ISSN 2385-4138 (digital)
https://doi.org/10.5565/rev/isogloss.346

To be or *to have*? That is the question: Auxiliary Selection in Italian

Greta Viale
University of Verona & Sorbonne Université
greta.viale@univr.it

Andrea Briglia
Univ. Grenoble Alpes, CNRS
andrea.briglia@grenoble-inp.fr

Massimo Mucciardi
University of Messina
massimo.mucciardi@unime.it

Anne Carlier
Sorbonne Université
anna.carlier@sorbonne-universite.fr

Received: 12-04-2023
Accepted: 19-02-2024
Published: 09-05-2024

How to cite: Viale, Greta, Briglia, Andrea, Mucciardi, Massimo, & Anna Carlier. 2024. *To be* or *to have*? That is the question: Auxiliary selection in Italian. RLLT22, eds. Anna Gavarró, Jaume Mateu, Jon Ander Mendia & Francesc Torres-Tamarit. Special Issue of *Isogloss. Open Journal of Romance Linguistics* 10(3)/7, 1–30.
DOI: https://doi.org/10.5565/rev/isogloss.346

Abstract

For the first time, systematic research of auxiliary selection in Italian is proposed using corpus analysis and natural language processing (NLP). By combining these methods,

we seek to find the most significant factors that influence the choice of auxiliary in intransitive verbs with double auxiliation. These verbs have often been studied in the literature (e.g., *peripheral verbs* [Sorace 2000]), but they have never been addressed in a comprehensive way (Giancarli 2015). The findings emphasize the most significant factors influencing the choice of 'be' or 'have' based on semantic, syntactic, and morphological aspects. On the basis of corpus analysis and statistical tools (CHAID and Random Forest) evidence, we propose the internal cause and the human trait as the possible factors useful in untangling the knot of auxiliary selection within Italian verbs with double auxiliation. This article also presents a reflection on semi-auxiliary verbs, a particular group of Italian verbs that operate as semi-auxiliary by being followed by an infinitive. For this group of verbs, we propose that auxiliary selection depends not only on the semantics of the verb or of the subject, but mainly on the auxiliary selection of the infinitive.

Keywords: auxiliary selection, Italian, NLP, morphosyntax, semantics.

1. Research question

Auxiliary selection is a well-known phenomenon comprising the alternation of 'have' and 'be' in the perfect tense. In Italian, as in certain other Romance languages (e.g., French), the perfect tense can be formed either with the auxiliary 'be' or with the auxiliary 'have'. The selection of the auxiliary depends in part on the syntactic construction of the verb. Transitive verbs form the perfect tense invariably with 'have':

(1)	Giorgio	**ha**	**comprato**	una	nuova	casa[1].
	G.	have.PRS.3SG	buy.PTCP	INDF.ART	new	house
	'Giorgio bought a new house.'					

The phenomenon of auxiliary split therefore only concerns intransitive verbs. Italian is a particularly interesting expression of the phenomenon. Not only does it have intransitive verbs that take either the auxiliary 'be' or the auxiliary 'have' in the perfect tense, but it also has certain verbs that can take both auxiliaries. A case in point is the Italian verb *saltare* 'to jump'.

(2) a.	Veronica	**ha**	**saltato**	sul	letto[2].
	Veronica	have.PRS.3SG	jump.PTCP	on.DEF.ART	bed
	'Veronica jumped on the bed.'				

b.	Veronica	**è**	**saltata**	sul	letto[3].
	Veronica	be.PRS.3SG	jump.PTCP.F.SG	on.DEF.ART	bed
	'Veronica jumped on the bed.'				

The variation of auxiliary selection raises several questions. Why do the auxiliaries 'be 'and 'have' compete for selection in some languages? Which verbs

[1] Data are ours (speakers come from Piedmont and Tuscany).
[2] Data are ours (speakers come from Piedmont and Tuscany).
[3] Data are ours (speakers come from Piedmont and Tuscany).

select ‘be’ rather than ‘have’? For languages that have this double auxiliation, is the class of verbs selecting the auxiliary ‘be’ rather than ‘have’ similar from one language to another?

Research carried out on these questions has highlighted that auxiliary split is particularly relevant to the linguistic theory, by “standing at the intersection between syntax, lexical and clausal semantics and morphology” (McFadden 2007: 674). However, there is no consensus on how these different dimensions interact or are correlated, as demonstrated by the major syntactic (Perlmutter 1978) and semantic hypotheses on auxiliary split (Van Valin 1990, Sorace 2000, Bentley and Eythòrsson 2004).

So far, the research on auxiliary selection appears to have focused on intransitive verbs that select dominantly either ‘have’ or ‘be’ and aims to identify the elements that play a role in the selection of the auxiliary. The work of Sorace (2000) plays a pivotal role by proposing an auxiliary selection hierarchy (*ASH*), later renamed Split Intransitivity Hierarchy. The hierarchy is built on the concept of gradience. The core-verbs are found at the two extremes of the hierarchy: at one end we find the verbs that correlate the most with ‘be’ because of telicity and at the other we find those that correlate the most with ‘have’ because of agentivity. The center consists of the verbs most subjected to auxiliary selection, the so-called *peripheral verbs* (Sorace 2000:860).

Studies are typically based on acceptability assessments and selected instances, the technique of collection of which is not always specified (see Sorace 2000). Acceptability assessments are a major source of data, and this data was frequently acquired using sound experimental procedures based on statistical generalizability and supplemented with various sorts of outcomes.

However, if such studies fail to assess how much different factors matter for different verbs in occurrences, parameter choices can be inconsistent (e.g., *agentivity* for *prevalere* ‘prevail’, *animacy* for non-volitional process verbs as *tentennare* ‘waver’, see Sorace 2000), which makes unclear what impact each factor may have on the choice of the auxiliary.

Moreover, as Giancarli (2015) demonstrates, many studies on auxiliary selection fail to investigate how particular factors affect the so-called *peripheral verbs*. The main studies on auxiliary selection do not analyze intransitives with double auxiliation in regard to the factors that determine ‘have’ or ‘be’.

By focusing on these peripheral verbs most subject to auxiliary alternation, our approach fills this gap in previous studies. Thanks in part to the tools now available to us, we aim to carry out a quantitative study of the relative frequency of both auxiliaries. We considered a corpus-driven analysis to respond to the constraint of reproducibility of the results. We will consistently analyze the verbs according to the same set of parameters, allowing to demonstrate how much weight certain factors have for each verb and how they globally influence the auxiliary selection for verbs allowing both auxiliaries.

After describing the method of analysis, the purpose of this work is to demonstrate how the concept of *internal cause* (Mateu 2009), rather than agentivity, is important in auxiliary selection when applied at the level of *peripheral verbs.*

2. Methodology: sample size and representativeness

Before any kind of scientific generalization, it is necessary to reflect upon the representativeness of the sample used. Auxiliary selection seems to deal with many observations: in Italian, the number of sentences containing auxiliaries cover a range that is difficult to estimate. A reliable and robust model must consider as many sentences as possible to provide a trustworthy explanation both for sentences that are included in the corpus and for those that are not included. As is usual in the field of linguistics, the requirement of having a big sample is in tension with the time needed for manual annotation.

As far as the authors know, the way in which the sentences have been annotated in this work represents a rather qualitative interpretation that would not have been possible to do automatically. There is no NLP library capable of automatically attributing any of the parameters that have been used in this work.

To give an example, the authors of this paper are unaware of an NLP-task capable of spotting the difference between sentences (6) and (7) in the same way that they have been delineated in this paper, despite the huge advances in NLP tasks. It is thus pertinent to reflect upon the relation between the sample considered and its respective population (*i.e.*, the whole language).

Crucial is the relation between the rate of occurrence of a given auxiliary form in the texts available in Sketch Engine (Jakubíček 2013) and the size of our sample[4]. This relation gives us a way to estimate the probability of having targeted a statistically representative number of occurrences of a given phenomenon.

It is known that word frequency distributions in any human language follow a power-law called "Zipf's law" (Zipf 1949). This has been verified more recently and with new insights by Ferrer i Cancho and Solé (2003). Verbs follow the same frequency distribution: a small number of verbs are highly frequent, and a large number of verbs are quite rare.

A general picture of frequency distribution of verbs in Italian can be drawn on the basis of some basic statistics of the chosen corpus (ItTenTen16). The corpus numbers 4989729171 tokens, of which *essere* is by far the most frequent verb, with 69621692 occurrences. *Avere,* meanwhile, is the fourth most frequent verb, accounting for roughly four and a half times fewer occurrences than *essere* with 14876951 occurrences.

Among the thirteen verbs included in our list, *continuare* is the most frequent, holding the 43rd position in the rank with 1970184 occurrences; it is followed by *iniziare*, holding the 53rd position with 1653634 occurrences.

How can we assess the representativeness of these differently occurring verbs in relation to the size of our sample?

In fact, all the verbs considered occur frequently enough to allow the auxiliary alternation variability to be caught in a sample of the same size, comprising 100 occurrences per verb (50 transitive and 50 intransitive verbs). To respond to the above questions, the observed differences in the frequency of occurrence therefore will not bias the results.

[4] Sketch Engine is a corpus manager tool for text analysis. Unless specified otherwise, all our data comes from the repository of this software, in particular from the ItTenTen(16) corpus mentioned in the article.

