, CD proposes that the complementizer is present (let us call it the simpler, 'canonical' configuration) but then it is omitted (more complex,

'marked' configuration); on the other hand, CR stipulates that the verb moves ('canonical'), but if the movement does not take place, a complementizer emerges ('marked').

Table 1 Sets of bridge verbs and 'license' verbs lemmas in Italian and German and their English glosses

English Gloss	(Old) Italian	German
Bridge Verbs		
believe	credere	glauben
know	sapere	wissen
think	pensare	denken
hope	sperare	hoffen
'License' Verbs		
regret	dispiacersi	bedauern
notice	accorgersi	bemerken
confirm	confermare	bestätigen
doubt	dubitare	zweifeln

Different hypotheses can be carried out and all of them have a crosslinguistic nature. A non-trivial first hypothesis is related to the probability of the presence of the complementizer (Comp) with bridge and license verbs. The two models assume two different generative processes: CD stipulates that the presence of the complementizer (Comp) is easier than the movement of the verb (V-to-C), while CR postulates the opposite. Therefore, in both cases we expect asymmetric distributions. Due to the quantitative nature of this study, we can also test how much our results are given by exploring a binomial distribution (in line with Samo, Merlo 2019).

The second hypothesis is related to Sub in bridge verbs: CR makes clear predictions with respect to the movement of the verb inflected with subjunctive mood: bridge verbs should favor this configuration. In order to detect forms of preferences, we need to create simulated counts representing a baseline, such as expected counts (Exp) on the basis of the probability distribution of subjunctive forms in the entire treebank. This comparison between observed counts in the given configuration (e.g. subjunctive mood of the embedded verb selected by a bridge verb), and expected counts, an imputed count on the basis of the mere probability of an event to occur (e.g. the probability of a verb of being subjunctive in a given dataset) has been successfully explored across phenomena and languages in quantitative computational syntax (see the overview in Merlo, Samo forthcoming; see also Van Craenenbroeck, Van Koppen 2022; Samo, Merlo 2019; 2021; Merlo, Samo 2022).

No other asymmetry is predicted, therefore we should expect that for verbs introduced by complementizers for CR, and both configurations for CD, the observed counts should be similar (≈) to the expected one. The hypotheses are summarized in **[tab. 2]**.

Table 2 Hypotheses and predictions for each model; > stands for higher probability

Model	Comp vs. V-to-C	Sub vs. Non-Sub
CD	Comp > V-to-C	$Sub_{Comp} \approx Sub_{Exp}$ $Sub_{v\text{-}to\text{-}C} \approx Sub_{Exp}$
CR	V-to-C > Comp	$Sub_{Comp} \approx Sub_{Exp}$ $Sub_{V\text{-}to\text{-}C} > Sub_{Exp}$

The materials and methods of the study are presented in section 4.

4 Materials & Methods

We explored seven treebanks (three for Italian and German, one for Old Florentine) annotated following the guidelines of UD. The treebanks belong to different registers and genres including, but not limited to, newspapers, legal texts, encyclopedic entries and social media. One treebank for German (LIT) and the Old Florentine datasets contain poetry and literature. In particular, the Old Florentine treebank (labeled as Old Italian in the UD community) represents the syntactically annotated corpus of Dante's Divine Comedy. Although we recognize this factor as a limitation, due to the constraints that poetic texts inherently bear for generative analysis, we do believe that our results are indicative - we are looking to some syntactic aspects triggered by the lexicon, which will not be extremely affected by the text genre. From a replicability point of view, the process can be fully automatized and not rely on additional manual annotation **[tab. 3]**.

We automatically retrieved the counts via a python script from *grew.count.fr*. All the queries and scripts are available as supplementary files. Relevant examples of structures from the Italian treebank ISDT are given in **[tab. 4]**.