In comparison with non-peripheral verbs, *peripheral verbs* show a quite large range of different contexts of use. For this reason, we used the CQL query on SketchEngine. This tool allowed us to obtain a sample covering the full range of different possible contexts in order to be robust and reliable enough for accurately predicting which kind of auxiliary a given verb would have in a given sentence. A random sample would not have been guaranteed to target all the possible occurrences.

The CQL query function provided by Sketch Engine enables us to address the problem of capturing enough targets in any given verb sample. By using this function, we can look for the verbs we want to analyze, reducing in this way the inherent variability of the corpus.

We checked whether the distribution of transitivity and intransitivity in sentences containing verbs that allow for auxiliary selection is significantly different according to the ki square test. We sampled 50 random occurrences obtained through a specific command in the Sketch Engine CQL query and we manually annotated the two features (type of auxiliary and (in)transitivity) in a spreadsheet. We then computed the ki square test by using JASP[5]. For the verbs *contare, cambiare, cedere, cominciare, continuare, diminuire, fallire, iniziare, procedere, suonare* the distribution is significantly different, while for other verbs such as *finire, galleggiare, pesare, proseguire* the distribution of the features is not significantly different.

By way of illustration, table 1 presents the contingency table (cross tabulation) for the verb *cambiare*. There are 12 transitive sentences with ‘have’ and 38 intransitive sentences with ‘be’ (Chi-square (1) = 50.00; p-value < 0.001).

Table 1. Contingency table for *cambiare*

Contingency Table			
	TRANS		
AUX	**0=intransitive**	**1=transitive**	**Total**
0=have	0	12	12
1=be	38	0	38
Total	38	12	50

This information allows us to contextualize the results developed in our paper: before starting to analyze how intransitive verbs select the auxiliary, we wanted to know how auxiliaries are distributed among those verbs (as well as other control verbs) and verify how (in)transitivity is distributed between the two alternative forms of perfect tense.

To conclude, we have described the inherent variability of the verbs that will be analyzed, and we have demonstrated how the *corpus* we have built can be considered as statistically representative of the phenomenon under investigation.

[5] JASP Team (2023). JASP (Version 0.18) [Computer software].

3. Corpus analysis

3.1. Data collection

The strength of our analysis lies in the large quantity of data examined. By analyzing around a thousand sentences, we have detected unexpected data, and uncovered recurrent patterns, which would be impossible to trace on the basis of a smaller sample. ItTenTen(16), a corpus available on SketchEngine, provided us a large number of occurrences of the linguistic phenomenon under study, drawn from texts representing a variety of registers. Corpus Query Language (CQL) option was used to find several intransitive verbs, and for each of them we then analyzed 50 occurrences with 'have' and 50 with 'be'.

This number of items was chosen because it was necessary to have enough data for a reliable analysis, including the parameters that condition the use of one or the other auxiliary in the *peripheral verbs* we have considered. Thus, we choose to analyze 13 verbs (*cambiare* 'change', *continuare* 'continue', *iniziare* 'begin', *cominciare* 'begin', *contare* 'count' 'matter', *suonare* 'ring' 'sound', *procedere* 'proceed', *proseguire* 'proceed', *cedere* 'surrender', *fallire* 'fail', *finire* 'finish', *pesare* 'weight' 'matter', *prevalere* 'prevail'), for a total of 1583 occurrences studied (1183 main verbs and 400 semi-auxiliaries). Four of these verbs are also used as (semi-) auxiliaries[6], in combination with an infinitive. A case in point is the verb *continuare* 'continue':

(3) Questa percentuale **ha** **continuato** **a** crescere.
This percentage have.PRS.3SG continue.PTCP to grow.INF
'This percentage has continued to grow.'

(4) Il prezzo **è** **continuato** **a** salire.
DEF.ART price be.PRS.3SG continue.PTCP to rise.INF
'The price has continued to rise.'

In both examples, the subject is inanimate, and combines with an intransitive verb that selects 'be' in Italian. However, in (3), 'have' is chosen and in (4), 'be' is chosen. We hypothesize that the type of infinitive that follows the semi-auxiliary affects the choice of the auxiliary itself, as will be demonstrated later. If this is true, the aspectual auxiliary verb is transparent to the 'have/be' selection of the infinitive. The feature of transparency with respect to auxiliary selection suggests an advanced grammaticalization as an auxiliary.

3.2. Parameter Analysis

We decided to consider various parameters that could influence the selection of the auxiliary. These parameters, listed in table 2, are either semantic or syntactic (sometimes further distinguished on an inherently semantic basis). They have been examined for each sentence. Additionally, we examined a third factor that possibly could

[6] We are aware of the 'restructuring' phenomenon discussed in the literature (Cinque 2004, Rizzi 1982). However, for the sake of this paper, we use the terminology 'semi-auxiliary' to emphasize a peculiarity of this group of verbs analyzed in our corpus regarding transparency. The larger phenomenon of restructuring is not addressed here.

influence auxiliary selection, namely the morphological markedness of the participle because of the agreement with the subject.

Table 2. Parameters (main verbs).

Semantic parameters		**Syntactic parameters (Adverbials)**		**Syntactic parameters (Others)**
Subject [+ Human]	***Subject [+ Animate]***	***[+ Adverbial of manner]***	***[+ Adverbial of time]***	***[+ Zero] (Free-adverbials context)***
Subject [+ Agentive]	***Subject [+ Internal cause]***	***[+ Adverbial of quantity]***	***[+ Adverbial of time + duration]***	***[+ Agreement subject – past participle]***
		[+ Argument Adverbial]	***[+ Locative Adverbial] Static***	***[+ Direct Object implied]***
		[+ Aspectual Adverbial] Telic	***[+ Locative Adverbial] [-endpoint]***	
		[+ Aspectual Adverbial] atelic	***[+ Locative Adverbial] [+endpoint]***	
		[+ Aspectual Adverbial] unmarked		

Some specific parameters deserve more attention. A first case is represented by verbs that allow an implicit Direct Object (DO). While we have tried to exclude transitive phrases, some sentences, without an explicit DO and/or followed by an adverb are ambiguous. For example, in (5) *molto* can be both an adverb or a pronoun.

(5) Luigi **ha** **cambiato** molto.
Luigi have.PRS.3SG change.PTCP ADV (1) / ADJ (2)
'Luigi has changed a lot' (1)/ 'Luigi has changed many things' (2)

According to the latter reading, the sentence means 'Luigi has made numerous changes' and, thus *cambiare* is analyzed as a transitive construction. This is *a priori* the preferred interpretation with 'have', but we cannot completely rule out the other option without considering more data. The fact that many of these verbs can be used in both transitive and intransitive contexts increases the possibility of ambiguity.

We decided not to exclude these data entirely because analyzing them with actual transitive uses allows us to better understand which parameters influence selection, effectively acting as a sort of control test. This allows us to draw the thin line that separates transitive and intransitive interpretation. Starting from syntactic contexts, we considered the numerous syntactic situations in which the occurrences could be found, namely the adverbial of manner, the adverbial of quantity and the argument adverbial. This last adverbial can be described as an argument that cooccurs regularly with the verb, even when it should have been an expression that has become fixed.

An important distinction has been made between aspectual and temporal adverbials. For the aspectual adverbials, we distinguish between telic, atelic and

unmarked adverbials. As for temporal adverbials, a distinction is established between simple temporal adverbials and temporal plus duration adverbials, which do not solely perform a temporal localization but also provide information on the internal composition of the verbal situation, specifically its duration (e.g., *da allora a oggi* 'from then to now'). Among spatial adverbials we found locational adverbials and directional adverbials, with and without endpoints.

It is important to consider syntactic parameters for semi-auxiliary verbs, especially the constructional properties of the infinitive before we delve into the semantic criteria they share. The infinitive can be transitive, or intransitive. Among the intransitives, those selecting 'have' (e.g., *lavorare* 'work') have been differentiated from those selecting 'be' (e.g., *arrivare* 'come').

Table 3. Parameters (semi-auxiliaries).

Semantic factors		**Syntactic factors**	
Subject [± Human]	***Subject [± Animate]***	***V AUX + V INF Trans***	***V AUX + V INF Intransitive Aux. 'have'***
Subject [± Agentive]	***Subject [± Internal cause]***		***V AUX + V INF Intransitive Aux. 'be'***

In terms of semantic parameters, we considered the subject's semantics, resulting in the traits [± Human], [± Animate], [± Agentive] and [± Internal Cause]. There are two crucial distinctions to make: animacy *versus* agentivity (6), and agentivity *versus internal cause* (7). Animacy and agentivity are related, but while it is impossible for inanimate subjects to be agentive, not all animate subjects are agentive. This is where our analysis adds a second division, between agentivity and *internal cause*. For this reason, we decided to deepen the analysis by adding the difference between agentivity and *internal cause*, which we interpreted as a subset of agentivity.

We adopted the concept of *non-volitional internal cause* from Mateu (2009), who accounts for intransitives that choose 'have' by means of this concept rather than agentivity. We decided to apply it not only to verbs that ordinarily select 'have', but also to verbs showing split intransitivity, thus combining Mateu's concept with Reinhart's (2000, 2002) hypothesis defining agentivity by the presence of 'c' and 'm', namely 'cause change' and 'mental state' involved.

In our analysis, therefore, *internal cause* can thus be a trait of an instrument, a natural force, or an intrinsic property of the subject based on the semantics of the verb. As a result, an agentive subject must be characterized by *internal cause* (since we view agentivity as volition plus cause), whereas *internal cause* does not imply agentivity.

Consider for example the following sentences:

(6)[7] Il vento **ha** **aperto** la porta.
DEF.ART wind have.PRS.3SG open.PTCP DEF.ART door
'The wind has opened the door'

[7] Data are ours (speakers come from Piedmont and Tuscany).

(7)[8] Lucia **ha** **contato** molto per me.
L. have.PRS.3SG matter.PTCP a lot for me
'Lucia has mattered a lot to me.'

In (6) the wind performs the action and thus has power over it, but it does not have voluntary control over it, whereas in (7) Lucia wields considerable power over the experiencer, but she could be completely unaware of it. Both subjects are responsible for causative – but not agentive – behavior.

4. Predicting Parameters

4.1. Verb semantics, agentivity and internal cause

The distinction between agentivity and *internal cause* was crucial in our investigation into the parameters that influence the choice of 'have'. Indeed, we observed split intransitivity in the 13 verbs studied in the absence of agentivity. 'Have', as is well known, corresponds more easily to agentivity and 'be' is correlated to non-agentive subjects. *Internal cause*, on the other hand, permits us to explain not only why some verbs that normally select 'be' occasionally select 'have', but also why verbs with non-agentive subjects choose 'have' by default.

Consider the verb *cambiare* 'change' as an example of the first situation.

CAMBIARE

(8) La mia vita **è** **cambiata** da quando
DEF.ART my life be.PRS.3SG change.PTCP.F.SG since when
ho deciso di interessarmi a questa vicenda.
have.PRS.1SG decide.PTCP to take.an.interest.INF to this matter
'My life has changed since I decided to take an interest in this matter.'

(9) Quando i padri mi hanno regalato la
When DEF.ART fathers to me have.PRS.3PL give.PTCP DEF.ART
sedia a rotelle, la mia vita **ha** **cambiato** […]
wheelchair DEF.ART my life have.PRS.3SG change.PTCP
'When the mission fathers gave me the wheelchair my life changed […].'

We considered the possibility of interpreting a transitive meaning with 'have' in (9), but when comparing these two sentences, we are obliged to deny this option: the only difference between the two sentences is the auxiliary. If we look at two additional sentences, the line is even more subtle.

[8] Data are ours (speakers come from Piedmont and Tuscany).

(10) La storia **ha** **cambiato** un po', invece di
DEF.ART story have.PRS.3SG change.PTCP a bit instead of
salvare le uova di Yoshi, Mario dovrà trovarle.
save.INF DEF.ART eggs of Y. M. need.FUT.3SG find.INF.them
'The story has changed a bit, instead of saving Yoshi's eggs, Mario will have to find them.'

(11) Durante il Cretaceo,[…] la situazione **ha**
during DEF.ART Cretaceous DEF.ART situation have.PRS.3SG
cambiato molto, sia pure gradatamente.
change.PTCP a lot even if albeit gradually
'During the Cretaceous, […] the situation changed greatly, albeit gradually.'

Both (10) and (11) could in principle have an ambiguous interpretation (provided by 'have' combined to a quantifying adverbial) between transitive and intransitive construction, but we will rather consider *cambiare* as intransitive in both cases. Our analysis relies on the *ne*-cliticization test proposed by Burzio (1986): the Italian partitive pronoun *ne* 'of it/them' can be used in place of an internal argument: the direct object of a transitive verb or the unique argument of unaccusatives (Bentley 2003: 221), provided that this internal argument is indefinite. *Ne*-cliticization does not support the analysis of *cambiare* as transitive in (10) and (11), as evidenced by (12a), (12b), (13a) and (13b):

(12) a. *Ne **ha** **cambiato,** la storia.
of.it have.PRS.3SG change.PTCP DEF.ART story
'It changed it, the story'
b. ? Ne **ha** **cambiate** un po', la storia.
of.it have.PRS.3SG change.PTCP.F.PL a few (things) DEF.ART story
'It has changed some, the story.'

(13) a. *Ne **ha** **cambiato**, la situazione.
of.it have.PRS.3SG change.PTCP DEF.ART situation
'It changed it, the situation.'
b. ?Ne **ha** **cambiate** molte, la situazione.
of.it have.PRS.3SG change.PTCP.F.PL many (things) DEF.ART situation
'It has changed some, the situation.'

Ne-cliticization in (12a) and (13a) is ungrammatical, and hence *cambiare* is not transitive in these examples. If we accept and make explicit the transitivity of (10) and (11) by assuming *un po'* 'a little' as un *po' di cose* 'a few things' in (12b) and *molto* as *molte cose* 'many things' in (13b), the sentence would be more grammatical, but would still make no sense with the rest of the phrase, confirming the non-transitivity.

This is particularly highlighted in (11) by the presence of *gradatamente.* This adverb highlights that the subject is undergoing change, and hence excludes the possibility of an object complement. The non-agentive and non-human subject intensifies this reading: we do not exclude this type of subject occurring in transitive sentences, but in the case of a verb such as *cambiare* that selects almost exclusively 'be' in its intransitive reading, we expect the auxiliary 'have' to be more likely to occur when the subject is human and agentive or considered as 'causer'. Indeed, according

to the analysis, 'be' is the default choice for the intransitive use of *cambiare* (as it has been shown in Table 1). Witness example (14):

(14) Adesso la situazione **è** **cambiata**.
Now DEF.ART situation be.PRS.3SG change.PTCP.F.SG
'Now the situation has changed.'

If we compare (10) and (11) with (14), we infer that 'have' may be used when the subject is non-agentive. However, the subject can present an ambiguity that prevents it from being clearly analyzed either due to the porous boundary with its transitive construction or because of internal causality.

We argue that when the subject conveys the feature of internal causality, the use of 'have' is more plausible, as evidenced with the quantitative data (figure 1 & table 4) and sentences below (15,16,17).

Figure 1. Percentages of data presenting lack of agentivity and *internal cause* in *cambiare.*

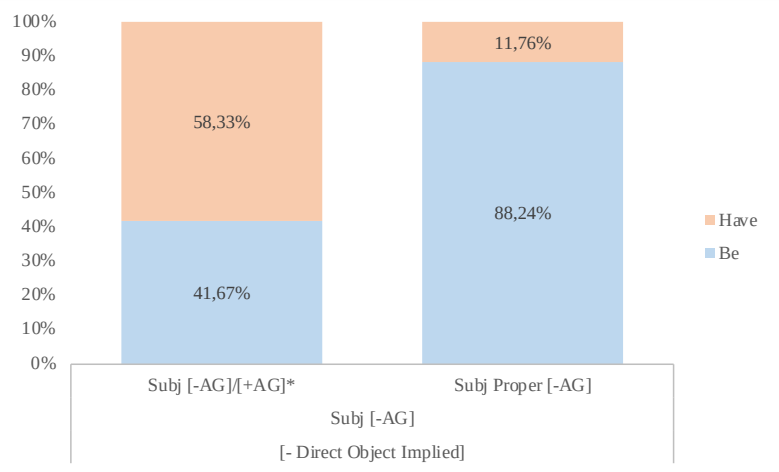


Table 4. Data presenting lack of agentivity and *internal cause* in *cambiare.*

Parameters	**Aux. 'be'**	**Aux. 'have'**	**Total**
[-DO] Subj [-AG] [-AG]/[+AG] *	**5**	**7**	**12**
[-DO] Subj Proper [-AG]	45	6	51
Total	*50*	*13*	***63***

In fact, we first distinguished agentives from non-agentives, and then non-agentives were further subdivided into pure non-agentives (Subj proper [-AG]) and subjects who could propose an ambiguous interpretation (Subj [-AG]/[+AG] *= *internal cause*), which occurs when the subject does not convey the feature of volition but can somehow control the action.

(15) Nel frattempo **è** **cambiato** anche caratterialmente.
in.DEF.ART meantime be.PRS.3SG change.PTCP also temperamentally
'In the meantime, he has also changed temperamentally.'

(16) Il suo sguardo **ha** **cambiato** e da
DEF.ART his gaze have.PRS.3SG change.PTCP and from
sospettoso è diventato curioso.
suspicious be.PRS.3SG become.PTCP curious
'His gaze has changed and from suspicious became curious.'

(17) Ma mai l'animale-uomo **ha** **cambiato**, e
But never DEF.ART.animal-man have.PRS.3SG change.PTCP and
cambierà, nelle sue forme di sostentamento.
change.FUT.3SG in.DEF.ART its forms of sustenance
'But never has animal man changed, and will change, in his forms of sustenance.'

Even in these sentences, we have interpreted these subjects as having *internal cause* because it is impossible to identify whether they are rather *affected by change* or rather *causes of change*. This is the crucial point: they can neither be considered as pure agentive subjects nor as non-agentive subjects, because of their causative meaning.

The concept of *internal cause*, although it relies on interpretation, allows us to understand why certain verbs tend to select 'have'. After this brief study of a verb that predominantly selects 'be' as an auxiliary, namely *cambiare*, we will now take a closer look at a verb with 'have' as a default auxiliary, namely *contare* 'count'.