Table 3 Treebanks, size in terms of tokens and trees and references

Language	Treebank	Size (tokens)	Size (trees)	References
Italian	ISDT v.2.13 [l, n, w]	278,461	14,167	Bosco et al. 2014
	VIT v.2.13 [n, nf]	259,625	10,087	Alfieri, Tamburini 2016
	PoSTWITA v.2.13[sm]	119,334	6,712	Sanguinetti et al. 2018
German	HDT v.2.13 [n, nf, web]	3,399,390	189,928	Borges Völker et al. 2019
	GSD v.2.13 [n, r, w]	287,721	15,590	See caption
	LIT v.2.13 [nf]	40,340	1,920	
Old Florentine	Italian-Old v.2.13 [p]	80,694	2,402	Corbetta, Passarotti, Moretti 2024[8]

Table 4 Examples of conditions, queries and output sentences (with their ID)

Condition	Query	Example (ID)
Bridge + Comp	pattern { verb [lemma= credere | sapere | pensare | sperare]; verb -[ccomp]-> CP2; CP2 -[mark]-> Comp }	*Spero che sarà esaminata con uno spirito positivo* 'I hope that it will be looked at in a positive spirit' (2_Europarl-42)
License + Comp	pattern { verb [lemma= dispiacere | accorgere | confermare | dubitare]; verb -[ccomp]-> CP2; CP2 -[mark]-> Comp }	*altri documenti confermano che Piero fu suo assistente* 'other documents confirm that Piero was his assistant' (tut-3318)
Bridge + SUB + Comp	pattern { verb [lemma= credere | sapere | pensare | sperare]; verb -[ccomp]-> CP2; CP2 -[cop | aux]-> aux; aux [Mood = Sub]; CP2 -[mark]-> Comp }	*Credo che certe cose possano pure stancare* 'I believe that certain things can also be tiring' (isst_tanl-2463)
Bridge + SUB + V-to-C	pattern { verb [lemma= credere | sapere | pensare | sperare]; verb -[ccomp]-> CP2; CP2 [Mood = Sub]} without { CP2 -[mark]-> Comp }	*Spero non si arrabbino quelli che mi danno da mangiare (il gruppo sportivo Carabinieri) ma io voto dall'altra parte* 'I hope those who feed me (the Carabinieri sports group) don't get angry but I vote the other way' (isst_tanl-2235)[9]

We explore all the treebanks for the first hypothesis (comp vs. V-to-C), while for the second hypothesis we also manually observed the quality of the annotation. As a matter of fact only, the German treebank GSD clearly marked subjunctive forms. We therefore have decided to only work GSD and on the Italian treebank ISDT to maintain

8 For the references of GSD and LIT see the relevant treebank hub pages: `https://universaldependencies.org/treebanks/de_gsd/index.html`. Genres: l = legal, n = news, nf = nonfiction, p = poetry, sm = social media, r = reviews, w = wiki, web = web.

9 Please note that the form *arrabbino* 'to get angry' is present in the original naturally occurring example.

comparable sizes since (*circa* 15,000 trees). The results are presented and discussed in section 5.

5 Results & Discussion

All data points are available in the Appendix. We here present the relevant data for the two hypotheses. As stated in section 3, we compare the two models with respect to the probability of the presence of the complementizer in bridge verbs. Table 5 presents the probability in bridge verbs compared to a random group (given by an exact binomial p) and the license group. We aggregate the counts of all treebanks for each language.

Table 5 Probability of complementizer presence with bridge verbs, license verbs and a 'random' control group explored with the binomial

Language	Bridge	bin. p (random)	License
Italian	0.77	< 0.00001	0.92
German	0.75	< 0.00001	0.81
Old Florentine	0.74	0.00004	1.00

As Table 5 shows, the probability of the presence of the complementizer with bridge verbs is similar across languages (Italian 77%, German 75% and Old Florentine 74%). Our results also demonstrate that bridge verbs are different from license verbs with respect to the presence of the complementizer (Italian 92%, German 81% and Florentine 100%) and from a 'random' group that would have established a 50% probability of presence of the complementizer in the three languages. In other words, we observe a clear tendency to rule out complementizer omission in bridge verbs, supporting the predictions of the CD model.

Let us move to the second hypothesis related to the presence of the subjunctive inflection in the embedded verb. In this case, we adopt the simple computational model based on the comparison between an observed distribution and the expected distribution, as discussed in section 3. The results are summarized in **[fig. 1]** and can be read as follows. The data from German clearly confirm the predictions of the CD model, in line with the literature on Germanic (see section 2): the expected counts are similar to the observed counts (17% vs. 17% and 18%). The Italian data display a higher distribution of subjunctive in embedded clauses introduced by bridge verbs signaling that the selection ability of the bridge verb does play an important role. Finally, the Old Florentine data show an intriguing result: the observed counts (11%) for V-to-C are similar to German (18%), while

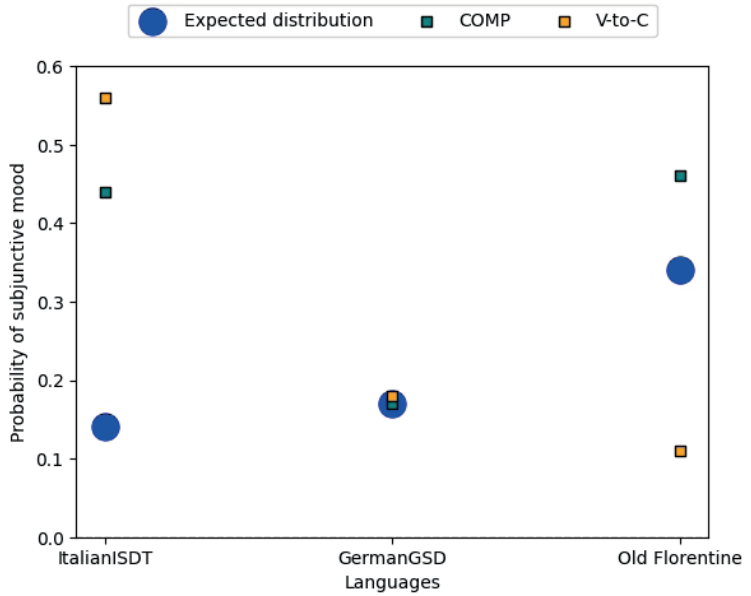

Figure 1 Probability of subjunctive mood in main/embedded verbs in the entirety of the treebank (expected) and in those embedded contexts with (Comp) or without (V-to-C) complementizer introduced by the bridge verb.

the Comp figures (46%) are similar to Italian (44%). Future studies should investigate such behavior.

Let us compare the predictions of each model given in **[tab. 2]** with the output of our study, as given in **[tab. 6]**.

Table 6 Hypotheses and confirmation

Model	Comp vs. V-to-C	Hypothesis confirmed	Sub vs. Non-Sub	Hypothesis confirmed
CD	Comp > V-to-C	✔	$Sub_{Comp} \approx Sub_{Exp}$ $Sub_{v\text{-}to\text{-}C} \approx Sub_{Exp}$	Only for German
CR	V-to-C > Comp	X	$Sub_{Comp} \approx Sub_{Exp}$ $Sub_{V\text{-}to\text{-}C} > Sub_{Exp}$	Partially for Italian

While the investigation of the first hypothesis (Comp vs. V-to-C) shows clear results and a winning model (CD), the exploration of the second hypothesis reveals meaningful points of analysis. The resulting asymmetry between German and the two Romance languages seems to comply with the early observation by Giorgi and Pianesi (1997). The CR model's predictions are partially corroborated by the Italian data.

We believe that the preliminary results of this study might be of interest for the theoretical community working on complementizer omission and on optionality. Future studies should improve the methodology and in line with other works on quantitative analyses of optionality (Samo, Si 2024) explore even larger and non-syntactically annotated datasets which may provide a richer quantitative dimension.

6 Conclusions

In this paper we have presented a methodology borrowed from quantitative computational syntax to test a fine-grained research question on two different models trying to explain the presence or the absence of the complementizer in embedded clauses selected by bridge verbs. Specifically, we have tested two models: a Complementizer Deletion and a Complementizer Rise model.