CONTARE

(18) I nostri clienti **hanno** **contato** ripetutamente
DEF.ART our clients have.PRS.3PL count.PTCP repeatedly
su di noi per soddisfare o superare le loro aspettative.
on of us to meet.INF or exceed.INF DEF.ART their expectations
'Our clients have repeatedly counted on us to meet or exceed their expectations.'

(19) Una donna che nella sua vita **ha**
a woman that in.DEF.ART his life have.PRS.3SG
contato molto e con cui vorrebbe tornare.
count.PTCP a lot and with whom want.COND.3SG return.INF
'A woman who counted a lot in his life and with whom he would like to return.'

(20) Il pubblico di Vasco era lì solo
DEF.ART audience of V. be.IPFV.3SG there only
il resto **ha** **contato** poco.
DEF.ART rest have.PRS.3SG count.PTCP little
'Vasco's audience was there only for him; the rest counted for little.'

(21) Solo una cosa **era** **contata** più di
only one thing be.IPFV.3SG count.PTCP.F.SG more than
qualunque altra.
any other
'Only one thing counted more than any other.'

(22) Temo di dichiararle tutto il mio amore,
be.afraid.PRS.1SG to declare.INF.to.her all DEF.ART my love,
di farle capire che nessun'altra in cinque anni
to make.INF.her understand.INF that no.one.else in five years
è **contata** veramente più di lei.
be.PRS.3SG count.PTCP.F.SG really more than her.
'I am afraid to declare all my love to her, to make her understand that no one else in five years has really counted more than her.'

Only the first of these sentences contains a fully agentive subject to whom intentionality can be imputed (*clients* repeatedly exercise this voluntary act of trust). In other circumstances, whether the subject is human and animate (19 and 22) — the *woman* exerts influence, but lacks will — or inanimate (20 and 21), the semantics of the verb itself provides an *internal cause*. Because it 'matters', the subject exerts a certain amount of *force* even though there is no intention.

The borderline is sometimes very thin. However, linking back to earlier discussion of *control*, we argue that the subject of *contare* does not have to be agentive — especially when it is inanimate and thus cannot be agentive but is more likely to be characterized as *causative*.

The data represented in Figure 2 and Table 5 show that all subjects are either fully agentive or characterized by *internal cause*. The presence of 'be' appears only in the latter case, where the interpretation may be equivocal.

Figure 2. Percentages of data presenting lack of agentivity, *internal cause* and human trait in *contare*.

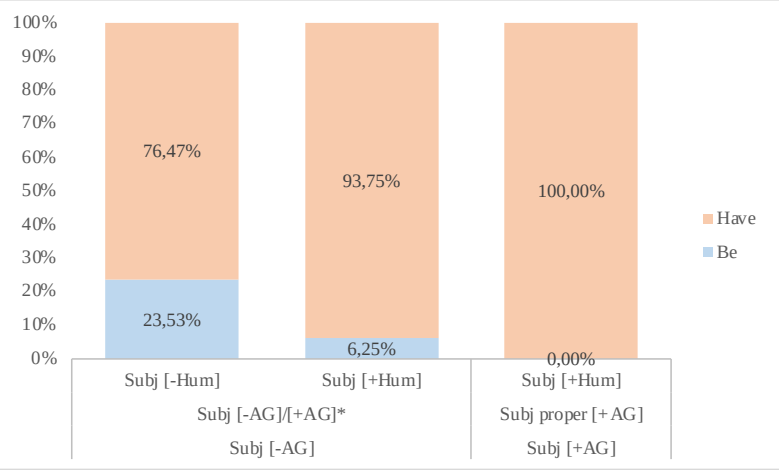


Table 5. data presenting lack of agentivity, *internal cause* and human trait in *contare*.

Parameters	**Aux. 'be'**	**Aux. 'have'**	**Total**
Subj [-AG] [-AG]/[+AG] *[-Hum]	***8***	***26***	***34***
Subj [-AG] [-AG]/[+AG] *[+Hum]	1	15	16
Subj [+AG] Subj proper [+AG] [+Hum]	/	9	9
Total	*9*	*50*	***59***

Figure 2 and Table 5 include another parameter, namely the human trait. As shown before, auxiliary alternation only occurs when the subject is non-agentive. With the verb *contare*, the subject is systematically regarded as an *internal cause*, because of the semantics of the verb itself. Nevertheless, when the subject is non-human, the presence of 'be' increases. This tendency highlights that 'have' is linked to a semantics of control or 'cause', as well as the role of human trait in triggering this control interpretation.

The concept of *internal cause* is not only useful for verbs that dominantly select either 'be or 'have' but can help us explain circumstances where there is 50-50 auxiliary selection, as *suonare* 'ring'.

SUONARE

(23) L'allarme **ha** **suonato** alle 2.40 di stanotte.
DEF.ART.alarm have.PRS.3SG ring.PTCP at.DEF.ART 2:40 of tonight
'The alarm rang at 2:40 a.m. tonight.'

(24) L'allarme del negozio **è** **suonato** costringendoli a scappare.
DEF.ART-alarm of.DEF.ART store be.PRS.3SG ring.PTCP forcing.them to leave.INF
'The store alarm rang forcing them to leave.'

(25) Ma io vi dico che il momento è giunto, l'ora **è** **suonata**.
but I to.you say.PRS.1SG that DEF.ART time be.PRS.3SG come.PTCP DEF.ART.hour be.PRS.3SG ring.PTCP. F.SG
'But I say to you that the time has come, the hour has come.'

Although the same subject appears in (23) and (24), 'have' is selected in one example and 'be' in the other. We explain this difference through a possible interpretation of *internal cause*, outlined above: remember that the feature combination Subj [-AG]/[+AG] * indicates the ambiguity of the interpretation, i.e. the subject is non-agentive but can be considered as a *cause* in some sense. In fact, the verb *suonare* 'ring' lends itself well to this dual interpretation: we can think of the subject of ringing as the one creating the sound, and therefore regard it as that which *causes* the sound, but we can also conceive of the event as triggered by some external cause.

(25), in contrast, is viewed as not agentive and less causative because the sense is metaphorical and can be glossed as 'the time has come'; therefore, there is not the same possibility of considering the subject as responsible for the event.

Figure 3. Percentage of data presenting lack of agentivity and *internal cause* in *suonare.*

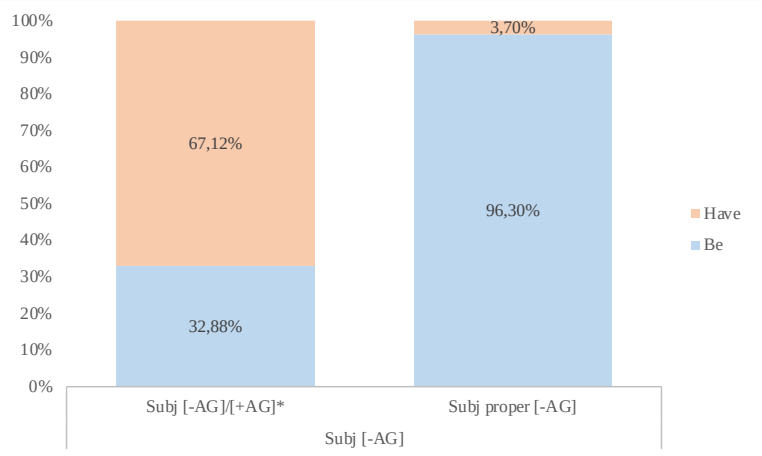


Table 6. Data presenting lack of agentivity and *internal cause* in *suonare.*

Parameters	**Aux. 'be'**	**Aux. 'have'**	**Total**
Subj [-AG] [-AG]/[+AG] *	***24***	***49***	***73***
Subj [-AG] Subj proper [-AG]	26	1	27
Total	*50*	*50*	***100***

For this verb, we only have inanimate subjects in our corpus, and we can observe how the percentage of 'have' and 'be' increases and drops depending on the *internal cause*, and vice versa.

4.2. The semantic features of the subject [+human] [+animate] and their interaction with internal cause

Having demonstrated the relevance of *internal cause*, we now turn to animacy as a parameter that may help to solve the puzzle. The reasoning may appear to be circular, making it difficult to assess these concepts. Human and animate subjects are more easily associated with agentivity, or at least with internal causation, and thus make 'have' more acceptable. However, the presence of 'have' itself tends to induce the transitive interpretation, unless we have elements that make transitive use obvious (see the examples with *cambiare* mentioned above).

The analysis is complicated by the fact that several of the *peripheral verbs* also have transitive use, and that the availability of the transitive construction may have an impact on auxiliary selection. The identification of the semantic features conveyed by the subject, although straight-forward at first sight, proves to be often complex for the data provided by the corpus.

First, there are subjects that, by themselves, would not be human, but become so through metonymy: this is the case with *sguardo* 'gaze' (see examples with *cambiare*, *occhi* 'eyes'). There is not much of a difference between (26) and (27): the eyes stand in for the person who, unwillingly, gives in. The feature of volition with respect to human subjects can also raise subtle distinction. In (27), we see a person who is hesitant to forgive (*il povero uomo* 'the poor man'), but eventually relents and forgives. Likewise, in (28), the person did not want to surrender but was forced to do

so due to *force majeure*. However, the selected auxiliary is 'be' in (27) and 'have' in (28).