We explored seven large-scale treebanks and simple computational models to compare the predictions of the two models under investigation. Our results show that the Complementizer Deletion model maps the results in a precise way, at least for German, while the Complementizer Rise model seems to be, partially, a good model for Italian.

The processes of model comparison and model selection represent methodologies to test, in a quantitative and computational way, the predictions of formal theories and, ultimately, to understand their learnability (cf. Merlo 2016). However, the comparison and the selection should always depend on factors of grammaticality since simulated data, as expected counts, are built on grammatical clauses extracted from large-scale datasets. Dialogue with frameworks like cartography and their strong empirical predictive power allows for such experiments. Recently, these works have tested the generation site of constituents and the functional lexicon (e.g., testing whether initial non-arguments are generated or moved; Samo 2022) and locality effects (bottleneck effects and locality in V2 languages; Samo 2023). The creation of dedicated annotated corpora might be an additional layer of work that prevents full automation of the process. However, translations of available sources (Samo 2019b) or the exploration of large-language models (see details in Merlo, Samo, forthcoming; Wilcox, Futrell, Levy 2023) are viable solutions. With respect to complementizer deletion, such explorations, once all external factors are accounted for, might represent forms of improvement of the methodology outlined here.

Bibliography

Alfieri, L.; Tamburini, F. (2016). "(Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format". Corazza, A.; Semeraro, G., Montemagni, S. (eds), *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016* (Naples, 5-7 December 2016). Turin: Accademia University Press, 19-23. CEUR Workshop Proceedings 1749. https://cris.unibo.it/handle/11585/592497.

Borges Völker, E. et al. (2019). "HDT-UD: A Very Large Universal Dependencies Treebank for German". Rademaker, A.; Tyers, F. (eds), *Proceedings of the Third Workshop on Universal Dependencies* (UDW, SyntaxFest, Paris, 29-30 August 2019), 46-57. https://aclanthology.org/W19-8006.pdf.

Bosco, C. et al. (2014). "The EVALITA 2014 Dependency Parsing Task". Basili, R.; Lenci, A.; Magnini, B. (eds), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014* (Pisa, 9-11 December 2014). Pisa: Pisa University Press, 1-8. https://www.torrossa.com/gs/resourceProxy?an=3044378&publisher=F46792.

Cocchi, G.; Poletto, C. (2002). "Complementizer Deletion in Florentine: The Interaction Between Merge and Move". Beyssade, C. et al. (eds), *Romance Languages and Linguistic Theory 2000*. Amsterdam: John Benjamins Publishing Company, 57-76. Current Issues in Linguistic Theory 232. https://doi.org/10.1075/cilt.232.05coc.

Cocchi, G.; Poletto, C. (2007). "Complementizer Deletion and Double Complementizers". Picchi, M.C.; Pona, A. (eds), *Proceedings of the "XXXII Incontro di Grammatica Generativa"* (Florence, 2-4 March 2006). Alessandria: Edizioni dell'Orso, 49-62.

Colasanti, V. (2018). "La doppia serie di complementatori nei dialetti del Lazio meridionale: un approccio microparametrico". *Revue de linguistique romane*, 82, 65-91.

Corbetta, C.; Passarotti, M.; Moretti, G. (2024). "The Rise and Fall of Dependency Parsing in Dante Alighieri's Divine Comedy". Sprugnoli, R.; Passarotti, M.C. (eds), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024* (Turin, 25 May 2919). ELRA-ICCL, 50-6. https://aclanthology.org/2024.lt4hala-1.7/.

De Marneffe, M. et al. (2021). "Universal Dependencies". *Computational Linguistics*, 47(2), 255-308.

Den Besten, H. (1983). "On the Interaction of Root Transformations and Lexical Deletive Rules". Abraham, W. (ed.), *On the Formal Syntax of the Westgermania. Papers from the 3rd Groningen Grammar Talks (3e Groninger Grammatikgespräche), Groningen, January 1981*. Amsterdam: John Benjamins Publishing Company, 47. Linguistik Aktuell/Linguistics Today 3. https://doi.org/10.1075/la.3.03bes.