CEDERE

(26) Quegli stupidi dei miei occhi **hanno** **ceduto**;
those stupid of.DEF.ART my eyes have.PRS.3PL surrender.PTCP
è durato solo un attimo […].
be.PRS.3SG last.PTCP only a moment
'Those stupid eyes of mine gave way; it lasted only a moment […].'

(27) […] finché il povero uomo **è** **ceduto** e
until DEF.ART poor man be.PRS.3SG surrender.PTCP and
ha deciso di perdonare […].
have.PRS.3SG decide.PTCP to forgive.INF
'[...] until the poor man broke down and decided to forgive […]'

(28) Vero, non ce l'ho fatta. Su questo punto **abbiamo** **ceduto.**
true not could.do.1SG.it on this point have.PRS.1PL surrender.PTCP
'True, I couldn't do it. On this point we gave in.'

Consider the verb *fallire* 'fail', instead: there are cases where the subject is objectively human and others where the interpretation is more complex.

FALLIRE

(29) […] il Sindaco **ha** **fallito**, meglio non far […] nulla.
DEF.ART mayor have.PRS.3SG fail.PTCP better not do.INF nothing
'[...] the mayor has failed, better to do […] nothing.'

(30) Noi ci siamo, e siamo intenzionati a proseguire
we LOC be.PRS.1PL and be.PRS.1PL intent.PTCP.1PL to continue.INF
una storia segnata da aziende […] convinte di superarci,
INDF.ART history marked by companies convinced to surpass.INF.us
ma che alla fine **hanno** **fallito**.
but that ultimately have.PRS.3PL fail.PTCP
'We are there, and we are intent on continuing a history marked by companies […] that were convinced they would surpass us but ultimately failed.'

(31) Aveva detto di avere a portata di mano, come
have.IPFV.3SG say.PTCP to have.INF on hand as
compratori, e invece la compagnia **è** **fallita.**
buyers and instead DEF.ART company be.PRS.3SG fail.PTCP.F.SG
'He said he had a group of [daring Italian patriots on hand] as buyers, but instead the company went bankrupt.'

(29) depicts a subject who is unmistakably human and animate, and who, though not willing to fail voluntarily, is nonetheless responsible for failure, and hence conveys the feature of *internal cause*.

(30) and (31), on the other hand, present an ambiguous case: in (30), 'company' could refer either to the physical company or to the people who work there. In this case, we prefer this latter interpretation for what it follows. In (31), however, the physical company (which has gone bankrupt) appears to be emphasized, and, as a result, to be considered inanimate. In (29) and (30), the interpretations correlate with internal cause because subjects can be classified as either internal cause or non-agentive in terms of failure semantics. (31), instead, presents a not agentive and not causative interpretation. An inanimate subject can still be defined by *internal cause* (as we saw with *contare* 'count'), but as shown in these examples, the more animate the subject, the more it can be thought of as actively causing something.

Another phenomenon to mention is *shifted intentionality*, which characterizes inanimate subjects with a degree of *internal cause.* Action is attributed to the inanimate item but refers to the people behind it. Somehow, there is a shift from the human referent to the inanimate subject, thus becoming internal cause. For example, the sentences (32), (33) and (34) contain inanimate subjects corresponding to nouns evoking an activity performed by humans, such as *studio* 'study', *ricerca* 'research', and *lavoro* 'work'. The connection with humans is less prevalent with subject such as 'trade', 'business' *attività* (35).

CONTINUARE

(32) Numerosi studi **hanno** **continuato** nel 1990 e
numerous studies have.PRS.3PL continue.PTCP in.DEF.ART 1990 and
anche dopo la morte del dottor Atkins.
even after DEF.ART death of.DEF.ART dr. A.
'Numerous studies continued in the 1990s and even after Dr. Atkins' death.'

(33) Dopo […] le ricerche **sono** **continuate**.
after DEF.ART research be.PRS.3PL continue.PTCP.F.PL
'After[...], the research continued.'

(34) Anche la ricerca **ha** **continuato** nel 2005.
even DEF.ART search have.PRS.3SG continue.PTCP in.DEF.ART 2005
'The search also continued in 2005.'

(35) Tali attività **sono** **continuate** e si
these activities be.PRS.3PL continue.PTCP.F.PL and REFL
sono consolidate nel corso del 2011.
be.PRS.3PL consolidate.PTCP.F.PL in.DEF.ART during of.DEF.ART 2011.
'These activities continued and were consolidated during 2011.'

Shifted intentionality is also relevant with subjects denoting instruments. A case in point are vehicles (36), which are inanimate but can convey by the features of the person driving them (hence, inanimate but *internal cause*).

(36) Dopo una breve sosta […], il veicolo **aveva**
After INDF.ART brief stop DEF.ART vehicle have.IPFV.3SG
continuato su... ma solo per un minuto.
continue.PTCP on but only for one minute.
'After a brief stop [...], the vehicle had continued... but only for a minute.'

Another instance of subjects endowed with the feature of *internal cause* are force of nature subjects, exemplified by (37).

(37) Il vento **è** **iniziato** a soffiare verso le 23.
DEF.ART wind be.PRS.3SG start.PTCP to blow.INF around DEF.ART 11p.m.
'The wind started blowing around 11 p.m. [...].'

In this case, the subject is inanimate and does no longer have any connection to human semantics or intentionality.

Finally, inanimate subjects may be endowed with the feature of *internal cause* because of the semantics of the verb.

PESARE

(38) Certo, l'assenza di questa autorità **ha** **pesato**.
certainly DEF.ART.absence of this authority have.PRS.3SG matter.PTCP
'Certainly, the absence of this authority mattered'.

(39) Ma un'assenza **è** **pesata** più di altre.
but INDF.ART.absence be.PRS.3SG matter.PTCP.F.SG more than others.
'But one absence mattered more than others.'

Pesare 'matter' is a verb whose semantics implies that the subject, even if inanimate, has power over the action without being aware of it. These examples clearly demonstrate how a subject characterized by *internal cause* can choose either 'have' or 'be', but nevertheless favours 'have' precisely because of this property of the subject (out of 75 cases with *pesare*, 50 are with 'have').

Table 7. List of subjects that can be *internal cause*

Parameters	**Type of subjects**	
[+Hum] [+Anim]	***People***	***Metonymical subj. (e.g. eyes)***
[-Hum] [-Anim]	***Instruments (e.g., vehicles)***	***Human activities (e.g. research)***
	Forces of nature (e.g., wind)	***Subj. with inherent internal cause because of verb semantics***

This table summarizes the subjects likely to be considered internal cause. In the cells in orange, we see human and animate subjects, divided into people and metonymical subjects. In the cells in green, we find the inanimate subjects divided into instruments, forces of nature, human activities, and subjects with inherent internal cause because of verb semantics.

Internal cause can be useful to explain the choice of some auxiliaries even with subjects that we would have hardly consider internal cause, for example the inanimate ones. Without realizing it, by talking we can attribute animate properties to inanimate things: this is what happens in (40).

(40)[9] Il negozio **ha** **chiuso** alle 20.
DEF.ART store have.PRS.3SG close.PTCP at.DEF.ART 8 p.m.
'The store closed at 8 p.m.'

This is a perfect example of internal cause: the subject is inanimate, but with a strong reference to a human subject. This is the reason why 'have' is selected.

4.3. The type of infinitive as predicting parameter for semi-auxiliary verbs

Among the verbs examined, four of them (viz. *continuare* 'continue', *iniziare* 'begin', *cominciare* 'begin', *finire* 'finish'), besides their use as main verbs, also function as semi-auxiliaries, that is, they partially perform the auxiliary function.

In earlier studies devoted to auxiliary selection of Italian verbs, these semi-auxiliary uses have never been considered separately. We put forward the hypothesis that the infinitive has a crucial impact on the choice of the auxiliary.

Consider the verb *iniziare* 'begin'.

INIZIARE

(41) […] chi da pochi mesi o giorni **ha** **iniziato** ad
who since a.few months or days have.PRS.3SG start.PTCP to
investire nella criptovaluta.
invest.INF in.DEF.ART cryptocurrency
'[...] those who have only been investing in cryptocurrency for a few months or days.'

(42) A un certo punto, però, i suoi toni **hanno**
at INDF.ART some point however DEF.ART his tones have.PRS.3PL
iniziato **a** debordare
start.PTCP to overflow.INF
'At some point, however, his tones began to overflow.'

(43) La frequenza della malattia **ha** **iniziato** **a**
ART frequency of.DEF.ART disease have.PRS.3SG start.PTCP to
decrescere dal 1995.
decrease.INF since.DEF.ART 1995
'The frequency of the disease began to decrease since 1995.'

(44) La mortalità **è** **iniziata** **a** scendere sensibilmente.
ART mortality be.PRS.3SG start.PTCP.F.SG to drop.INF significantly
'The mortality rate began to drop significantly.'