Giorgi, A.; Pianesi, F. (1997). "On the Semantics and Morphosyntax of the Italian Subjunctive". Giorgi, A.; Pianesi, F. (eds), *Tense And Aspect From Semantics to Morphosyntax*. New York: Oxford University Press, 93-279. https://doi.org/10.1093/oso/9780195091922.003.0005.

Holmberg, A. (2015). "Verb Second". Kiss, T.; Alexiadou, A. (eds), *Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and*

Communication Science (HSK) 42/1. Berlin; München; Boston: De Gruyter, 342-83. https://doi.org/10.1515/9783110377408.342.

Isolani, E. (2023). "Verb Movement in Florentine: The Case of Complementizer Deletion Under a Parametric Approach". *Isogloss. Open Journal of Romance Linguistics*, 10(1), 1-34. https://doi.org/10.5565/rev/isogloss.369.

Ledgeway, A. (2005). "Moving Through the Left Periphery: The Dual Complementiser System in the Dialects of Southern Italy". *Transactions of the Philological Society*, 103(3), 339-96. https://doi.org/10.1111/j.1467-968X.2005.00157.x.

Leu, T. (2015). "Generalized x-to-C in Germanic". *Studia Linguistica*, 69(3), 272-303. https://doi.org/10.1111/stul.12035.

Merlo, P. (2016). "Quantitative Computational Syntax: Some Initial Results". *IJCoL. Italian Journal of Computational Linguistics*, 2(2-1), 1-20. https://journals.openedition.org/ijcol/347.

Merlo, P.; Samo, G. (2022). "Exploring T3 Languages with Quantitative Computational Syntax". *Theoretical Linguistics*, 48(1-2), 73-83. https://doi.org/10.1515/tl-2022-2032.

Merlo, P; Samo, G. (forthcoming). "Generative Computational Modelling". Leivada, E.; Grohmann, K.K. (eds), *The Cambridge Handbook of Minimalism and Its Applications*. https://lingbuzz.net/lingbuzz/007675.

Obenauer, H. (2001). *Alternative Checkers in the Left Periphery of Pagotto*. Ms., CNRS, Paris.

Paoli, S. (2007). "The Fine Structure of the Left Periphery: COMPs and Subjects". *Lingua*, 117(6), 1057-79. https://doi.org/10.1016/j.lingua.2006.05.007.

Poletto, C. (1995). "Complementizer Deletion and Verb Movement in Italian". *Working Papers in Linguistics*, 5(2), 1-15.

Poletto, C. (2014). *Word Order in Old Italian*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660247.001.0001.

Rizzi, L. (1997). "The Fine Structure of the Left Periphery". Haegeman, L. (ed.), *Elements of Grammar*. Dordrecht: Springer Netherlands, 281-337. https://doi.org/10.1007/978-94-011-5420-8_7.

Rizzi, L.; Bocci, G. (2017). "Left Periphery of the Clause: Primarily Illustrated for Italian". Everaert, M.; Riemsdijk, H.C. (eds), *The Wiley Blackwell Companion to Syntax, Second Edition*. Hoboken, NJ: John Wiley & Sons, 1-30. https://doi.org/10.1002/9781118358733.wbsyncom104.

Samo, G. (2019a). *A Criterial Approach to the Cartography of V2*. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Samo, G. (2019b). "Cartography and Locality in German: A Quantitative Study with Dependency Structures". *Rivista Di Grammatica Generativa/Research in Generative Grammar*, 41(5), 1-26.

Samo, G. (2022). "Moved to ModP or Base-generated in FrameP? A Quantitative Cartographic Study in Romance". *Revue Roumaine de Linguistique*, LXVII, 4, 345-61.

Samo, G. (2023). "Testing Cartographic Proposals on Locality Effects in V2: A Quantitative Study". *Journal of Historical Syntax*, 7(24), 1-32.

Samo, G.; Merlo, P. (2019). "Intervention Effects in Object Relatives in English and Italian: A Study in Quantitative Computational Syntax". Chen, X.; Ferrer-i-Cancho, R. (eds), *Proceedings of the First Workshop on Quantitative Syntax* (Quasy, SyntaxFest, Paris, 26-30 August 2019), 46-56. https://aclanthology.org/W19-7906.pdf.

Samo, G.; Merlo, P. (2021). "Intervention Effects in Clefts: A Study in Quantitative Computational Syntax". *Glossa: A Journal of General Linguistics*, 6(1), 1-39. https://www.glossa-journal.org/articles/10.16995/glossa.5742/.

Samo, G.; Si, F. . (2022). "Optionality of De in Chinese Possessive Structures: A Quantitative Study". *Quaderni di Linguistica e Studi Orientali*, 8, 37–53. https://doi.org/10.13128/qulso-2421-7220-13602.

Sanguinetti, M. et al. (2018). "PoSTWITA-UD: An Italian Twitter Treebank in Universal Dependencies". Calzolari, N. et al. (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Phoenix Seagaia Conference Center Miyazaki, Japan, 7-12 May 2018), 1768-75. https://aclanthology.org/L18-1279.pdf.

Van Craenenbroeck, J.; Van Koppen, M. (forthcoming). "Quantitative Approaches to Syntactic Variation". Barbiers, S.; Corver, N.; Polinsky M. (eds), *The Cambridge Handbook of Comparative Syntax*. Cambridge: Cambridge University Press, 1-23. http://jeroenvancraenenbroeck.net/s/CUP_chapter_qual_quant.pdf.

Vikner, S. (1994). "Finite Verb Movement in Scandinavian Embedded Clauses". Hornstein, N.; Lightfoot, D. (eds), *Verb Movement*. Cambridge: Cambridge University Press, 117-47.

Wilcox, E.G.; Futrell, R.; Levy, R. (2024). "Using Computational Models to Test Syntactic Learnability". *Linguistic Inquiry*, 55(4), 805-48. https://doi.org/10.1162/ling_a_00491.

Xu, D. (1993). "A CP Analysis of Mandarin Chinese". *Linguistics in the Netherlands*, 10(1), 189-200.

Zanuttini, R. (1997). *Negation and Clausal Structure: A Comparative Study of Romance Languages*. Cary: Oxford University Press, Incorporated.

Appendix

Table 7 Raw data.

Corpus	**# sentences**	**Bridge_ Tot**	**Bridge_ Comp**	**Br_ CompSUBAux**	**Br_ CompSUBLex**	**Br_V2C_ SUBAux**	**Br_V2C_ SUBLex**
UD_Italian-ISDT@2.13	14167	139	112	31	18	13	2
UD_Italian-VIT@2.13	10087	157	134	8	4	1	1
UD_Italian-PoSTWITA@2.13	6712	217	149	25	33	13	20
Total ITA	30966	513	395	64	55	27	23
UD_German-HDT@2.13	189928	763	591	0	0	0	0
UD_German-GSD@2.13	15590	46	24	4	0	3	1
UD_German-LIT@2.13	1920	23	13	0	0	0	0
Total DEU	207438	832	628	4	0	3	1
UD_Italian-Old@2.13	1228	68	50	8	15	1	1
		Lic_Tot	Lic_Comp	Lic_ CompSUBAux	Lic_ CompSUBLex	Lic_ VerbSUBAux	Lic_ VerbSUBLex
UD_Italian-ISDT@2.13		8	8	1	0	0	0
UD_Italian-VIT@2.13		20	18	0	0	0	0
UD_Italian-PoSTWITA@2.13		8	7	1	1	0	0
Total ITA		36	33	2	1	0	0
UD_German-HDT@2.13		57	46	0	0	0	0
UD_German-GSD@2.13		2	2	0	2	0	0
UD_German-LIT@2.13		0	0	0	0	0	0
Total DEU		59	48	0	2	0	0
UD_Italian-Old@2.13		5	5	0	0	0	0

Corpus	# sentences	SUBAux	SUBLex	Aux_Tot	Lex_Tot
UD_Italian-ISDT@2.13	14167	412	1093	9468	1093
UD_German-GSD@2.13	15590	539	769	7142	769
UD_Italian-Old@2.13	1228	79	430	1067	430