In (41) the infinitive is a transitive verb and 'have' is selected for the semi-auxiliary, the infinitive is an intransitive 'have'-selection verb in (42) and 'have' is equally selected for the semi-auxiliary. In (43) and (44), the infinitive is an intransitive 'be'-selection verb, while the auxiliary is 'have' in (43) and 'be' in (44). Our

[9] Data are ours (Piedmontese and Tuscanian speakers).

hypothesis is that the infinitive has a major impact on the selection of the semi-auxiliary. Hence, a transitive infinitive or an intransitive with 'have' selection strongly disfavors 'be' for the semi-auxiliary.

However, the infinitive with 'be'-selection does allow the selection of 'have' for the semi-auxiliary. As a result, it can be hypothesized that 'have' is undergoing grammaticalization and expanding its use in these semi-auxiliary constructions with infinitive.

If we compare it to a semi-auxiliary verb with a different semantics such as *continuare* 'continue', we see that the infinitive plays a role as well.

Figure 4. Percentage of data presenting infinitives transitive and intransitive in *continuare*.

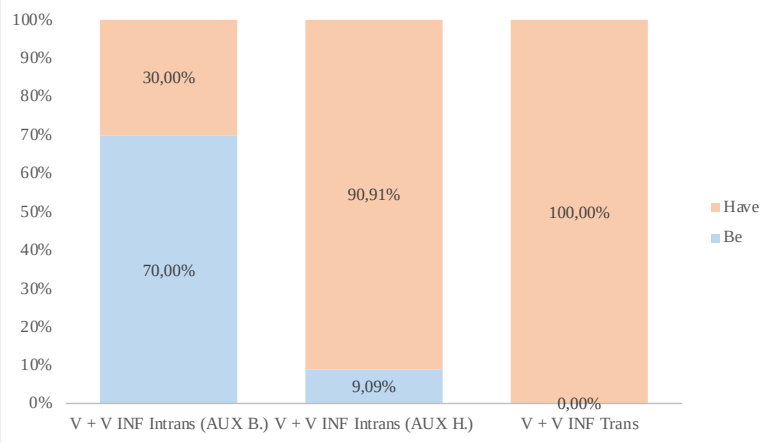


Table 8. Data presenting infinitives transitive and intransitive in *continuare*.

Parameters	**Aux. 'be'**	**Aux. 'have'**	**Total**
V + V.INF Intrans (Aux. 'be')	***49***	***21***	***70***
V + V.INF Intrans (Aux. 'have')	***1***	***10***	***11***
V + V.INF Trans	–	19	19
Total	*50*	*50*	***100***

The data show that when the infinitive is transitive, the auxiliary used for the semi-auxiliary is always 'have'. When the infinitive is an intransitive verb selecting 'have', the auxiliary used for the semi-auxiliary is 'have' in all but one case. When the infinitive is an intransitive verb selecting 'be', we have 49 occurrences with 'be' and 21 with 'have', hence a great variation.

Again, internal causation can help us understand auxiliary selection.

Figure 5. Percentages of data presenting the intransitive infinitive selecting 'be' and internal cause-agentivity-lack of agentivity parameters in *continuare*.

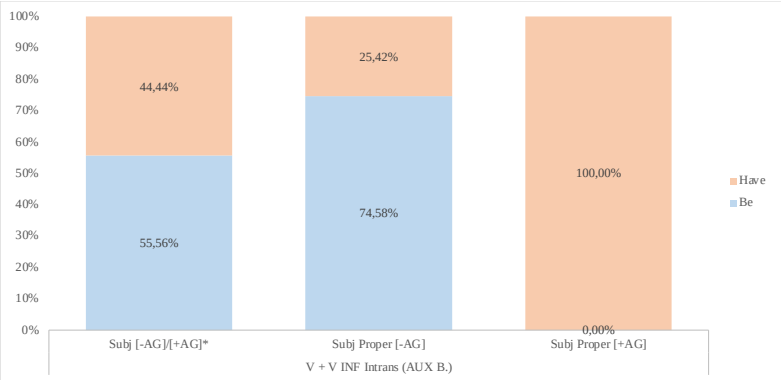


Table 9. Data presenting the intransitive infinitive selecting 'be' and internal cause-agentivity-lack of agentivity parameters in *continuare*.

Parameters	**Aux. 'be'**	**Aux. 'have'**	**Total**
V + V INF Intrans (AUX.B.) Subj [-AG] /[+AG]*	***5***	***4***	***9***
V + V INF Intrans (AUX.B.) Subj proper [-AG]	44	15	59
V + V INF Intrans (AUX.B.) Subj proper [+AG]	*/*	*2*	***2***
Total	*49*	*21*	***70***

We have considered only the infinitive whose verb is an intransitive selecting 'be', given that it is the only one showing variation. When the subject is properly agentive, there is 100% 'have' selection (2 occurrences). When the subject is properly not agentive, 'be' prevails. When the subject is characterized by *internal cause*, there is a considerable 'have' and 'be' variation.

Although the semantics of the semi-auxiliary, in this case *continuare* 'continue', plays a role, the type of the infinitive combined with the semi-auxiliary *continuare* and the feature of internal cause are major parameters. The prevalence of 'be' is obvious when the infinitive selects 'be'. However, a subject conveying the feature of *internal cause* increases the relative frequency of 'have'.

(45) Il fiumiciattolo **ha** **continuato** **a** scorrere.
DEF.ART little river have.PRS.3SG continue.PTCP to flow.INF
'The little river continued to flow.'

(46) I tir **sono** **continuati** **ad** entrare.
DEF.ART trucks be.PRS.3PL continue.PTCP.M.PL to enter.INF
'The trucks continued to enter.'

In (45) the verb *continuare* may give the impression of a subject in control of its action, but actually the semantic features of the subject play a role: a force of nature which is an internal cause. (46) also contains an internal cause subject, *i.e.* a vehicle that is forcedly driven by a person. The combination of causer's semantics, the semantics of *continuare* and the fact that the infinitive would normally select 'be' is likely to make both auxiliaries acceptable.

The discourse is complex and necessitates more pages to go into further depth, but we wanted to show how the concept of internal cause may be interesting and useful to explain auxiliary selection. This section also highlights the phenomenon of semi-auxiliaries in Italian from a new perspective.

5. CHAID and Random Forest: a statistical analysis

CHAID (acronym for Chi Squared Automatic Interaction Detection) (Kass 1980) is a type of decision tree algorithm used for predictive modeling and classification. In this contest, CHAID aims to create a tree structure that predicts the target variable (AUX) based on the values of predictor variables (*i.e.*, the sixteen parameters).
To do so, the algorithm iteratively forms subsequent smaller sub-groups: by maximizing the differences between the segments and by minimizing the difference within a given segment, the algorithm chooses the strongest predictor capable of splitting a node.

The developer defines his method in the following way: "CHAID proceeds in steps. First the best partition for each predictor is found. Then the predictors are compared and the best one chosen. The data are subdivided according to this chosen predictor. Each of these subgroups are re-analyzed independently, to produce further subdivisions for analysis. The type of each predictor determines the permissible groupings of its categories, so as to build the contingency table with the highest significance level according to the Chi-square test. This implies that there are enough observations to ensure the validity of this test" (Kass 1980: 2).

We used this model classification because it offers criteria to create homogeneous groups of sentences by maximizing inter-variability between groups and minimizing intra-variability within groups. This method gives us a solid basis to demonstrate the validity of our hypothesis in a grounded way.
The CHAID analysis was performed using the SPSS ver. 26 and STATA software ver. 15.

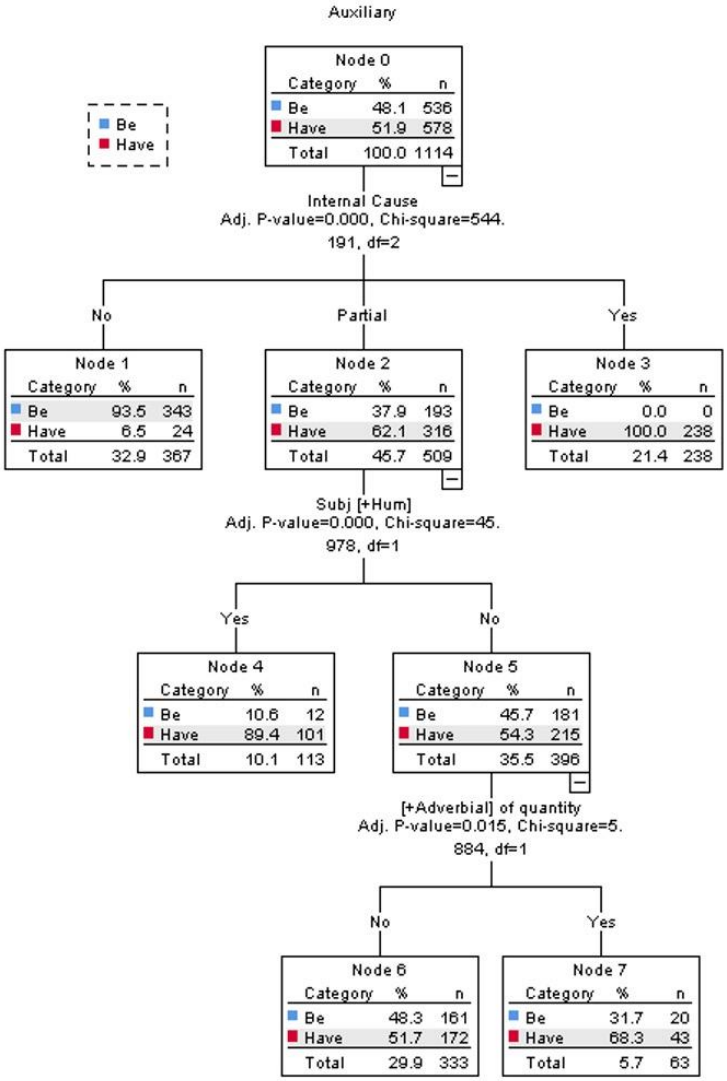
The first node (Node 0) represents the whole corpus of intransitives: 1114 is the total number of sentences containing the 13 intransitive uses of the verbs). It is split into three sub-nodes according to what CHAID estimates to be the most significant predictor (p-value < 0.001) of Auxiliary selection, viz. internal cause. When the subject is properly agentive (Node 3), all the sentences contain 'have'. On the contrary, when the internal cause is absent, and the subject is completely non-agentive, 'be' scores for 93.5% of sentences. When the subject is causative without being agentive (Node 2, characterized by internal cause without being agentive), the ratio according to auxiliary selection is approximately 60/40% with a higher percentage of 'have'.

It is worth noting that Node 2 contains far more occurrences than Node 1 and Node 3. For this reason, CHAID can find another highly significant (p-value < 0.001) parameter (Subj[+Hum]) that further splits Node 2 into two subnodes: [+Hum] subjects show a clear majority of sentences choosing 'have'. In contrast, [-Hum] subjects show an almost equal proportion between the two auxiliaries.

Finally, concerning Node 7 (subjects characterized by adverbial of quantity) and Node 6, the absence of adverbials of quantity does not impact the auxiliary selection, whereas its presence yields a higher probability of 'have'. This could be

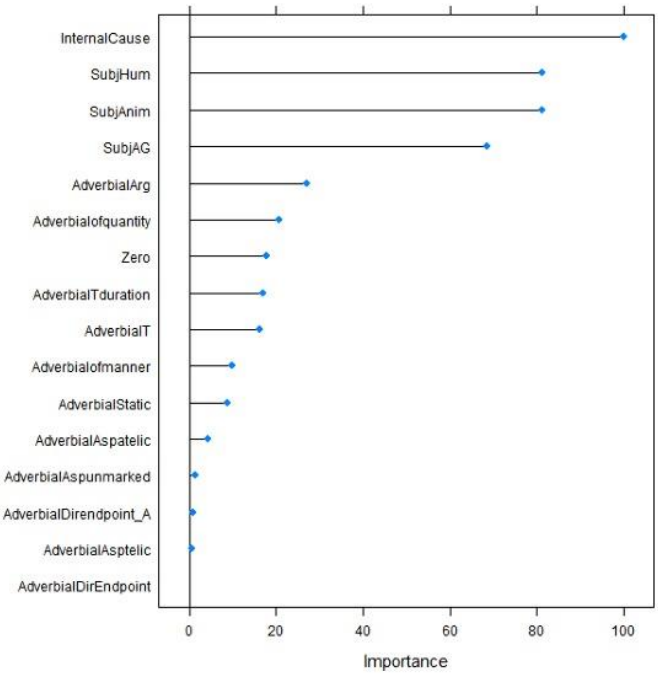
because an adverbial of quantity may be misinterpreted as an object complement, as we have seen with *cambiare*.

Figure 6. The decision tree of the 13 verbs sample.



To visualize in an alternative way the importance of each predictor in the analysis, we submitted the same data set to the Random Forest (RF) model algorithm (Breiman 2001; Liaw and Wiener 2002). RF shows the significance of different predictors in determining the target variable (AUX). So, it helps in understanding which variables influence the classification most. This procedure was performed through R-package "caret" ver. 6.0 (Kuhn, 2008). The outcome reveals that (Figure 7) in the ranking of the 16 parameters, 4 emerged as the most important ones: internal cause, human trait, animacy and agentivity (or lack thereof).

Figure 7. Ranking of the 16 parameters.



The fact that both methods (CHAID and RF) identified the same parameters as dominant corroborates the correctness of our analysis. To evaluate the performance of the CHAID model, we create a classification table in which the prediction made by the model is compared to the known results in the dataset.

The overall percentage of correctly classified cases demonstrates the validity of our model. In 80% of cases, the parameters we attribute to the sentences of our sample allow CHAID to correctly predict which auxiliary a given verb from the list will choose. This means the sixteen parameters can explain more than eighty percent of auxiliary alternation for the verbs analyzed in this study[10] (Table 10). In other words, the recursive partitioning performed by CHAID confirms our intuition: quantitative results thus confirm qualitative interpretation about auxiliary selection[11].

Table 10. Classification table for CHAID model

	Predicted		
Observed	Be	Have	Percent Correct
Be	343	193	64.0%
Have	24	554	95.8%
Overall Percentage	32.9%	67.1%	80.5%

[10] The RF classification results are similar and are not reported due to page limit constraint.

[11] We were also unable to include CHAID results with semi-auxiliaries due to space constraints. We chose to prioritize clarity and present the analysis with main verbs in greater detail to demonstrate the method's significant contribution.

6. Conclusions

Auxiliary selection is a very complex topic, at the interface of syntax and semantics. It is well-known that auxiliary selection in the perfect tense has partly a syntactic basis: transitive verbs select 'have' whereas intransitive verbs select either 'have' or 'be' or may combine with both auxiliaries. Moreover, it is argued that for intransitive verbs the choice of auxiliary can be accounted in terms of two parameters, telicity and agentivity (Sorace 2000: 861-862).

Although considerable research attention has been devoted to these parameters having an impact on auxiliary selection, especially for Italian, the present study aims to innovate in several respects. First, it focuses on a grey area in this research, the so-called *peripheral verbs* (Sorace 2000: 860), *i.e.* verbs that occur both with 'have' and 'be' in the perfect tense. Moreover, it highlights an important, hitherto unnoticed difference between these verbs used as main verb or as a semi-auxiliary with respect to auxiliary selection. Finally, it brings new empirical data and unveils recurring patterns, thanks to its methodological approach based on a large-scale corpus, and combining qualitative and quantitative analysis using state-of-the-art statistical tools.

With respect to the verbs with double auxiliation investigated in the present study, the major parameter accounting for the auxiliary selection proved to be the feature of *internal cause*, rather than agentivity. As shown in figure 6, when the subject is negatively marked with respect to this feature, there is a very high probability that the auxiliary is 'be'; when the subject is positively marked with respect to this feature and is moreover agentive, the auxiliary 'have' is chosen; however, even with a subject devoid of agentivity but positively marked with respect to the feature 'internal cause' the auxiliary 'have' is also dominant. Secondly, our statistical results show that the feature [±human] is another key factor accounting for auxiliary selection for the verbs with double auxiliation investigated in the present study, and that it operates independently of agentivity: for non-agentive subjects that convey the features of internal cause factor and [+human], the auxiliary 'have' is highly probable. Thirdly, and unexpectedly, with respect to the aspectual parameter of telicity, this parameter was found to have no significant effect on auxiliary selection for the double auxiliation verbs investigated in the present study.

Besides these results, which concern the verbs under study when they are used as the main verb, our study also yielded original results for the same verbs when used as a semi-auxiliary. In particular, it is empirically shown that the semi-auxiliary has a high degree of transparency with respect to the auxiliary selection features of the infinitive, *i.e.* the semi-auxiliary selects 'have' when the infinitive is transitive, it also selects 'have' when the verb in the infinitive is a 'have'-selecting intransitive verb, and it tends to select 'be' when the verb in the infinitive is a 'be'-selecting intransitive verb.

From a methodological point of view, the combination of qualitative corpus analysis and statistical analysis was crucial. It enabled us to not only corroborate the relevance of the factors studied, but also to demonstrate the statistical significance and thus their validity as predictors. Two important statistical tools (CHAID and RF) found both internal cause and the human trait as relevant, confirming our hypotheses with respect to the parameters having an impact on auxiliary selection for double auxiliation verbs.

Acknowledgments

We would like to thank the University of Verona (Italy) for the fully funded PhD position of Greta Viale and Sorbonne Université (France) for funding the participation at Going Romance 2022 conference. A previous version of this work was presented at Going Romance 2022 at Universitat Autònoma de Barcelona. We are grateful for the interesting exchange we had with the audience.

References

Ackema, Peter, & Antonella Sorace. 2017. Auxiliary selection. In M. Everaert, & H. van Riemsdijk (eds), *The Blackwell Companion to Syntax*. 2nd ed., 424–455. Wiley-Blackwell. https://doi.org/10.1002/9781118358733.wbsyncom072

Amato, Irene. 2022. Auxiliary selection is Agree: person-driven and argument-structure-based splits. In O. Matushansky, L. Roussarie, M. Russo, E. Soare, & S. Wauquier (eds), Selected papers from Special issue of *Isogloss Open Journal of Romance Linguistics* 8(2)/10: 1–20. https://doi.org/10.5565/rev/isogloss.131

Barbiers, Sjef, & Rint Sybesma. 2004. On the different behavior of auxiliaries. *Lingua* 114(4): 389–398. http://dx.doi.org/10.1016/S0024-3841(03)00065-2

Bentley, Delia, & Thórhallur Eythórsson. 2004. Auxiliary selection and the semantics of unaccusativity. *Lingua* 114(4): 447–471. https://doi.org/10.1016/S0024-3841(03)00068-8

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324

Burzio, Luigi. 1986. *Italian syntax: A Government-Binding approach*. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-4522-7

Carlier, Anna, & Laure Sarda. 2010. Le complément de la localisation spatiale: Entre argument et adjoint. In F. Neveu., V. Muni-Toké., J. Durand., T. Klingler, L. Mondada, & S. Prévost (eds), *Actes du CMLF'10*, 2057–2073. Paris: ILF. https://dx.doi.org/10.1051/cmlf/2010251

Cinque, Guglielmo. 2004. "Restructuring" and Functional Structure. In A. Belletti (ed.), *Structures and Beyond: The Cartography of Syntactic Structures*, vol. 3, 132–191. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780195171976.003.0005

D'Alessandro, Roberta. 2017. When you have too many features: auxiliaries, agreement and clitics in Italian varieties. *Glossa: A Journal of General Linguistics* 2(1): 50, 1–36. https://doi.org/10.5334/gjgl.102

DeLancey, Scott. 1984. Notes on Agentivity and Causation. *Studies in Language* 8(2): 181–213. https://doi.org/10.1075/sl.8.2.05del

Dowty, David R. 1979. *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ.* Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-9473-7

Ferrer i Cancho, Ramon, & Ricard V. Solé. 2003. Least effort and the origins of scaling in human language. *PNAS* 100(3): 788–791. https://doi.org/10.1073/pnas.0335980100

Finocchiaro, Chiara. 2002. Sensitivity to the verb [±agentive] feature: the case of an aphasic subject. *Journal of Neurolinguistics* 15(3/5): 433–446. https://doi.org/10.1016/S0911-6044(01)00033-1

Flaux, Nelly, & Danièle Van de Velde. 2000. *Les Noms en Français: Esquisse de Classe*. Collection L'essentiel français. Paris: Ophrys. https://doi.org/10.1017/S0959269501290267

Giancarli, Pierre-Don. 2015. Auxiliary Selection with intransitive and reflexive verbs: the limits of gradience and scalarity, followed by a proposal. In M. Rosemeyer, & R. Kailuweit (eds), *Auxiliary Selection Revisited. Gradience and Gradualness*, 79–123. Berlin: De Gruyter. https://dx.doi.org/10.1515/9783110348866-004

Gillmann, Melitta. 2015. Auxiliary selection in closely related languages. The case of German and Dutch. In M. Rosemeyer, & R. Kailuweit (eds), *Auxiliary Selection Revisited. Gradience and Gradualness*, 333–358. Berlin: De Gruyter. https://doi.org/10.1515/9783110348866-012

Grimshaw, Jane. 1990. *Argument Structure*. The MIT Press. CA: Massachusetts https://doi.org/10.1093/jos/11.1-2.103

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, & Vít Suchomel. 2013. The TenTen corpus family. *7th International Corpus Linguistics Conference* CL: 125–127.

Kass, Gordon V. 1980. An exploratory technique for investigating large quantities of categorical data. *App. Statist* 29(2): 119–127. https://doi.org/10.2307/2986296

Keller, Frank, & Antonella Sorace. 2003. Gradient auxiliary selection and impersonal passivization in German: an experimental investigation. *Journal of Linguistics* 39(1): 57–108. https://doi.org/10.1017/S0022226702001676

Kuhn, Max. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28(5): 1–26. https://doi.org/10.18637/jss.v028.i05

Ledgeway, Adam. 2019. Parameters in the development of Romance perfective auxiliary selection. In M. Cennamo, & C.Fabrizio (eds), *Historical Linguistics 2015*:

22nd International Conference on Historical Linguistics. Naples July 2015, 343–384. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/CILT.348.17LED

Legendre, Geraldine. 2007. Optimizing auxiliary selection in Romance. Split Auxiliary Systems: A cross-linguistic perspective. In R. Aranovich (ed.), 145–180. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.69.08leg

Levin, Beth, & Malka Rappaport Hovav. 1995. *Unaccusativity. At the Syntax-Lexical Semantics Interface.* Cambridge, *MA:* MIT Press. https://doi.org/10.1017/S0022226796276571

Levin, Beth, & Malka Rappaport Hovav. 1998. Building verb meanings. In M. Butt, & W. Geuder (eds), *The Projection of Arguments: Lexical and Compositional Factors*, 97–134. Stanford, CA: CSLI Publications.

Liaw, Andy, & Matthew Wiener. 2002. Classification and Regression by RandomForest. *R News* 2(3): 18–22. http://CRAN.R-project.org/doc/Rnews/

Martin, Fabienne. 2020. Aspectual differences between agentive and non-agentive uses of causative predicates. In E.A. Bar-Asher Siegal, & N. Boneh (eds), *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop*, 257–294. Cham: Springer. https://doi.org/10.1007/978-3-030-34308-8_8

Martin, Fabienne, & Florian Schäfer. 2017. Sublexical modality in defeasible causative verbs. In A. Arregui, M. Rivero, & A. Salanova (eds), *Modality Across Syntactic Categories*, 87–108. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198718208.003.0006

Mateu, Jaume. 2009. Gradience and auxiliary selection in Old Catalan and Old Spanish. In P. Crisma, & G. Longobardi (eds), *Historical Syntax and Linguistic Theory*, 176–193. Oxford: Oxford University Press. https://doi.org/10.1093/ACPROF%3AOSO%2F9780199560547.003.0011

McFadden, Thomas. 2007. Auxiliary selection. *Language and Linguistics Compass* 1(6): 674–708. https://doi.org/10.1111/j.1749-818X.2007.00034.x

McLendon, Sally. 1978. Ergativity, case, and transitivity in Eastern Pomo. *International Journal of American Linguistics* 44(1):1–9. https://doi.org/10.1086/465512

Perlmutter, David M. 1978. Impersonal passives and the Unaccusative Hypothesis. *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*, 157–190. https://doi.org/10.3765/bls.v4i0.2198

Reinhart, Tanya. 2016. The theta system: syntactic realization of verbal concepts. In M. Everaert, & M. Marelj (eds), *Concepts, Syntax, and Their Interface: The Theta*

System, 1–112. Cambridge, MA: MIT Press Scholarship Online. https://doi.org/10.7551/mitpress%2F9780262034135.003.0001

Reinhart, Tanya. 2002. The theta system: an overview. *Theoretical Linguistics* 28(3): 229–290. https://doi.org/10.1515/thli.28.3.229

Rizzi, Luigi. 1982. *Issues in Italian Syntax.* Dordrecht: Foris. https://doi.org/10.1515/9783110883718

Sorace, Antonella & Legendre, Géraldine. 2003. Auxiliaries and intransitivity in French and in Romance. (English version of 'Auxiliaires et intransitivité en français et dans les langues romanes'). In D. Godard (ed.), *Les Langues Romanes: Problèmes de la Phrase Simple*, 243–268. Paris: CNRS Edition. https://doi.org/10.1093/acprof%3Aoso%2F9780199257652.003.0010

Sorace, Antonella. 2015. The cognitive complexity of auxiliary selection: from processing to grammaticality judgements. In M. Rosemeyer, & R. Kailuweit (eds), *Auxiliary Selection Revisited. Gradience and Gradualness*, 23–43. Berlin: De Gruyter. https://doi.org/10.1515/9783110348866-002

Sorace, Antonella. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76(4): 859–890. https://doi.org/10.2307/417202

Talmy, Leonard. 1988. Force dynamics in language and cognition. *Cognitive Science* 12: 49–100. https://doi.org/10.1016/0364-0213(88)90008-0

Tenny, Carol L. 1994. *Aspectual Roles and the Syntax-Semantics Interface*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-1150-8

Ter Meulen, Alice G. B. 2004. The dynamic semantics of aspectual adverbs. In O. Bonami, & P. Cabredo Hofherr (eds), *Empirical Issues in Formal Syntax and Semantics* 5: 241–253. http://www.cssp.cnrs.fr/eiss5/ter-meulen/ter-meulen-eiss5.pdf

Van Valin, Robert D. 1993. *Advances in Role and Reference Grammar*. Amsterdam/ Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1016/0024-3841(95)90011-X

Van Valin, Robert D. 1990. Semantic parameters of split intransitivity. *Language* 66(2): 221–260. Linguistic Society of America. https://doi.org/10.2307/414886

Veecock, Candace. 2012. *Agentivité, Modalités de Contrôle et Subjectivité*. Ph.D. thesis. Université Michel de Montaigne – Bordeaux III. HAL Id : tel-00910818 version 1

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2): 143–160. https://doi.org/10.2307/2182371

Vlach, Frank. 1993. Temporal adverbials, tenses and the perfect. *Linguistics and Philosophy* 16: 231–283. http://dx.doi.org/10.1007/bf00985970

Washio, Ryuichi. 2004. Auxiliary selection in the East. *Journal of East Asian Linguistics* 13: 197–256. https://doi.org/10.1023/B:JEAL.0000038249.86375.a5

Zipf, George Kingsley. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio Books. https://doi.org/10.1037/h0052803